Edited by
Joseph Emonds and Markéta Janebová

Language Use and Linguistic Structure

Proceedings of the Olomouc Linguistics Colloquium 2016

OLOMOUC
MODERN
LANGUAGE
SERIES
VOL. 5

Language Use and Linguistic Structure

Proceedings of the Olomouc Linguistics Colloquium 2016

Edited by Joseph Emonds and Markéta Janebová

Palacký University
Olomouc

2017

OLOMOUC MODERN LANGUAGE SERIES (OMLS) publishes peer-reviewed proceedings from selected conferences on linguistics, literature, and translation studies held at Palacký University Olomouc, Czech Republic.

Published so far:
OMLS, Vol. 1: *Teaching Translation and Interpreting in the 21st Century* (2012)
OMLS, Vol. 2: *Tradition and Trends in Trans-Language Communication* (2013)
OMLS, Vol. 3: *Language Use and Linguistic Structure. Proceedings of the Olomouc Linguistics Colloquium 2013* (2014)
OMLS, Vol. 4: *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structur*e (2014)

OLOMOUC MODERN LANGUAGE SERIES
Vol. 5

Language Use and Linguistic Structure

Proceedings of the Olomouc Linguistics Colloquium 2016

organized by

Department of English and American Studies
Faculty of Arts, Palacký University Olomouc, Czech Republic
June 9–11, 2016

Edited by Joseph Emonds and Markéta Janebová

Palacký University
Olomouc
2017

Reviewer of the volume: Mojmír Dočekal (Masaryk University, Brno)

Each of the contributions was peer-reviewed by two anonymous reviewers in addition to the main reviewer prior to the publication of this volume.

First Edition

ISBN 978-80-244-5173-2
(online: PDF; available at http://anglistika.upol.cz/olinco2016proceedings/)

ISBN 978-80-244-5172-5
(print)

Table of Contents

Syntactic Features and Their Interpretations

Word Study and the Lexicon: Phonological Approaches

Word Study and the Lexicon: Corpus Approaches

Alphabetical List of Authors

Gábor Alberti
University of Pécs
Pécs, Hungary

Tania Avgustinova
Saarland University
Saarbrücken, Germany

Anna Babarczy
Research Institute for Linguistics
Hungarian Academy of Sciences
Budapest, Hungary

Giulia Bellucci
University of Florence, Italy
Florence, Italy

Ágnes Bende-Farkas
Department of Cognitive Science (BME)
Budapest University of Technology
and Economics
Budapest, Hungary

Pavel Caha
Masaryk University
Brno, Czech Republic

Péter Csatár
University of Debrecen
Debrecen, Hungary

Lena Dal Pozzo
University of Florence, Italy
Florence, Italy
Pontifical Catholic University
of Rio de Janeiro
Rio de Janeiro, Brazil

Tomáš Duběda
Charles University in Prague
Prague, Czech Republic

Joseph Emonds
Palacký University
Olomouc, Czech Republic

Judit Farkas
Research Institute for Linguistics
Hungarian Academy of Sciences
Budapest, Hungary

Ludovico Franco
Nova University of Lisbon
Lisbon, Portugal

Volker Gast
Friedrich Schiller University
Jena, Germany

Wojciech Guz
The John Paul II Catholic University
of Lublin
Lublin, Poland

Kateřina Havranová
Palacký University
Olomouc, Czech Republic

Anders Holmberg
Newcastle University
Newcastle upon Tyne, Great Britain
University of Cambridge
Cambridge, Great Britain

Jen Ting
National Taiwan Normal University
Taipei, Taiwan

Enikő Tóth
University of Debrecen
Debrecen, Hungary

Jorge Vega Vilanova
University of Hamburg
Hamburg, Germany

Ludmila Veselovská
Palacký University
Olomouc, Czech Republic

Susi Wurmbrand
University of Connecticut
Storrs, CT, USA

Joanna Zaleska
University of Leipzig
Leipzig, Germany

Acknowledgements

The editors are grateful to all those who have helped make this book a reality. Above all, we would like to thank all the authors for both their enthusiastic participation in the conference and their cooperation in the time consuming editorial process. We would also like to express gratitude to our colleagues and students from the Faculty of Arts of Palacký University, Olomouc, for their efforts related to the organization of the Olomouc Linguistics Colloquium (OLINCO) 2016 conference and the subsequent publishing activities. We greatly appreciate the assistance of Eva Nováková, Irena Pauková, Monika Pitnerová, and Andrea Ryšavá, without whose tireless devotion to the editing work the proceedings would never have come into existence.

We would also like to express our immense gratitude to all the reviewers who devotedly participated in the process of accepting and reviewing the papers for the conference and later another round of the peer-reviewing process for the proceedings. Special thanks are also due to Mojmír Dočekal of Masaryk University, Brno, for the overall review of the proceedings.

Joseph Emonds
Markéta Janebová

Introduction

The articles in this volume are based on papers and posters presented at the Olomouc Linguistics Conference (OLINCO) at Palacký University in the Czech Republic in June 2016. This conference welcomed papers that combined analyses of language structure with generalizations about language use. The essays here represent, we think, the best of the conference contributions. All these papers have been doubly reviewed, with one reviewer always external to Palacký University, and revised on the basis of these reviews. The sections in the Table of Contents have been determined, in the final analysis, by their subject matter rather than by a priori "areas." What follows is the briefest of synopses of each of the papers, grouped into the areas reflected in the Table of Contents.

Morphosyntax of Agreement Features

Syntacticians are always drawn to constructions involving "agreement," i.e., multiple constituents that co-vary along specifiable formal lines, and the contributions to this volume testify to their continued efforts to clarify this broad issue: how and in what ways do constituents in different positions come to agree?

Susi Wurmbrand shows how gender mismatch, pluralia tantum, and polite pronouns affect German agreement in attributive, predicate, pronominal and ellipsis contexts. These patterns argue for two types of nominal ellipsis and for a dual feature system that justifies an Agreement Hierarchy, with room for language-specific deviations. **Jorge Vega Vilanova** proposes, based on new data, that grammaticalization with a current theory of Agree accounts for how Old Romance past participle agreement contracts into more restricted modern uses. He links loss of this agreement to direct object specificity, differential object marking and the emergence of clitic doubling.

Ludmila Veselovská uses extensive corpus data to show how two Czech quantifiers, *mnoho*/*málo* "a lot / few" in oblique cases support Pesetsky's recent Case theory, in particular for the category Q. In these terms, she further proposes a new account of the previously unexplained "adverbial" inflection on Czech Qs. **Krisztina Szécsényi and Tibor Szécsényi** argue for a cyclic rather than long distance account of Hungarian definiteness agreement. They show that properties of objects in multiple infinitives support a covert agreement analysis even when overt morphology for it is lacking.

Anders Holmberg and On-Usa Phimsawat contrast the properties of overt and null inclusive generic pronouns. Using data from several languages with and without agreement, they argue that their restriction to human reference crucially depends on the presence of agreement. Their explanation is based on feature architecture. **Joseph Emonds** analyzes French verbal clitics without movement devices, using four lexical entries whose forms are determined by principles of grammatical lexicons. In this system, each morpheme spells out at most one marked feature. He also argues that all such clitics replace clause-mates of their verbal host and never result from raising.

Syntactic Derivations

Since most current models of formal grammar involve a sentence's syntactic structure being modified by operations that take place at different "derivational levels," several contributions here focus on the derivational architecture of this model and propose modifications as to how the levels affect syntactic structure.

Mark Newson and Márton Kucsera argue against the view that multiple Wh-constructions involve Wh-elements reinterpreted as quantifiers. They propose instead that underlying universal quantifiers are realized as Wh-elements. They claim that their hypothesis radically simplifies the grammar of these constructions. **Ángel L. Jiménez-Fernández** uses the sub-extraction case of *wh*-movement in Spanish psych-verb constructions to determine the categorial nature of the P *a* in accusative and dative objects. He argues that the sub-extraction criterion argues for analyzing *a* as a Kase marker with an edge feature that permits extraction during the derivation.

Kateřina Havranová compares two much discussed types of Dutch nominalization, bare nominal infinitives and infinitives introduced by a definite article. She shows that their internal differences can be explained by application at slightly different levels of a single operation combing "Merge" and "Categorial Switch." **Pavel Caha** weighs the issue of whether Bobaljik's Root Suppletion Generalization constitutes evidence for the view that such principles should be in the lexicon. He presents evidence for an alternative non-lexical mechanism for blocking suppletion that crucially involves adjacency. **Gábor Alberti and Judit Farkas** argue against a head-final analysis of Hungarian, proposing instead that raising into specifiers accompanied by remnant movement improves analyses of aspectual and de-verbal nominal constructions.

Syntactic Features and Their Interpretations

Grammarians of every strip, including both those inclined to formalism and those less so, want to find the semantic "essence" of what they study, both to clarify the nature of the basic elements and to better understand the mapping between form and function. Several of the papers in this volume concern themselves centrally with this issue.

Enikő Tóth and Péter Csatár conclude, based on a sentence-picture verification task, that exhaustivity and expectedness interpretations do not distinguish Hungarian preverbal and syntactically unmarked focus. Both can be exhaustive, and counter to earlier views, exhaustivity of preverbal focus is rather a pragmatic phenomenon. On this same topic, **Tamás Káldi, Anna Babarczy, and Ágnes Bende-Farkas** propose that the pragmatic inference of exhaustivity in Hungarian preverbal focus results from scalar implicature generation. They confirm this hypothesis by finding a strong context dependence and predicted delays in eye-tracking experiments.

M. Rita Manzini and Leonardo M. Savoia investigate varieties of the Romance feminine *-a*, which has additional dialectal uses for "cohesive" plurals and for singulars interpreted like the Italian plural *-a*. Their data supports their claim that *-a* can be

specified as [aggregate] for mass nouns and as [⊆] for plurals. **Ludovico Franco, Giulia Bellucci, Lena Dal Pozzo, and Rita Manzini**, using comparative Uralic evidence, argue that Finnish "inner case" (both genitive and *-l, -s*) are best characterized as a part-whole / zonal inclusion relator, while traditionally labelled Uralic adpositions are best characterized rather as Axial Parts.

Jen Ting studies the Taiwan Mandarin expression *bucuo*-V "good to V," showing first that it is a word rather than syntactic and then that its morphology is Taiwan Southern Min based on *bebai*/*bephai*-V "good to V," which Taiwan Mandarin has borrowed via language contact. **Wojciech Guz** analyzes Polish relatives in terms of a head noun's definiteness and specificity. Corpus data and complementary tests of constructed examples strongly correlate *co* relatives with definite/specific NP heads (and realis clauses), while *który* relatives tend towards indefinites, often non-specific heads (and irrealis clauses).

Word Study and the Lexicon: Phonological Approaches

As with numerous linguistics conferences in recent decades, the OLINCO organizers would like to see more focus on phonology. So we are happy to have two papers in phonology, but at the same time disappointed not to have more.

Joanna Zaleska uses informant data from a devised word game, based on Pig Latin, which helps to settle the issue of whether Polish [i] and [ɨ] are underlyingly the same or different. On the basis of this data, she argues in favor of distinct underlying sources. **Tomáš Duběda** formulates several principles of phonological borrowing and categorizes them as either "integrative" or "modular." He provides quantitative evidence for their relative scope and formulates psycholinguistic hypotheses for an adaptation model of borrowing.

Word Study and the Lexicon: Corpus Approaches

Volker Gast uses the BNC data to show that the two operators differ primarily in terms of the downward entailing operators they are licensed by. While *even* tends to occur more frequently in the scope of local negation than *so much as*, the latter operator is more commonly found in conditionals and without-PPs. A certain effect of the category of the co-constituent can also be observed.

The contribution by **Klára Jágrová, Irina Stenger, Roland Marti, and Tania Avgustinova** is a contrastive one: national corpora of four Slavic languages, Bulgarian, Czech, Polish, and Russian, are compared to identify the share of cognates between the languages. The paper aims at discovering the mechanisms by which intercomprehension in these closely related languages works: the measures of lexical and orthographic distance serve as predictors for the performance of monolingual Slavic readers in their attempt to understand a related Slavic language. Lexical asymmetries for all language combinations and directions of reading are observed. The Czech subjectivity lexicon

is examined in the paper by **Jana Šindlerová and Aleš Tamchyna**. The aim is to document the behavior of verb valency complementations regarding the position of the target of evaluation within the valency frame. The authors classify the types of evaluative meaning expressed by the verbs and identify shared characteristic features considering the valency patterns of the verbs.

The next two papers use data from the parallel corpus InterCorp: in the first one, **Denisa Šebestová and Markéta Malá** analyze the communicative polyfunctionality of the affix *-pak* and of the three particles containing it: the contrastive data reveal that the *-pak* particles have both content/speaker-related functions and communication/addressee-oriented functions (Kranich and Gast 2015). **Olga Nádvorníková** explores reasons for and consequences of shifts in the segmentation of sentences, i.e., the joining and splitting of sentences, in translations into English, Czech and French. The author focuses on two different explanations of these shifts: the hypothesis of information density and the theory of translation universals.

The section closes with **Magdalena Szczyrbak**'s contribution, which examines the patterns of use involving the verb *say* in police interviews carried out in a homicide investigation. The aim is to establish how legal professionals and laypersons deploy say in interaction and to compare selected "saying" routines in police in trial data.

We hope that all readers will find several papers here to be of interest to them and their fellow researchers. It was both challenging and gratifying to organize and participate in the conference in person, and now we want to extend the challenges and the results of this linguistics forum to a wider audience of those who can participate via the written word, which was, after all, invented by our species so that the pleasures and benefits of speech and hearing could be extended to the widest possible audience.

Joseph Emonds
Markéta Janebová
Michaela Martinková

Morphosyntax of Agreement Features

Formal and Semantic Agreement in Syntax: A Dual Feature Approach

Susi Wurmbrand

University of Connecticut, Storrs, USA

susanne.wurmbrand@uconn.edu

Abstract: This paper surveys the distribution of formal and semantic agreement in German, using three types of trigger nouns (gender mismatch nouns, pluralia tantum nouns, and polite pronouns) in four syntactic contexts (attributive, predicate/T, pronouns, and nominal ellipsis). The distribution of agreement is shown to be dependent on the properties of the controller and the target, as well as the type of agreement dependency. The paper provides new evidence for the existence of two types of nominal ellipsis, and establishes a context in which predicative agreement can be tested in German. The findings lead to a refined Agreement Hierarchy, and a dual feature system is proposed which derives the basic tendencies of the Agreement Hierarchy and leaves room for language-specific deviations.

Keywords: semantic agreement; agreement mismatches; agreement hierarchy; nominal ellipsis; phi-features

1. Introduction

The phenomenon of *formal* (= morphological) vs. *semantic agreement* is wide-spread cross-linguistically. Formal agreement is used to refer to agreement with the formal features of the controller/trigger, whereas semantic agreement refers to agreement with semantic features of the controller. In most cases, formal and semantic agreement look the same, however, configurations involving controllers with mismatching formal and semantic features allow us to tease apart the two forms of agreement. If an agreement target realizes a feature value that is different from the morphological feature value expressed by the controller, we speak of an agreement mismatch. In this paper, I summarize the distribution of agreement mismatches in German and provide new data from nominal ellipsis showing that when agreement is not determined NP-internally,

predicate agreement must be semantic agreement. I will show how the new observations can be aligned with Corbett's (1979; 2006) *Agreement Hierarchy* and sketch a feature approach to derive the patterns.[1]

2. German Agreement Mismatches and the Agreement Hierarchy

German is a language with grammatical gender, which means that nouns are lexically specified for a particular (formal) gender, which cannot always be related to the semantic properties of the noun (e.g., there are two nouns corresponding to "car," *Wagen* and *Auto*, however they differ in formal gender—the first one is masculine whereas the second one is neuter). An example of a noun which shows mismatching formal and semantic gender is *Mädchen* "girl," which is formally neuter but semantically feminine. Such nouns allow either formal or semantic agreement when they control agreement on a pronoun. As shown in (1), a pronoun bound or co-referent with an NP headed by the noun *Mädchen* can occur either as neuter (formal agreement, [1a]) or feminine (semantic agreement, [1b]).

(1) (a) Das Mädchen genießt seinen Urlaub.
the.N.SG girl enjoys its.N.SG vacation
"The girl is enjoying her vacation."

(b) Das Mädchen genießt ihren Urlaub.
the.N.SG girl enjoys her.F.SG vacation
"The girl is enjoying her vacation."

Agreement mismatches are not possible in every agreement configuration, and languages differ regarding which dependencies can display semantic agreement. The cross-linguistic distribution follows the *Agreement Hierarchy* in (2) (Corbett 1979, 204; Corbett 2006, 207), an implicational hierarchy which states that the further right an element is on this hierarchy the more likely it is to allow semantic agreement. Furthermore, if in a language an element (anywhere on the scale in [2]) allows semantic agreement, all elements to the right of that element also allow semantic agreement, and, conversely if an element does not allow semantic agreement, all elements to its left also do not allow semantic agreement.

(2) [formal] ⇠ attributive — predicate — relative — personal PRON ⇢ [semantic]

1 This paper does not offer room to discuss other languages. In addition to German, so far, the paradigms presented in this paper have been tested and replicated in Dutch, Slovenian, Czech, and Greek, and similar effects have been observed in these languages. For an account covering the similarities and differences, see Wurmbrand (2016b).

Relative pronouns differ from personal pronouns in German in not allowing semantic agreement, which is illustrated in (3). A relative clause modifying an NP headed by the noun *Mädchen* must occur with neuter—i.e., formal—agreement on the relative pronoun, and feminine is impossible. Note however that, as shown in (3a), a possessive pronoun within the relative clause is still free to choose semantic agreement.

(3) (a) Das Mädchen, das ihren Urlaub genießt . . .
the.N.SG girl that.N.SG her vacation enjoys . . .
"The girl that is enjoying her vacation."

(b) *Das Mädchen, die ihren Urlaub genießt . . .
the.N.SG girl who.F.SG her vacation enjoys . . .

The impossibility of semantic agreement on relative pronouns leads to the expectation that predicate and attributive agreement can also only realize formal agreement in German. This is shown to be the case for attributive adjectives and determiners in (4a, b) and for verb (i.e., T-) agreement in (4c, d). Collective nouns such as "committee" allow semantic plural agreement in certain languages, however, this is not possible in German, (4c). Similarly, polite pronouns are formally plural, even when they are used to address a single person. As shown in (4d), the polite pronoun *Sie* "you.polite" can only trigger plural agreement on the finite verb in German and using semantic singular agreement (to indicate a single addressee) is not possible.

(4) (a) ein nettes Mädchen / *Frau
a.N.SG nice.N.SG girl.N / *woman.F
"a nice girl/woman"

(b) eine nette *Mädchen / Frau
a.F.SG nice.F.SG *girl.N / woman.F
"a nice girl/woman"

(c) Das Komittee hat / *haben getagt
the committee has.3.SG / *have.PL met
"The committee has/*have met"

(d) Sie haben / *hat gewonnen.
ADDRESSEE.POL have.3.PL / *has.3.SG won
"You (pol.) have one."

The split between semantic and formal agreement in German is thus between relative and personal pronouns as indicated in (5).

(5) [formal] ⇠ attributive — predicate — relative || personal pron ⇢ [semantic]

In addition to the agreement hierarchy in (2), the category "predicate" involves a set of elements, which also follow an implicational hierarchy, namely: verb » participle » adjective » noun (Comrie 1975). Above, we have seen examples of verb/T-agreement. Since participles and predicative adjectives do not agree in German, these categories cannot be tested for agreement mismatches. Predicative NPs/DPs, on the other hand, can be shown to not require formal agreement. In (6a), a 2nd person pronoun is in a predicative relation with a 3rd person DP, thus there is a person mismatch. In (6b), we find a gender mismatch, since a masculine pronoun is in a predicative relation with a neuter DP. Finally in (6c), when addressing a single person, the polite plural pronoun can only be associated predicatively with a singular DP, thereby yielding a number mismatch between the subject (controller) and the target predicate (see Wechsler 2011; Wechsler and Hahm 2011).

(6) (a) Du bist das Mädchen, das . . .
you.2.SG are.2.SG the.N.SG girl who.N . . .

(b) Er ist das Opfer.
he.M.3.SG is.3.SG the.N.SG victim.

(c) Sie sind der Verlierer / *die Verlierer.
3.PL (POL) are.3.PL the.M.SG loser / *the.PL loser
"You (addressing a single person politely) are the loser."

In the next section, I turn to another predicative DP configuration, one in which the predicate DP involves nominal ellipsis. We will see that an interesting agreement pattern arises, which will lead to two observations. First, agreement in predicate nominal contexts exists in German. Second, confirming the suspicion noted in Corbett (2006, 233), the relative ranking of the predicate hierarchy is to some extent independent of the non-predicate elements of the agreement hierarchy in that predicate nouns undergo semantic agreement more frequently than personal pronouns.

3. Agreement in Nominal Ellipsis

3.1 Two Types of Nominal Ellipsis

Before looking at the details of agreement, we need to have a brief look at the properties of nominal ellipsis. Nominal ellipsis, like verbal ellipsis, comes in two types—surface

and deep anaphora (Hankamer and Sag [1976]; Merchant [2014]; see also Merchant [forthcoming] and Saab [forthcoming] for overviews). Surface ellipsis involves deletion, possibly at PF, of an N, NP, or *n*P in the presence of a parallel antecedent. In this form of ellipsis, the elided part thus contains a specific noun during the syntactic derivation, and this noun feeds into the interpretation. This is illustrated in (7). If the configuration contains an elided N(P) as in in (7a), the sentence is interpreted as in (7b)—i.e., the *the only* phrase singles out one boy from a group of boys.

(7) (a) This boy is the only ~~boy~~ who is nice. ~~boy~~ ➞ one
(b) This boy is the only boy who is nice.

Following Merchant (2014), deep ellipsis, on the other hand, involves an abstract null noun, which does not correspond to a specific noun but is only specified as [±ANIMATE] (see Saab [forthcoming] for a similar proposal). I provide further motivation for this proposal in section 4. In a context such as (8) where there is only a single boy in the comparison group, the interpretation corresponding to N(P) ellipsis in (8a) is infelicitous since the comparison set triggered by *the only* does not include any boys. Instead, the interpretation is as in (8b) where ellipsis is best understood as "the only person."

(8) Context: a group of women and one boy
The boy is the only one who is nice.
(a) #The boy is the only ~~boy~~ who is nice.
(b) The boy is the only $\emptyset_{[+ANIM]}$ who is nice.
This boy is the only person <animate/human entity> who is nice.

The two interpretations are available in German as well. A sentence such as (9a) can refer to either context given above (for the N(P) ellipsis situation an element indicating contrastive focus is necessary, e.g., a demonstrative, modifier of "boy" etc.). Thus, both structures in (9b) are available.

(9) (a)

Der	Bub	ist	der	Einzige,	der	nett	ist.
the.M.SG	boy	is	the.M.SG	only.SG	who.M.SG	nice	is

"The (this) boy is the only boy who is nice."
"The boy is the only person who is nice."

(b)

Der	Bub	ist	der	Einzige,	~~Bub~~/$\emptyset_{[+ANIM]}$	der . . .
the.M.SG	boy	is	the.M.SG	only.SG	~~boy~~/$\emptyset_{[+ANIM]}$	who.M.SG . . .

German nominal ellipsis raises an interesting question regarding agreement. As shown in (9b), agreement on the remnants is obligatory in both cases (no other feature combination

is possible). For an N(P)-ellipsis derivation, agreement is easily achieved via the elided noun. However, for the deep ellipsis configuration, something else needs to be at work to equip the ellipsis remnant with the necessary features. In the next section, we will see that associating the nominal deep anaphor with a (personal) pronoun is not sufficient for nominal ellipsis.

3.2 Mismatches in Nominal Ellipsis

In this section, we will consider agreement in predicate ellipsis constructions of the "the only" type with three kinds of trigger nouns in the antecedent—mismatching nouns, pluralia tantum nouns, and polite pronouns—in deep and surface ellipsis. The conclusion will be that the generalization in (10) holds.

(10) In German predicate constructions, formal agreement between the subject and the ellipsis remnant is only possible when the interpretation is compatible with N(P) ellipsis.

The first situation is given in Figure 1: the context group for the sentence includes a single girl who is dressed entirely in blue, and all other individuals are male and not dressed in blue. In this context, ellipsis cannot be interpreted as N(P) ellipsis (the girl is the only girl that is dressed in blue), but only as deep ellipsis (the girl is the only person that is dressed in blue).

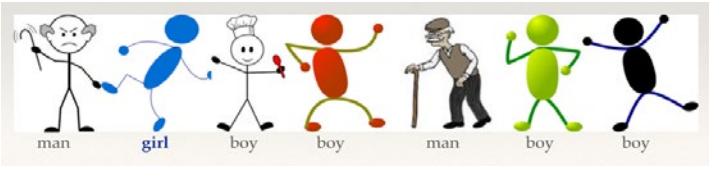


Figure 1. Deep ellipsis with animate mismatch noun

As shown in (11), in this context it is not only possible to use semantic agreement, (11a), but it is necessary; formal agreement, an option that is otherwise always available with mismatch nouns, is excluded, (11b).

(11)	(a)	Das	Mädchen	ist	die	Einzige,
		the.N.SG	girl	is	the.F.SG	only.SG
		die	blau	angezogen	ist.	
		who.F.SG	blue	dressed	is	

"The girl is the only one who is dressed in blue."

(b)	*Das	Mädchen	ist	das	Einzige,
	the.N.SG	girl	is	the.N.SG	only.SG
	das	blau	angezogen	ist.	
	who.N.SG	blue	dressed	is	

"The girl is the only one who is dressed in blue."

The agreement pattern changes if a context as in Figure 2 is considered where the group used as a comparison set for "the only" consists of only girls and only one girl, the second one, is dressed in blue.

Figure 2. N(P) ellipsis with animate mismatch noun

As shown in (12a), the formal agreement option is now the preferred option. Semantic agreement, (12b), is also still available, due to the entailment that in the context in Figure 2 the second girl is also the only person who is dressed in blue. Thus this situation is also compatible with a deep ellipsis configuration, however, the N(P) ellipsis interpretation is more informative and may therefore be preferred.

(12)	(a)	Das	zweite	Mädchen	ist	das	Einzige,
		the.N.SG	second	girl	is	the.N.SG	only.SG
		das	blau	angezogen	ist.		
		who.N.SG	blue	dressed	is		

"The second girl is the only one who is dressed in blue."

(b)	?	Das	zweite	Mädchen	ist	die	Einzige,
		the.N.SG	second	girl	is	the.F.SG	only.SG
		die	blau	angezogen	ist.		
		who.F.SG	blue	dressed	is		

"The second girl is the only one who is dressed in blue."

The effect that formal agreement disappears when the interpretation is not compatible with N(P) ellipsis (i.e., generalization in [10]) is also observable when the ellipsis antecedent contains an inanimate noun. The situation in Figure 3 describes a context in which waiter trainees need to set a table with all the items given. The items *Kerze* "candle," *Serviette* "napkin," *Gabel* "fork," *Vase* "vase," *Flasche* "bottle" are all feminine nouns in German. The numbers indicate how many trainees put the relevant item on the table, thus none of the trainees forgot to put the fork on the table.

Figure 3. Deep ellipsis with inanimate mismatch noun

In this situation, the remnants of (deep) ellipsis must occur with neuter agreement as in (13a), and it is not possible to realize formal agreement (i.e., feminine) matching the gender of *Gabel* "fork" (and all the other items in the context). In section 4, I will suggest that (13a), like the example in (11a) with *Mädchen*, involves semantic agreement and that neuter is the default realization of nominal elements lacking semantic gender (i.e. all [−ANIMATE] entities including events and actions).

(13) (a) Die Gabel ist das Einzige,
the.F.SG fork.F is the.N.SG only.SG
das/was niemand vergessen hat.
that.N.SG/what nobody forgotten has
"The fork is the only one/thing that nobody forgot."

(b) *Die Gabel ist die Einzige,
the.F.SG fork.F is the.F.SG only.SG
die niemand vergessen hat.
that.F.SG nobody forgotten has

Turning to an N(P) ellipsis context, consider the situation depicted in Figure 4. In this case, waiter trainees have to name different types of forks. A checkmark above a fork indicates that the trainees recognized the fork, whereas a cross mark shows that they could not name that type of fork.

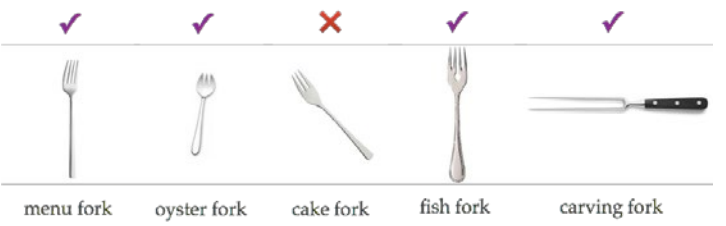


Figure 4. N(P) ellipsis with inanimate mismatch noun

In this context, formal feminine agreement as in (14a) is possible (and preferred) since the interpretation is compatible with an N(P) ellipsis interpretation. As before, semantic neuter agreement in (14b) is not excluded but marked.

(14)	(a)	Die	Kuchengabel	ist	die	Einzige,
		the.F.SG	cake.fork.F	is	the.F.SG	only.SG
		die	niemand	erkannt	hat.	
		that.F.SG	nobody	recognized	has	
		"The fork is the only one that nobody recognized."				

	(b)	?Die	Kuchengabel	ist	das	Einzige,
		the.F.SG	cake.fork.F	is	the.N.SG	only.SG
		das/was	niemand	erkannt	hat.	
		that.N.SG/what	nobody	recognized	has	

Another type of noun that can be described as involving a feature mismatch are pluralia tantum nouns like *Augengläser* "glasses" which are formally plural but can refer to a single item. In the situation in Figure 5, someone is looking for all the items displayed but he found only the glasses.

Figure 5. Deep ellipsis with pluralia tantum noun

Since there is only a single pair of glasses in the context, the sentence in (15) is not compatible with an N(P) ellipsis configuration. As in the other deep ellipsis cases above, formal agreement is impossible and only the default neuter singular form can be used on the ellipsis remnants. Note that the finite verb, on the other hand, obligatorily shows plural agreement in (15a).

(15)	(a)	Die	Augengläser	sind	das	Einzige,
		the.PL	glasses.PL	are.PL	the.N.SG	only.SG
		das/was	er	gefunden	hat.	
		that.N.SG/what	he	found	has	
		"The glasses are the only thing he found."				

(b) *Die Augengläser sind die Einzigen,
the.PL glasses.PL are.PL the.PL only.PL
die er gefunden hat
that.PL he found has
"The glasses are the only thing he found."

Once again the situation changes when the context leads to an N(P) ellipsis interpretation as in Figure 6, where an optometrist is looking for several pairs of glasses.

Figure 6. N(P) ellipsis with pluralia tantum noun

In this context, formal plural agreement, (16a), is the preferred option to refer to a specific pair of glasses and the default neuter version in (16b) is infelicitous and marked.

(16) (a) Die grünen Augengläser sind die Einzigen,
the.PL green.PL glasses.PL are.PL the.PL only.PL
die er gefunden hat.
that.PL he found has
"The green glasses are the only ones he found."

(b) ?Die grünen Augengläser sind das Einzige,
the.PL green.PL glasses.PL are.PL the.N.SG only.N.SG
das/was er gefunden hat.
that.N.SG/what he found has
"The green glasses are the only ones he found."

The last controller type is polite pronouns. As shown in (17), when referring to a single person, the polite pronoun must trigger singular agreement on the ellipsis remnant and plural agreement is only possible when addressing several people.

(17) (a) Sie sind der Einzige, der gelacht hat.
you.PL are.PL the.M.SG only.SG who.M.SG laughed has
"You (pol.) are the only one who laughed."

(b) Sie sind die Einzige, die gelacht hat.
you.PL are.PL the.F.SG only.SG who.F.SG laughed has
"You (pol.) are the only (female) one who laughed."

(c) Sie sind die Einzigen, die gelacht haben.
you.PL are.PL the.PL only.PL who.PL laughed have.PL
*"You (pol.) are the only one who laughed." (single addressee)
OK "You (pol.) are the only ones who laughed." (multiple addressees)

3.3 Summary

Table 1 summarizes the distribution of formal and semantic agreement in German. Gender mismatches cannot be tested in verb/T-agreement configurations, since verbs do not inflect for gender in German. Pluralia tantum nouns do not allow semantic agreement for referential pronouns. By definition, these nouns do not have singular forms, and since gender is only distinguished in the singular in German, pluralia tantum nouns are not specified for gender. I tentatively assume that the lack of gender is the reason for why referential pronouns associated with a DP antecedent headed by a pluralia tantum noun cannot realize singular agreement but instead use the other (formal) agreement option. In deep ellipsis contexts, on the other hand, formal agreement is not available (see the next section), and hence a default option kicks in which yields the neuter singular form.

	attributive	**predicate (T)**	**relative**	**personal pronoun**	$\emptyset_N$
mismatch noun	formal	N/A	formal	formal or semantic	semantic
pluralia tantum	formal	formal	formal	formal, semantic N/A (no gender)	semantic
polite pronoun	N/A	formal	N/A	formal, semantic N/A (no polite SG form)	semantic

Table 1. Formal and semantic agreement with different N controllers

The last row shows the agreement options for polite pronoun controllers. Pronouns generally do not occur with other elements in the noun phrase and thus agreement with attributive elements and relative pronouns cannot be tested. The only elements that may be considered modifiers of pronouns are affective adjectives (Wechsler and Hahm 2011) such as *Sie Armer*/*Arme!* "You.POL poor.M.SG/F.SG/*PL" (Poor you!). As indicated, the form used on the adjective reflects semantic agreement and formal agreement is impossible. However, it is not clear that such constructions involve a single DP structure in German. Adjectives must occur pre-nominally in German, but the word order in these PRON + ADJ examples cannot be changed (i.e., **Arme Sie!*). I therefore assume that these constructions are not single DPs but involve an elliptical appositive DP modifying the entire pronominal DP. Semantic agreement is then expected since these constructions fall under the $\emptyset_N$ category. Lastly, referential pronouns associated (bound by or co-referent)

with a polite pronoun antecedent can only show formal agreement since there is no honorific singular pronoun that could be used to refer back to a politely addressed participant.

The overall agreement pattern in German can thus be summarized as in Table 2 which will be the empirical basis for the account sketched in the next section.

attributive	**predicate (T)**	**relative**	**personal pronoun**	$\emptyset_N$
formal	formal	formal	formal or semantic	semantic

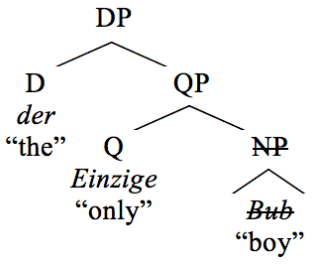
Table 2. Formal and semantic agreement in German

4. Deriving the Distribution of Formal and Semantic Agreement

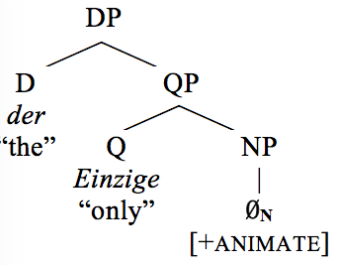
4.1 Ellipsis Structures and Agreement

Before providing an account of the distribution of formal and semantic agreement in Table 2, I lay out simplified structures for the two types of ellipsis. As illustrated in (18), the main difference is that in the N(P) ellipsis configuration the syntactic structure involves an actual noun which contributes the lexical, syntactic and semantic properties associated with that noun (except its phonological properties) to the remnant DP. In deep ellipsis, on the other hand, there is no actual noun but an abstract zero N head (see also Merchant [2014]) which is only equipped with the feature [±ANIMATE/HUMAN].

(18) (a) N(P) ellipsis



(b) Deep ellipsis



What I refer to as a zero noun in (18b) is often treated as a null pronoun (see among others Lobeck 1995, Kester 1996, Corver and van Koppen 2011, and Saab, forthcoming). Since pronouns cannot occur with determiners and modifiers (cf. **the only he*), but the null element in deep ellipsis does, such a null pronominal would have to be of a different nature than personal pronouns or argumental *pro*. Furthermore, as we have seen in German, the remnants of both N(P) ellipsis and deep ellipsis obligatorily agree, which goes against the observation made by Corver and van Koppen (2011), that the pronominal variant of ellipsis typically occurs without agreement of the remnant. Lastly, as Table 2 has shown, pronouns and the null element in deep ellipsis show different agreement

properties: clear cases of pronouns always allow formal agreement (or require it in case of relative pronouns), whereas deep ellipsis only allows semantic agreement.

I propose further that the agreement properties of $\emptyset_N$ reflect genuine agreement rather than simply a semantic property. There are two pieces of motivation for the claim that there is agreement in deep ellipsis contexts. First, as shown in (19) ([19a] is repeated from [11]), semantic agreement, which is the required form in a deep ellipsis context in (19a) (Figure 1 above), becomes unavailable when the antecedent DP does not c-command the deep ellipsis $\emptyset_N$, as is the case in the inverted order in (19b).

(19) (a)

Das	Mädchen	ist	die	Einzige,
the.N.SG	girl	is	the.F.SG	only.SG
die	blau	angezogen	ist.	
who.F.SG	blue	dressed	is	

"The girl is the only one who is dressed in blue."

(b)

?*Die	Einzige,	die	blau	angezogen	ist,
the.F.SG	only.SG	who.F.SG	blue	dressed	is
ist	das	Mädchen			
is	the.N.SG	girl			

"The only one who is dress in blue is the girl."

Second, following a similar argument provided in Corbett (2006, 233), there are languages that allow either formal or semantic agreement in deep ellipsis contexts. This is the case in Greek and possibly also in one variety of Czech. In these languages, there is a general preference for formal agreement, however, in exactly the deep ellipsis configurations, semantic agreement is allowed as well. Below I will suggest that the choice of agreement type is subject to a preference condition which favors semantic agreement in deep ellipsis contexts. However, if a language also has a preference condition for formal agreement (such as the *Agreement Marking Principle* in Wechsler [2011], Wechsler and Hahm [2011]), the tension between these two choices can be resolved by making available both options. In light of the cross-linguistic distribution of agreement in deep ellipsis contexts, relying solely on semantic properties is insufficient, but a uniform account is possible if the constructions in Tables 1, 2 all involve agreement.

4.2 Dual Feature System

The account of agreement mismatches I propose follows feature systems in which noun phrases involve two sets of φ-features (see Pollard and Sag 1994; Wechsler and Zlatić 2000, 2003; Wechsler 2011; Wechsler and Hahm 2011; Wurmbrand 2012; Smith 2015). The two feature types co-exist in syntax but are split at Spell-Out and sent to different interfaces. The specific approach I adopt is that a DP/NP has formal *u*φ-features which

feed (only) into PF and carry the values realized in morphology; and semantic *i***φ**-features which feed (only) into LF and carry the values interpreted in semantics. In contrast to DPs/NPs, φ-features on adjectives and verbs/T do not express semantic information on APs and T; these elements therefore only carry *u***φ**-features.

Syntactic agreement, I assume, is established via the operation Agree, and, in principle, an agreement target can copy either the values of the *u***φ** or the ones of the *i***φ**-features from the controller. If the *u***φ**-features of the controller are used, the target shows formal agreement; if the *i***φ**-features of the controller are used, the target shows semantic agreement. However, both types of agreement can be established in syntax (I continue to use the descriptive term "semantic" agreement, even though this relation is treated as a syntactic relation here).[2]

If both formal and semantic agreement can be established syntactically, the obvious question is how to restrict the system. Consider again the distribution of formal and semantic agreement in German as given in Table 3. If we add the feature types of the target elements, we see that there is a clear match. APs and T only require *u***φ**-feature values (**φ**-features are not interpreted on AP and T, only on the agreeing DP), and these elements only show formal agreement. Pronouns, being independent DPs, require both *u***φ** values and *i***φ** values, and pronominal targets can show either formal or semantic agreement. Lastly, the anaphoric $\emptyset_N$ in ellipsis is only visible semantically (it is phonetically zero and not visible at PF), hence it only requires *i***φ** values, and these elements only show semantic agreement.

	attributive	**predicate (T)**	**personal pronoun**	$\emptyset_N$
German	formal	formal	formal or semantic	semantic
Features of target	*u***φ**	*u***φ**	*u***φ** and *i***φ**	*i***φ**

Table 3. Target feature types

To implement the generalization observable in Table 3, but to also leave room for variation (see Wurmbrand 2016b), I assume that the choice between formal (*u*φ values of the controller) and semantic (*i*φ values of the controller) is subject to the preference condition in (20). The match condition in (20) yields, as a default, formal agreement for target elements with only formal *u*φ-features, semantic agreement for targets with only semantic *i*φ-features, and either form of agreement for targets with both types of features. As laid out in (20), A and B undergo Agree, which is subject to c-command and involves

2 Note that this does not mean that agreement has to apply in syntax. The claim is only that both formal and semantic agreement can be triggered in syntax. I assume that post-syntactic agreement is also an option. However, if agreement takes place at PF, only the formal features are available and only formal agreement will be triggered (see Bhatt and Walkow [2013], Wurmbrand [2012, 2016a] for evidence for PF-agreement).

establishing a link between the φ-features of A and B, if at least one of the feature sets is unvalued. At that point, the controller choice condition in (20) comes into play and temporarily inactivates the non-matching feature type on the controller (indicated as grey features in [20]). Feature copying then applies between B and the chosen feature of A. Feature inactivation is temporary and defined for each dependency separately. This is important for cases where one and the same controller triggers different types of agreement on different targets (e.g., T-agreement vs. agreement with pronouns).

(20) *Match preference for feature type of controller:*
Match the feature type of the target with the feature type of the controller.

$A_{controller}$ [*y*φ: val,	***x*φ**: val]	⟷ $_{Agree}B_{target}$ [*x*φ: ___]	Agree
$A_{controller}$ [*y*φ: val,	***x*φ**: val]	⇢ B_{target} [*x*φ: val]	Controller choice

As for German, the match condition in (20) is all that is required since, as shown in Table 3, the preferred feature type is exactly the feature type triggering agreement. This is, however, not the case in all languages. Interesting cross-linguistic variation can be found in the distribution of agreement on predicative adjectives and the agreement properties of polite pronouns (see also Comrie 1975; Corbett 1983, 2000, 2006; Hahm 2010; Wechsler 2011; Wechsler and Hahm 2011, among others). In Wurmbrand (2016b), I suggest that the more nuanced differences found cross-linguistically are attributed to the specific feature specifications of the different types of nominal elements, together with the concept that the *i*φ/*u*φ preference yielded by (20) can be overturned if the less preferred feature type constitutes a better source of features (similar to Wechsler [2011] and Wechsler and Hahm's [2011] *Agreement Marking Principle*).

As an example, in many languages, predicate APs show formal agreement with controllers headed by mismatch nouns, but semantic agreement when the controller is a polite pronoun, which is illustrated in (21) for Czech (see the references above).

(21) (a) To děvče je milé / *milá
this.N.SG girl.N.SG is nice.N.SG / *nice.F.SG
"This girl is nice." (Ivona Kučerová, pers. comm.)

(b) Vy jste čestný / čestná
you.2.PL be.2.PL honest.M.SG / honest.F.SG
"You (pol.) are honest." (Petr Biskup, pers. comm.)

In both cases, the match condition in (20) would favor formal agreement since APs only have *u*φ-features. This is what we find in (21a), but not in (21b), and I propose that when the controller is a polite pronoun, the *i*φ-features are a better match for the AP's

*u*φ-features due to a deficiency in the *u*φ-feature structure of polite pronouns.[3] Polite pronouns do not show morphological gender distinctions but they do involve person marking [3 (German), 2 (other languages)]. Assuming a markedness filter which prevents the combination of participant and gender features (cf. Calabrese 2011), the *u*φ-feature structure of a polite pronoun in Czech would be [2.PL]. The semantic features, on the other hand, do not include specific person features but rather a semantic property ADDRESSEE (which is then realized as either 2nd or 3rd person morphologically, depending on the language). Since markedness then does not apply, the *i*φ-feature structure of a polite pronoun is [ADDRESSEE (POLITE).SG/PL.M/F], depending on the gender and number of the addressee. Since AP targets require a gender value, we can now see why the *i*φ-features of polite pronouns are a better match than the *u*φ-features—the former contain a gender value, whereas the latter don't. I assume that this overrides the preference given by (20) and hence yields the difference in agreement for predicative APs in (21).[4]

As for the distribution of agreement in deep ellipsis contexts, I cannot review the various data and options here but only point out the generalizations I have encountered so far in testing agreement in ellipsis contexts (some details can be found in Wurmbrand 2016b). First, predicative DPs/NPs always allow (often require) semantic agreement, independent of the agreement properties in other constructions. Second, if a predicative DP/NP allows formal agreement with a particular controller, that controller (obligatorily) triggers formal agreement on predicative AP targets. While each language of course deserves its own special attention, these generalizations can nevertheless be taken as support for the feature system proposed here and the match condition in (20).

5. Conclusions

This paper has surveyed the distribution of formal and semantic agreement in German for three types of trigger nouns (gender mismatch nouns like *Mädchen* "girl," pluralia tantum nouns and polite pronouns) in four syntactic contexts (attributive, predicate/T, pronouns, and nominal ellipsis). The findings have led to the refined Agreement Hierarchy in (22).

3 An alternative (Jonathan Bobaljik, pers. comm.) would be to assume that predicate AP contexts are sometimes hidden NP/DP constructions involving a silent noun which undergoes semantic agreement like in deep ellipsis contexts. This is suggested in Wurmbrand (2016b) for the different agreement options arising in Russian predicative APs with long vs. short form adjectives. It remains to be seen whether a hidden N structure could be the source of all cases with semantic agreement.

4 This account is similar in spirit to the proposal in Wechsler (2011) and Wechsler and Hahm (2011), where it is proposed that polite pronouns are not specified for Concord features but do involve a plural Index feature. This feature structure is somewhat unintuitive since the Index feature represents semantic properties, however, polite pronouns are not plural semantically. The current proposal provides a more transparent morphology-semantics mapping.

(22) F ⇠attributive — predicate — relative — personal PRON — $\emptyset_N$/predicate DP ⇢ S

We have seen that the choice between formal and semantic agreement depends on the properties of the target (formulated as a preference condition for the feature type of the target to match the feature type of the controller), the feature structure of the trigger (e.g., underspecification, markedness effects), and possibly other language specific preferences such as a general preference for formal agreement. I have proposed a dual feature system that captures the basic tendency of the Agreement Hierarchy in (22) and leaves room for encoding differences attested across languages.

Works Cited

Bhatt, Rajesh and Martin Walkow. 2013. "Locating Agreement in Grammar: An Argument from Agreement in Conjunctions." *Natural Language and Linguistic Theory* 31 (4): 951–1013.

Calabrese, Andrea. 2011. "Investigations on Markedness, Syncretism and Zero Exponence in Morphology." *Morphology* 21 (2): 283–325.

Comrie, Bernard. 1975. "Polite Plurals and Predicate Agreement." *Language* 51 (2): 406–18.

Corbett, Greville G. 1979. "The Agreement Hierarchy." *Journal of Linguistics* 15 (2): 203–24.

Corbett, Greville G. 1983. *Hierarchies, Targets and Controllers: Agreement Patterns in Slavic*. London: Croom Helm.

Corbett, Greville G. 2000. *Number*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2006. *Agreement*. Cambridge: Cambridge University Press.

Corver, Norbert, and Marjo van Koppen. 2011. "NP-ellipsis with Adjectival Remnants: A Micro-comparative Perspective." *Natural Language & Linguistic Theory* 29: 371–421.

Hahm, Hyun-Jong. 2010. "A Cross-linguistic Study of Syntactic and Semantic Agreement: Polite Plural Pronouns and Other Issues." PhD diss., University of Texas, Austin, TX.

Hankamer, Jorge, and Ivan Sag. 1976. "Deep and Surface Anaphora." *Linguistic Inquiry* 7 (3): 391–426.

Kester, Ellen-Petra. 1996. "Adjectival Inflection and the Licensing of Empty Categories in DP." *Journal of Linguistics* 32: 57–78.

Lobeck, Anne. 1995. *Ellipsis: Functional Heads, Licensing, and Identification*. Oxford: Oxford University Press.

Merchant, Jason. 2014. "Gender Mismatches under Nominal Ellipsis." *Lingua* 151: 9–32.

Merchant, Jason. Forthcoming. "Ellipsis: A Survey of Analytical Approaches." In *Handbook of Ellipsis*, edited by Jeroen van Craenenbroeck and Tanja Temmerman. Oxford: Oxford University Press.

Pollard, Carl, and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: CSLI / University of Chicago Press.

Saab, Andrés. Forthcoming. "Nominal Ellipses." In *Handbook of Ellipsis*, edited by Jeroen van Craenenbroeck and Tanja Temmerman. Oxford: Oxford University Press.

Smith, Peter. 2015. "Feature Mismatches: Consequences for Syntax, Morphology, and Semantics." PhD diss., University of Connecticut, Storrs, CT.

Wechsler, Stephen. 2011. "Mixed Agreement, the Person Feature, and the Index/Concord Distinction." *Natural Language & Linguistic Theory* 29: 999–1031.

Wechsler, Stephen, and Hyun-Jong Hahm. 2011. "Polite Plurals and Adjective Agreement." *Morphology* 21: 247–81.

Wechsler, Stephen, and Larisa Zlatić. 2000. "A Theory of Agreement and Its Application to Serbo-Croatian." *Language* 76: 799–832.

Wechsler, Stephen, and Larisa Zlatić. 2003. *The Many Faces of Agreement*. Stanford, CA: CSLI Publications.

Wurmbrand, Susi. 2012. "Agreement: Looking Up or Looking Down?" Lecture given at Agreement Seminar, MIT, Cambridge, MA, March 12.

Wurmbrand, Susi. 2016a. "Agreement in Nominal Ellipsis—Consequences for the Agreement Hierarchy and the Direction of Agree." Talk given at an Agreement Workshop, Frankfurt, Germany, July 14.

Wurmbrand, Susi. 2016b. "Girls, Glasses, and You—The Distribution of Formal vs. Semantic Agreement." Talk given at Generative Grammatik des Südens (GGS), Leipzig, Germany, October 20–22.

A Number Constraint of Czech Quantified Nominals

Ludmila Veselovská

Palacký University, Olomouc, Czech Republic

ludmila.veselovska@upol.cz

Abstract: The paper summarizes data related to the occurrences of two Czech quantifiers—*mnoho* "a lot" and *málo* "few / a little" in oblique case contexts—as they are attested in the Czech synchronic corpus Syn2015. Assuming the category-based Case theory as in Pesetsky (2013) and the logic of his discussion of Russian quantifiers, the study argues in favour of the same analysis for the Q category in Czech. Moreover, concentrating on the constraint discovered here, which requires that in Czech the quantified nominal complex following the Qs be countable [+PLURAL] in oblique contexts, the paper also proposes a specific analysis of the so-called adverbial (or oblique) inflection of the Czech Qs.

Keywords: Czech quantifier; case; case-assignment; countable

1. Morphosyntactic Properties of Czech Quantifying Expressions

First let us briefly consider what is special about the properties of Czech quantifiers, compared with standard nouns N and Group nouns $N_{[Q]}$ followed by a postnominal genitive DP. Based on the discussion in Veselovská (2001) and Jiranová (2008), the Table in (2) shows a summary of the typical morphosyntactic characteristics of Czech quantifying expressions, with a division into three groups:

(1) Taxonomy of Quantifying elements (QE) in Czech:

(a) $Q_{[N]}$: a group of QE which show the most "nominal" characteristics,
(b) Q: Quantifiers (including Numerals "5 & up") and
(c) $Q_{[\varphi]}$: agreeing Quantifiers which show the most "adjectival" characteristics.

Some examples of purely nominal and purely adjectival Czech quantifiers are added at the top and the bottom of the table, and are to be compared with the characteristics of the $Q_{[\varphi]}$. The bold frames in the middle of Table (2) mark the two Czech quantifying elements—*mnoho* "a lot" and *málo* "few / a little"—which are going to be discussed in the following sections.

(2) **Morphosyntactic characteristics of Czech quantifying elements**
*non-interpretable plural (still reflected by predicate and AP agreement),
**-*o*/-*a* or Ø/-*i* variants only

Category	Czech quantificational expression	English translation	features of the QUANTIFYING ELEMENT			
			Grading	φ of the Q		
				Gender	Number	Case
$N_{[Q]}$	SKUPINA	group	-	F	S/P	N
	polovina, dvojice	a half / a couple	-	F	S/P	N
	tucet, kopa	a dozen	-	M/F	S/P	N
$Q_{[N]}$	milion/miliarda	a million / a billion	-	M	S/P	N
	miliarda	a billion	-	F	S/P	N
	sto/tisíc	100 / 1,000	-	N/M	S/P	N/Ø
	SPOUSTA	plenty	-	F	S (P*)	N
Q	**MÁLO**	few/little	+	N/Ø	Ø	N/Ø
	MNOHO	a lot of	+	Ø	Ø	**
	několik	several	-	Ø	Ø	**
	hodně/půl	a lot / half	-	Ø	Ø	Ø
	PĚT (5 & up)	5, 6, 7 & up	-	Ø	Ø	**
$Q_{[\varphi]}$	TŘI, ČTYŘI	3, 4	-	Ø	Ø	D
	dva	2	-	φN	Ø	D
	jeden	one (a/some)	-	φN	agree	D
	VŠICHNI	all	-	φN	agree	D
$A_{[Q]}$	mnohý etc.	plentiful	-	φN	agree	A
	ČTVRTÝ & up	4th & up	-	φN	agree	A

The properties of the quantifying elements which are marked on the left of Table (2) concern the features of the quantifying expressions themselves: e.g., their ability to be graded, and their definable paradigmatic φ feature content: marking that is available for a specific Gender, Number, and Case (N: nominal, D: pronominal, A: adjectival agreement). The columns in the middle field of Table (2), to the right of the English translations, show the ability of the quantifying element

- to license agreement with secondary predicates,
- to be relativized (using an agreeing relative pronoun), and
- to be an antecedent of a personal pronoun.

AP secondary predicate	Relativization	Antecedent to a pronoun	ADV pre Q	Q requires min. GEN cl.	QUANTIFIED NOMINAL (qN)			AGREEMENTS	
					Case of the qN		± Count NOM+ACC/OBL	AP premodifiers	Predicate
					NOM/ACC context	Oblique			
+	+	+	-	-	GEN	GEN	+	**	φQ
+	+	+	+	-	GEN	GEN	±	**	φQ
+	+	+	+	-	GEN	GEN	+	**	φQ
-	+	+	+	+	GEN	GEN	+	**	φQ
-	+	+	+	+	GEN	GEN	+	**	φQ
-	±	-	+	+	GEN	GEN/OBL	+	Q	Ø/φQ
+	±	-	+	+	GEN	GEN	±	**	Ø/φQ
-	-	-	+	+	GEN	OBL	±	N	Ø
-	-	-	+	+	GEN	OBL		N	Ø
-	-	-	-	+	GEN	OBL	+	N	Ø
-	-	-	+	+	GEN	OBL	±	N	Ø
-	-	-	+	+	GEN	OBL	+	N	Ø
-	-	-	+	-	NOM/ACC	OBL	+	N	φN
-		-	+	-	NOM/ACC	OBL	+	N	φN
-	-	-	+	-	NOM/ACC	OBL	+	N	φN
-	-	-	+	-	NOM/ACC	OBL	±	N	φN
-	-	-	-	-	NOM/ACC	OBL	±	N	φN
-	-	-	-	-	NOM/ACC	OBL	±	N	φN

Table (2) also shows whether the quantifying expression is pre-modified by an adverbial or adjective and whether it obligatorily subcategorizes for a genitive complement, i.e., whether it has to be complemented by at least a genitive clitic.

The columns toward the right of Table (2) concentrate on the properties of the quantified nominal complex (qN) which follows the Q, stating the case which is marked on the qN in

- nominative (NOM) and accusative (ACC) contexts,
- in the other case (Oblique) contexts, and
- it provides information about the marking of a possible Number feature on qN.

And lastly, the rightmost columns in Table (2) show the agreement, reflecting the features of the quantifying expressions: the agreement appearing on the pre-modifying adjectivals and on the predicate, when the quantified complex is in the position of a subject.

2. Heterogeneous and Homogenous Pattern of the Czech 5 & Up Numerals and Some Quantifiers

In Table (2) the two Czech quantifying elements *mnoho* "a lot" and *málo* "few / a little" belong to the category labelled here as "Q," which includes also the Czech Cardinals 5 & up, e.g., *pět* "five." The two examples in (3) and (4) demonstrate the case and agreement patterns of these Czech Qs in two distinct contexts. First, in (3a) the Qs are followed by a countable [+PLURAL] Noun and in (3b) by a non-countable (mass) Noun. We can see that the form of the Czech Q is the same with countable and non-countable qN, i.e., *mnoho* is the equivalent of both English *much* and *many* and *málo* of both *few* and *little*.

The examples in (3) are illustrated in the form which we can find in NOM/ACC ("structural") case contexts.[1] Notice that the only element in (3) which can be marked by the required NOM/ACC is the Q itself. The rest of the qN complex is expressed in a so-called **partitive GEN**, which in Czech is phonetically identical to any other GEN, such as a GEN selected by a Preposition, Verb or Noun.

(3) **NOM/ ACC context**: Heterogeneous Pattern

(a)	**Mnoh-o / mál-o /**	**pět-Ø**	**žlutých brouků**	**přiletěl-o**
	many$_{\textbf{NOM?}}$ / few$_{\textbf{NOM?}}$ /	five$_{\textbf{NOM?}}$	[yellow bugs]$_{\textbf{MP.GEN}}$	flew$_{\textbf{3SN}}$ in

"Many / few / five yellow bugs flew in."

1 In addition to an ACC selected by a verb, the same pattern appears after prepositions selecting ACC—as demonstrated in (6).

(b) **Mnoh-o / mál-o / *pět-Ø žlutého oleje přitekl-o**
much$_{\textbf{NOM?}}$ / little$_{\textbf{NOM?}}$ / *five [yellow oil]$_{\textbf{MS.GEN}}$ flowed$_{\textbf{3SN}}$ in
"Much / little / (*five) yellow oil flowed in."

The examples in (3) also show that if such a quantified complex appears in the subject position, the finite verb (and participle) reflects a default (=3SN) subject-predicate agreement features (ignoring thus the features of the qN). Because of the "dual" case pattern appearing inside the complex headed by Q—one part is marked by NOM/ACC and the other by GEN—this kind of case and agreement pattern is labelled as **Heterogeneous**.[2]

The examples in (4a) illustrate the same Qs in a context distinct from NOM and ACC, i.e., in the Oblique contexts (lexical case contexts), here following the preposition *s* "with" that requires Instrumental. Notice that there is no GEN inside the complex headed by the Q. The quantified nominal phrase qN (including its adjectival modifiers) is marked by the required Oblique (here Instrumental), and it is headed by a countable N which appears in [+PLURAL]. The Q thus acquires morphology distinct from the NOM/ ACC demonstrated in (3). The quantifier forms with the suffix *-a*/*-i* are traditionally labelled as oblique, and it is the same in all non-NOM/ACC case contexts. Because of the apparently uniform case-marking throughout the quantified complex, the kind of pattern in (4a) has been labelled as the **Homogenous** Pattern.[3]

(4) **Oblique context** (Instrumental): Homogenous Patterns[4]

(a) **Bojovali s mnoh-a / ?mál-o/*-a / pět-i (žlutými) brouky**
fought with [+INS] many$_{\textbf{OBL?}}$ / *few$_{\text{OBL?}}$ / five$_{\text{OBL?}}$ [yellow bugs]$_{\text{MP.INS}}$
"They fought with many / *few / five yellow bugs."

2 The terminology is taken from Baby (1985; 1987)and Franks (1994; 1995), who introduced similar data for other Slavic languages. For a detailed discussion, see also the earlier work of Pesetsky (1982) and for Czech, see Veselovská (2001).

3 A more detailed description of the variety of Czech "homogenous" agreeing patterns can be found in, e.g., Caha (2016).

4 The acceptability of the data here is marked according to the writer's intuitions and the data found in the Czech National Corpus Syn2015. In Google, however, some examples of fossilized *mál-o* "few" in oblique context with countable Nouns (esp. with *lidé* "people") can be found (*s málo lidma*$_{\text{INSTR}}$ "with few people," *o málo lidech* "about few people$_{\text{LOC}}$"). There is also minimally one example of *mál-o* "few" with non-countable on Google, namely *s málo olejem* "with few oil$_{\text{INSTR}}$." In all those examples the fossilized quantifiers (i.e., with the ending *-o*, not with the ending *-a*) are followed by an oblique case-marked nominal complex.

The two Q patterns illustrated in (3) and (4) are well known to exist in most Slavic languages and have been the topic of many research papers. In this text I will accept the overall framework introduced for Russian in Pesetsky (2013) for use also with the Czech data. In my study, however, I will concentrate on observations which to my knowledge have not been discussed yet, namely a curious number restriction which is attested in the qN complex in Oblique contexts. Notice that in (4b) no Q is acceptable with the non-countable qN *olej* "oil." This is not surprising with Numerals higher than 1, but given that in (3b) there is no restriction on Number of the qN, the ungrammaticality of the bold framed (4b) with either of the two available endings is puzzling.

(b)	**Bojovali s**	***mnoh-o/*-a /**	***mál-o/*-a /**	***pět-i**	**(žlutým) olejem**
	fought with [+INS]	plenty$_{OBL?}$ /	*little$_{OBL?}$ /	*five	[yellow oil]$_{MS.INS}$
	"They fought with plenty of / little yellow oil."				

In the following, I will provide corpus data which confirm that the observation that the Qs *mnoho* "a lot" and *málo* "few / a little" cannot be combined with non-countable [-PL] in oblique case contexts. In addition to the data search, I will provide an analysis which explains such a restriction. This explanation will comprise a claim concerning the character of the so-called Oblique case morphology with the Qs.

3. The So-Called Adverbial Paradigm of Q

Standard nominal paradigms in Czech have theoretically seven morphological case forms in both singular and plural. However, the many syncretisms reduce the 7 into 3–5 distinct endings. The Czech Qs, on the other hand, appear in only two forms, and therefore the paradigm of the Qs is traditionally labelled as "adverbial" although no reason is provided to why adverbs should have any case paradigm at all.[5] The specific paradigm of Slavic QEs historically developed from the grammaticalized nominal paradigms of feminine and neuter patterns.[6]

The two Qs discussed here follow the pattern developed from the neuter paradigm. The very standard Czech neuter paradigm follows the pattern of, e.g., *měst-o* "city" given in the left column of (5a). The example shows that the *-o* ending is the marking used in SG.NOM/ACC and the *-a* inflection belongs to either the GEN singular or NOM plural.

5 The normative *Dictionary of Czech Language* takes them for either nouns, numerals or adverbs (see http://ssjc.ujc.cas.cz/ or *Jazyková poradna Ústavu pro jazyk český AV ČR* (http://old.ujc.avcr.cz/jazykova-poradna/zajimave-dotazy/dotaz-tydne-2009.html).

6 In a generative framework the proposal is described as the change of the case assigner from a lexical into a functional head as, e.g., in Miechowicz-Mathiasen (2014). Using data from Polish, the author distinguishes the structural GEN assigned by Q, which is a functional head that has come into existence by a process of "numeralization" from a lexical GEN assigned by N, which is a lexical head.

With the Qs *mnoho* "a lot" and *málo* "few / a little," the *-o* ending also appears in NOM/ACC but the *-a* suffix is found in all oblique contexts.

A similar situation can be observed in the inflection of the cardinals 5 & up, as demonstrated in (5b). Diachronically they developed from the feminine paradigm *kost* "bone" given in the left column of (5b). This paradigm has a zero ending in SG.NOM/ACC, and an *-i* suffix in both GEN singular and NOM plural. With the Czech 5 & up cardinals the *-i* suffix gets generalized for all oblique contexts.

(5)

	SG.NOM/ACC	**SG.GEN/PL.NOM**		NOM/ACC	**OBL**
(a)	*měst-o / Ø*	*měst – a*	.....	*mnoh-o*	*mnoh - a*
(b)	*kost-Ø*	*kost – i*		*pět-Ø*	*pět - i*

The syncretism between GEN.SG and NOM.PL holds in Czech for all feminines, many of the neuters, and some few masculine nominal paradigms. Its systematic nature (attested moreover also in Latin and several not only Indo-European languages) has been the reason for attempts to find a systematic explanation. Some proposals assume that the syncretism is an accidental result of phonological processes, others suggest that it is a syncretism resulting from "polarity," i.e., allowed by the contradictory meaning of the involved categories.[7]

Using Czech data, the syncretism is discussed in detail in Caha (2016) in the framework of Nano-syntax. In this framework the syncretism is a result of the suffix spelling out GEN.SG and NOM.PL heads, which must appear structurally in very close proximity. Therefore the author proposes that plurals (at least in languages with the syncretism) result from bi-nominal recursive structures containing a (covert or overt) head with the meaning of "group." This abstract functional head requires a genitive complement which in turn agrees with this head. The plural morpheme then corresponds to the portmanteau realization of the genitive plus agreement.

With no detailed recourse to the nano-syntax of this inflection, I am nonetheless going to show in the following section that the oblique morpheme indeed represents either genitive or plural.

4. Searching for the Countability Constraint in Syn2015

To support the existence of the constraint requiring a countable characteristic of the nominal complex following the Qs in oblique contexts, i.e., the unacceptability of the Czech structures in (3) and (4), I searched in the Czech corpus for the forms of Qs

7 See Béjar and Hall (1999), Baerman et al. (2002), Wunderlich (2012), Lahne (2007), and Manzini and Savoia (2011).

mnoh-o/-a "a lot" and *mál-o/-a* "few / a little" following prepositions. The numbers provided in this section are taken from the Synchronic Representative Corpus (Syn2010).[8]

From the collected examples I manually excluded the examples of the Qs following preposition that select ACC, which systematically appears in the Heterogeneous pattern and which does not show any restriction on number—as we can see in (6), exactly as in (3), i.e., acceptable with both countable qN in (a) and uncountable qN in (b).

(6) An ACC context following preposition

(a)	Stěžoval si	na	mnoh-o / mál-o /	pět-Ø	žlutých brouků
	complained REFL	about	many / few$_{\textbf{NOM?}}$ /	five$_{\textbf{NOM?}}$	[yellow bugs]$_{\text{MP.}\textbf{GEN}}$
	"He complained about a lot / few / five yellow bugs."				

(b)	Stěžoval si	na	mnoh-o / mál-o /	*pět-Ø	žlutého oleje
	complained REFL	about	many / few$_{\textbf{NOM?}}$ /	five$_{\textbf{NOM?}}$	[yellow oil]$_{\text{MS.}\textbf{GEN}}$
	"He complained about a lot / little / *five yellow oil."				

Apart from the Q+qN combinations following prepositions I also searched for the Q+qN combinations with the Qs containing the oblique inflection *-a* and manually selected those examples where the structure was related to the verb selecting oblique. The results of the search are presented below.

5. Countability Constraint on Oblique *mnoha* "a lot$_{\text{OBL}}$" + [NP]

As demonstrated above in the contrast between (4a) and (4b), the Q *mnoh-a* "a lot" is *not* unacceptable in every oblique context. It appears to be ungrammatical only in those contexts where the following qN is uncountable.[9]

The next table shows the number of examples of the Q *mnoh-a* "a lot$_{\text{OBL}}$" in a variety of oblique case contexts following prepositions as they were found in the Syn2015 corpus.[10]

8 The corpus is a part of the Czech national corpus, and it has 121,667,413 positions—it is therefore a quite reliable source of modern Czech data. I am indebted to Monika Pitnerová for her help with the corpus search. Without her the data would not be complete.

9 In other words, while in NOM/ ACC form the Czech *mnoho* "a lot" may be translated like both "much/many," in oblique contexts (i.e., with the inflection *mnoh-a*) the reading "much" is not available.

10 The search was done concentrating on the prepositional obliques, but similar examples could be formed with verbs selecting oblique noun phrases, as the example below shows.

(7) The variety of Oblique contexts for the Q *mnoh-a* "a lot$_{OBL}$" after P

Case	Tokens	.....	Most frequent P		Tokens
GENITIVE total	1,729		from/of	*z* + Q + N_{GEN}	1,174
DATIVE total	155		to	*k* + Q + N_{DAT}	128
LOCATIVE total	4,277		in	*v*(*e*) + Q + N_{LOC}	2,798
INSTRUMENTAL total	1,154		with	*s*(*e*) + Q + N_{INS}	601
total	7,315				

All the above 7,315 examples were checked manually and all following nominals were both **countable** and [+**PL**].

The examples in (8) show some of the thousands of oblique forms of the Qs found in the corpus: a DAT assigned by a verb in (a), a LOC assigned by a preposition in (b) and also a post-nominal GEN in (c). In line with the generalization formulated here, all the examples found with the Q *mnoh-a* "a lot$_{OBL}$" were **countable** and [+**PL**].

(8) (a) za války pomohl mnoha Čechům
during war helped [+DAT] many Czechs$_{PL.DAT}$
"during the war he helped many Czechs"

(b) v mnoha případech stát reagoval opožděně
in [+LOC] many cases$_{PL.LOC}$ state reacted late
"in many cases the state reacted late"

(c) absorbuje vlivy mnoha kultur
absorbs influences [+GEN] many cultures$_{PL.GEN}$
"it absorbs the influences of many cultures"

The corpus search thus confirms the existence of a constraint which in Czech requires that the quantified nominal complex following the two Qs in oblique contexts is countable [+PLURAL].

I propose that this constraint can be explained using a small modification of the derivation of the structures containing Qs as in Pesetsky (2013). The steps are described in (9) in the paragraphs (a)–(i) and illustrated for scheme (10), which illustrates examples (11) = (4).

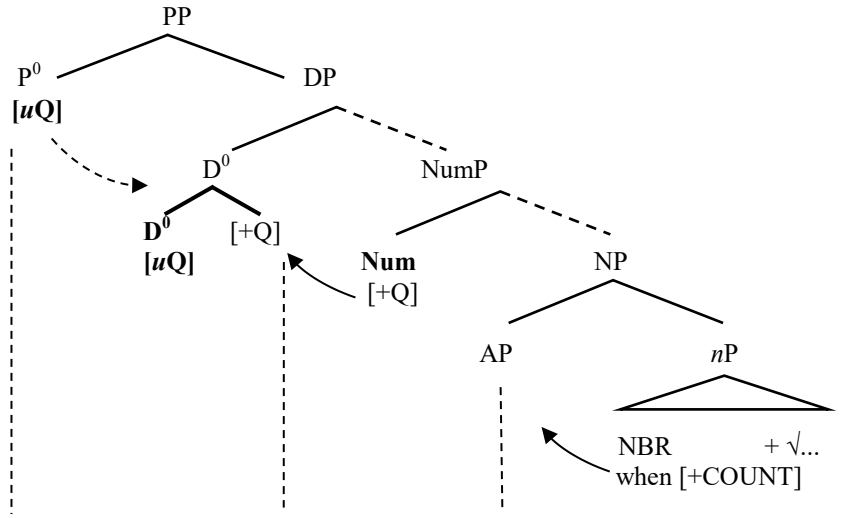
(i) To pomohlo mnoh-a našim lidem / *naší energii
it helped REFL[+DAT] many$_{OBL}$ our people$_{PL.DAT}$ / *our energy$_{SG.DAT}$
"It helped many of our people."

(9) **Derivation of the structures with Qs** (cf. Pesetsky 2013)

(a) Each nominal category is assigned a "primeval" GEN by some nominal categorial head. I assume the relevant categorial head is related to the Number head (NBR in [10]), which in Czech is merged very low and forms a part of the noun stem.
(b) After the merge of NBR, the projected N complex carries the primeval GEN which gets spread to its modifiers via NP-internal agreement.
(c) To project a structure with a Q, some Q (e.g., *mnoho/málo/pět* "a lot / little-few/ five" [5 & up]) merges in the high Num head position.
(d) Then an abstract of overt D head merges, which has the [*un*Q] feature.
(e) To satisfy the [*un*Q] feature of D, the closest [Q] (i.e., the Num) moves to D. It "undermerges" (i.e., right adjoins) to D and becomes its complement.
(f) The D category assigns a D-case (NOM) to its complement, i.e., to [Q]/Num.

In standard situations the D complement is constituted by the following (quantified) NP. In these structures with the undermerged Q, however, the complement of D is the right adjoined Q itself and therefore it is the Q that receives the D-case.

(10)



(11)

(a)	*s*	*mnoh-a* /	*?mál-o/*-a*	(*žlutými*) *brouky*
(b)	*s*	**mnoh-a* /	**mál-o/*-a*	(*žlutým*) *olejem*
	with	a lot of /	few-little	yellow bugs/oil

In a NOM context the NP retains Pesetsky's "primeval GEN." Although the whole nominal complex can later on appear in some other (later) case assigning contexts, the phase border makes the lower NP domain inaccessible for any other case marking.

In ACC contexts I accept Pesetsky's (2013) proposal that the V case (=ACC) is assigned (in Czech) only to a disjunction of [+FEM/+ANIM/+PRON/not -SG]. Because the Q is none of the above, the NP retains the primeval GEN which gets freezed by a phase border as in (g) above.

In Oblique contexts:

- P merges with DP/QP, which checks the [Q] feature,
- P assigns P-case to the DP/QP,
- Pesetsky's "One suffix rule" allows only the last Case (=P case) to be realized.

The above derivation (taken from Pesetsky [2013]) explains several points:

- Why the partitive (primeval) GEN disappears in Oblique contexts. It does so because P-Case overrides it.
- Why Qs cannot appear without the GEN cl.: This is because the [+Q] of D requires a complement.
- Why Subject-Verb agreement can only be default (3SN) with QP subjects: This is because the φ of D are deficient, given that they make the features of the lower domain inaccessible (they are blocked by the undermerge of Q).

However, the description of the derivation above, which follows Pesetsky's proposals, does not explain why the Qs *mnoho*/*málo* "a lot / few-little" require [+COUNT/+PL] in Oblique contexts. At the same time the above description has not pointed out sufficiently one aspect of Pesetsky's (2013) analysis. The author assumes that Russian paucals (the low cardinals 2–4) represent primitive features of NBR (a lower Number head) which move to the higher Num head (and then to D). In Czech, the paucals exhibit patterns, as in (2) above, and therefore there is no reason to assume the position of NBR in Czech is separate from the nominal stem. NBR forms a part of the φ features of the NP and as evidenced by the default agreement, those features are not accessible on the DP level when the QP is in the subject position.

On the other hand, based on the constraint on countability, I propose that the merge of P, which requires a [Q] feature, triggers the movement of the NBR features to the level of the DP. Crucially, this movement happens only when those NBR features are marked, i.e., countable and [-SG].

If the above proposals are on the right track, I conclude that the so-called "oblique" morphology on the Czech Qs, i.e., *-a*/*-i* mentioned in the discussion of (5), does not

realise any case at all, but it instead realises PLURAL, the marked feature (i.e., [-SG]) of NBR present on D/Num.

6. Paradigmatic Gap with the Q *málo* "few/ little" + [qN] in Obliques

This section summarizes the corpus data concerning the oblique form of the Q *málo* "few/little." Recall that in both (4a) and (4b), this Q is evaluated as unacceptable in all oblique contexts, i.e., irrespective of the number or countability of the following qN. Table (12) shows the varieties found in the total of 776 examples of the Q *málo* "few/little" (a) following oblique prepositions, (b) in the oblique form, i.e., with the inflection *-a.*

We can also see in (12a/ b) the most frequent structures containing a quantifier other than the searched for Q *málo.*

(12) ***málo*** "little/few" in Obliques (with a P that selects non-ACC)

	Tokens	**Collocation**	**Example**	**Type of pattern**	**See below**
(a)	727	one of few	*jeden z mál-a*	of GEN?	in (13)
(b)	4	next/ first (ADJ) of	*další/první z mál-a*	of GEN?	in (14)
(c)	16	P/V/(N?) + GEN	*mál-a*	GEN context	in (15)
(d)	28	some few =Q+Q	*několika mál-o*	frozen (*-o*) after Q?	in (17)
(e)	1	?? LOC	*mál-a*	5 & up Qs (error?)	in (i)
	776	Total			

In the lines (12a–b) above, the initial quantifiers *jeden*/*další* "one/ next" of the compound collocations select a partitive preposition *z* "of/from" which requires GEN. Therefore the following form of Q is *mál-a.* The illustrative examples below in (13) and (14) are taken from the Syn2016 search and represent this pattern. Notice that the presence of the inflected Q *mál-a* "few/little" is not needed for the homogenous agreement and case pattern because the GEN marked on qN can be theoretically assigned by either the Q or the P.[11]

(13) Jednou z (mál-a) výjimek jsou díly . . .
one$_{INS}$ of [+GEN] (few$_{GEN}$) exceptions$_{GEN}$ are volumes$_{P.NOM}$
"One of the (few) exceptions are the volumes . . ."

11 The English glosses show that the presence of the Q adds an evaluation of the quantity as "small."

(14) . . . byl prvním z (mál-a) myslitelů . . .
was first$_{INS}$ of [+GEN] (few$_{\mathbf{GEN}}$) thinkers$_{\mathbf{GEN}}$
"(he) was the first of the few thinkers . . . "

The numbers in the lines (12c–d) show that looking for the inflected oblique form *mál-a* followed by a qN, 16 tokens were found after verbs and other prepositions. Also here the presence of the Q *mál-a* "few/little" is not required for the homogenous case and agreement pattern because **all** those verbs and prepositions select GEN. In the example below taken from Syn2015, we can see the verb *dotknout se* "touch" in (15a) and the preposition *do* "into" in (15b).

(15)

(a) tyto objevy se dotkly jen (mál-a) lidí
these discoveries REFL touched[+GEN] only (few$_{\mathbf{GEN}}$) people$_{GEN}$
"these discoveries influenced only few people"

(b) cestovat jen do (velmi mál-a) zemí
travel only to [+GEN] (very few$_{\mathbf{GEN}}$) countries$_{\mathbf{GEN}}$
"to travel to only very few countries"

Although in all the examples in (13), (14) and (15) the GEN of the qN can be theoretically assigned by the Q as well as by the Verb or Preposition, the default agreement of the Q in (3) forces us to rank them among the Q with homogenous agreement in oblique contexts, and so the form *mál-a* can therefore be labelled as GEN. The fact that the same form was not found in any other case context argues in favour of the analysis which takes the *-a* inflection with *mál-a* "few / a little" as GEN only and not as a kind of universal oblique.[12]

On the other hand, the data also suggest that the GEN inflection *-a* is not a case which can be called **post-nominal** GEN or "the GEN found after the Qs" in the Heterogeneous pattern. The following (16) would then represent an example of GEN assigned

12 The only example found in the corpus was the following, which represents a structure at the edge of acceptability and most likely an error.

(i) jen v mál-a případech ho skutečně použijete
only in [+LOC] few$_{??}$ cases$_{LOC}$ it really use
"only in few cases can you really use it"

by a noun. No such an example was found in the corpus and its acceptability is very marginal.[13]

(16)	#Viděl	jsem	skupinu	mnoh-a /	?? mál-o/-a	chlapců.
	saw	Aux	[$_N$ group [+GEN]]	many$_{OBL}$ /	few$_{??}$	boys$_{GEN}$
	"I saw a group of many/few boys."					

In the few examples where a Q followed a pronoun that takes a GEN complement, the Q appeared *without* the *-a* inflection. It took the same form as in NOM/ACC, as demonstrated in (17) below.

(17)

(a)

v	několik-a	(mál-o)	minut-ách
in [+LOC]	some$_{OBL}$	(few$_{??}$)	minutes$_{LOC}$

"in some few minutes"

(b)

v	řádu	několik-a	(málo)	týdnů
in	[$_N$ scale[+GEN]]	some$_{OBL}$	(few$_{??}$)	weeks

"on the scale of some few weeks"

(c)

patřil	k	těm	několik-a	(mál-o)	cizinc-ům
belong	to [+DAT]	the$_{DAT}$	some$_{OBL}$	(few$_{??}$)	minutes$_{DAT}$

"he belonged to those several few foreigners"

To conclude, this corpus search has demonstrated that in Oblique case contexts, the combination with only Q *málo* "few/little" any qN is avoided. The structure Q+qN is replaced by a more complex idiomatic collocation *několik málo* "some few" in which the presence of the fossilised unmarked *mál-o* does not influence the morphology determined by the other lexical entries. The initial Q in the collocation (i.e., *několik* "some/several") takes a standard dual morphology (*-o*$_{NOM/ACC}$ and *-a*$_{OBL}$), as illustrated in (13). Because the oblique morphology of the initial Q requires a plural complement, the collocation is used with only a plural qN (according to the derivation proposed in the preceding section).

The only (but very frequent) context in which we can find the inflected "oblique" form of the Q *mál-a* "few/little" is a GEN context, where the form ***mál-a*** appears both after prepositions and verbs selecting GEN, but not in the post-nominal position of a GEN.

13 In contrast to the Q *mnoh-a* "much/many" which is fully acceptable in the same context under the condition that the qN is countable/plural (see [8]).

Therefore I propose that the Czech Q *málo* "few/little" exhibits a kind of paradigmatic gap. No oblique form of the Q is present in synchronic Czech, with the exception of post-verbal or post-prepositional GEN marked by a standard inflection *-a*.[14]

7. Conclusion

Based on the data found in the synchronic Czech corpus I have described a constraint which requires that the nominal complex following the quantifiers be in oblique contexts in order to be marked as plural.

I proposed that this constraint is the result of the derivation of the quantified structures as in Pesetsky (2013): assuming that oblique case represents a kind of case assigned by the category P, such a P selects a feature of [Q]. This feature, however, is visible on the D level only when the low embedded marked NBR head incorporates as a part of a high Num, which then becomes a part of the D level as demonstrated in the scheme in (10).

Based on the above analysis and the corpus data, I have argued that the inflection appearing on the Czech Quantifiers in oblique contexts is the realization of

a) a case-marking for GEN which is synchronically the only non-NOM/ACC case of the Czech Q *málo* "few/ little."
b) a marked feature of number NBR [-SG] with other Qs, especially the Q *mnoh-a* "a lot."

Funding Acknowledgement

The study was prepared thanks to the funding of the Project 16-16874S panel number P406, of the Grant Agency of the Czech Republic.

Works Cited

Babby, Leonard H. 1985. "Noun Phrase Internal Agreement in Russian." *Russian Linguistics* 9: 1–15.

Babby, Leonard H. 1987. "Case, Prequantifiers, and Discontinuous Agreement in Russian." *Natural Language and Linguistic Theory* 5: 91–138.

14 Why the gap includes a post-nominal (or post-pronominal or post Q) GEN is not clear to me. In any case the form is morphologically appropriate for GEN of a relevant nominal (Neuter) paradigm (as in *město* "city") as well as to oblique of standard Qs. This syncretism might support the usage of the form synchronically. On the other hand, the same morpheme is also used in compounds like *málo+mluvný* "speaking a little." In the compounds with prepositions, it has a case inflection, e.g., *po+málu* "a little" (literally "in little$_{LOC}$"), *bez+mála* "nearly" (lit. "without a little$_{GEN}$"). A frequent form of the prefix contains a short version of the morpheme, e.g., *malo+město* "small town," *malo+obchod* "small business" or *malo+věrný* "faithless." I do not discuss these forms here, but note only that the paradigmatic gap is synchronic.

Baerman, Matthew, Dunstan Brown, and Greville Corbett. 2002. "Case Syncretism in and out of Indo-European." In *CLS37: Main Session. Papers from the 37th Meeting of the Chicago Linguistic Society*, 15–28. Chicago: Chicago Linguistic Society.

Béjar, Susana, and Daniel Currue Hall. 1999. "Marking Markedness: The Underlying Order of Diagonal Syncretisms." In *Proceedings of the 1999 Eastern States Conference on Linguistics*, 1–12. Ithaca: Cornell University Press.

Caha, Pavel. 2016. "GEN.SG = NOM.PL: A Mystery Solved?" *Linguistica Brunensia* 64 (1): 25–40.

Caha, Pavel. Forthcoming. "Three Kinds of 'Homogenous' Patterns of Czech Numerals: A Phrasal Spell Out Account." In *FASL 24*, Michigan Slavic Publications.

Franks, Steven. 1994. "Parametric Properties of Numeral Phrases in Slavic." *Natural Language and Linguistic Theory* 12: 597–674.

Franks, Steven. 1995. *Parameters of Slavic Morphosyntax*. Oxford: Oxford University Press.

Jiranová, Pavlína. 2008. "Morfologická a syntaktická charakteristika českých číslovek vyjadřujících počet entit, jejich souborů a druhů." [Morphological and syntactical description of Czech numerals expressing number, sets and types of entities] MA thesis, Charles University in Prague.

Lahne, Antje. 2007. "On Deriving Polarity Effects." In *One-to-Many Relations in Grammar*, edited by Andreas Opitz and Jochen Trommer, 1–22. Leipzig: University of Leipzig.

Manzini, Rita M., and Leonardo M. Savoia. 2011. "Reducing Case to Denotational Primitives. Nominal Inflection in Albanian." *Linguistic Variation* 11 (1): 76–120.

Miechowicz-Mathiasen, Katarzyna. 2014. "Numeralization of Numeral Nouns in Polish." In *Nominal Structures: All in Complex DPs*, edited by Ludmila Veselovská and Markéta Janebová, 48–68. Olomouc: Palacký University.

Pesetsky, David. 1982. "Paths and Categories." Ph.D. diss., MIT, Cambridge, MA.

Pesetsky, David. 2013. *Russian Case Morphology and the Syntactic Categories*. Cambridge, MA: MIT Press.

Veselovská, Ludmila. 2001. "Agreement Patterns of Czech Group Nouns and Quantifiers." In *Semi-Lexical Categories: The Function of Content Words and the Content of Function Words*, vol. 59 of *Studies in Generative Grammar*, edited by Norbert Corver and Henk van Riemsdijk, 273–320. Berlin: Mouton de Gruyter.

Wunderlich, Dieter. 2012. "Polarity and Constraints on Paradigmatic Distinctions." In *The Morphology and Phonology of Exponence*, edited by Jochen Trommer, 160–94. Oxford: Oxford University Press.

Specificity and Past Participle Agreement in Catalan: A Diachronic Approach

Jorge Vega Vilanova

University of Hamburg, Hamburg, Germany

jorge.vega.vilanova@uni-hamburg.de

Abstract: Past participle agreement (PPA) was common in all Old Romance languages, but has been decreasing in the course of time. There are several approaches for the distribution of PPA in Modern Romance, but the trigger and path of this change is still unclear. In this paper, I provide a new explanation of PPA combining the concept of grammaticalization with a modern syntactic theory of Agree. On the basis of newly collected data, I show that the loss of PPA is linked to the specificity of the object. Consequently, PPA is tightly related to phenomena such as differential object marking and clitic doubling, but not directly to case assignment. Furthermore, I show that a view of grammaticalization taken as the formalization of semantic features is able to better account for the interplay between the progressive loss of PPA and the emergence of clitic doubling attested in Catalan.

Keywords: Agreement; Past Participle; Language change; Specificity

1. Introduction

Morphological agreement is a way of marking grammatical relations between two elements that may be non-adjacent but stand in a hierarchical relation to each other (cf. Corbett 2006). Subject-verb agreement is widespread among the languages of the world. Morphological markings for other grammatical relations are rarer. However, one can find different forms of direct object (DO) agreement. Hungarian has two different conjugation models according to certain features of the object, e.g., definiteness (É. Kiss 2002). Other languages use verbal affixes to mark case, person and number of the object, as does, e.g., Basque (Trask 1981). In the Romance languages too we find constructions where the DO agrees with the verb, namely, past participle agreement (PPA). In some syntactic contexts, gender and number of the DO are copied onto the past participle of compound

tense forms in Standard French (1) and Normative Italian (2).[1] As can be seen in the examples, these restrictions do not apply in the same way in all Romance languages: in French, PPA is obligatory with 3rd person clitics but optional with wh-elements; in Italian, PPA is obligatory with all kinds of clitics but ungrammatical with wh-elements. Spanish does not allow PPA in any context (3), and Catalan has different degrees of acceptability for optional PPA in different constructions (4a–c).

(1) (a) Ces sottises, Jean ne les Stand. French
These nonsense-F.PL J. NEG CL.ACC.3PL
a jamais **faites** (***fait**).
have-3SG never do-PP.F.PL do-PP.DEF
"Jean had never done such silly things." (Belletti 2006, 496–97)

(b) les sottises que Jean aurait **faites**/**fait**
the nonsense-F.PL REL J. have-SUBJ.3SG do-PP.F.PL/DEF
"the nonsense Jean would have done" (Belletti 2006, 496–97)

(2) (a) L'ho **vista** (***visto**). Normative Italian
CL.ACC.3SG.F-have-1SG see-PP.F.SG see-PP.DEF
"I've seen her." (Belletti 2006, 500)

(b) *I libri che ho letti (√ letto).
the book-M.PL REL have-1SG read-PP.M.PL read-PP.DEF
"The books I read." (Belletti 2006, 500)

(3) (a) Estas bobadas, Juan no las ha Spanish
this nonsense-F.PL J. NEG CL.ACC.3PL have-3SG
dicho (***dichas**) nunca.
say-PP.DEF say-PP.F.PL never
"Jean never told such silly things."

(b) ¿Cuántas sillas ha **pintado** (***pintadas**)?
how many chair-F.PL have-3SG paint-PP.DEF paint-PP.F.PL
"How many chairs does s/he paint?"

1 In all examples, the past participle is marked in **boldface** type, the DO is underlined.

(4) (a) Aquestes bajanades, les ha Catalan
this silly thing-F.PL CL.ACC.3PL have-3SG
dit/**dites** en Joan.
say-PP.DEF/F.PL the J.
"Joan told those silly things."

(b) les bajanades que ha **fet** (?**fetes**) en Joan.
the silly thing-F.PL REL have-3SG do-PP.DEF/F.PL the J.
"These are the silly things Joan did."

(c) Quantes bajanades ha **fet** (??**fetes**) en Joan?
how many silly thing-F.PL have-3SG do-PP.DEF/F.PL the J.
"How many silly things did Joan do?"

The phenomenon of past participle agreement is not only highly variable within a language (as seen in the attested data), but it is also variable across different languages of the same family as well as from a diachronic perspective. However, there is a cross-linguistic tendency in the Romance languages for PPA to disappear. Thus, in this paper I take a look at the restrictions in Old Catalan in order to identify the path and possible triggers for this language change. This, in turn, may provide new insights concerning the syntactic motivation of PPA and its restrictions in Modern Catalan. The paper is organized as follows: first, I expose some basic properties of PPA in the Romance languages. I then summarize some previous accounts and show that they are not able to properly explain the diachronic data. After showing some newly collected Old Catalan data, in Section 6, I provide a tentative explanation of this language change within a minimalist framework. I relate PPA to other phenomena like, e.g., differential object marking, clitic doubling, and scrambling, which also depend on the specificity of the DO. I then propose that the loss of PPA is the result of a grammaticalization process, whereas grammaticalization is understood as an ongoing change from semantic features to formal features (interpretable and uninterpretable), which in turn, due to economy reasons, end up disappearing, allowing a new formalization of the semantic features (semantic > formal > ø). I give some conclusions and point out some open issues in the last section.

2. Past Participle Agreement across Romance

Not all past participles in the verbal paradigm show agreement with the DO. In the Romance languages, mainly three kinds of syntactic contexts, excluding those cases with auxiliary BE and obligatory agreement, have been identified that trigger past participle agreement (PPA) (cf. Taraldsen 1987; Belletti 2006, among others): (i) when the object cliticizes to the left of the verb; (ii) when the object is fronted due to information structure

(topicalization or clitic left dislocation); (iii) when the object undergoes wh-movement, be it in relative or interrogative clauses. As can be seen, these conditions require that the object moves outside the VP complex, to the left of the verb. Nevertheless, these restrictions are also subject to parametrization. For example, French allows for PPA with wh-elements but not with 1st and 2nd person clitics, whereas Italian requires PPA with all kinds of clitics but it is ruled out with wh-elements.

The case of Catalan is particularly interesting. PPA is fully acceptable (although always optional) with 3rd person clitic objects. In all other contexts, PPA is rather dispreferred. This suggests that Catalan is placed at the end of the process before completely losing PPA. Hence, the more recent development of PPA in Catalan could provide new insights that help us understand how PPA disappeared in Romance in general. However, normative grammarians have usually interpreted the optionality of PPA in Catalan as a matter of stylistic variation (e.g., Fabra 1919). In doubtful cases, the decisive factor is not the spoken language but rather the French normative grammar, which has roughly been adapted to Catalan.

In Old Romance, by contrast, PPA was obligatory in almost all contexts, irrespective of the position of the DO (for Old Catalan, see Par 1928; Farreny Sistac 2004; Moll 1952; Solà 1972; Badia i Margarit 1981). The sentences in (5) show how PPA was realized with pre- and postverbal objects in Old Catalan.

(5) (a) Vols dir que d'autoritats prou
want-2SG say.INF that of-authority.F.PL enough
n'has **llestes**?
CL.ACC.PART-have-2SG read-PP.F.PL
"Do you think you have read enough authors of reference?" (14th century)

(b) que en Àsia dones no hagen
that in Asia woman-F.PL NEG have-3PL
edificades ciutats notables
build-PP.F.PL city prominent-F.PL
"that women in Asia wouldn't built up prominent cities" (14th century)

In the light of these facts, two questions arise: (a) how can we account for these additional contexts that require PPA, namely with post-verbal objects, in Old Romance? and (b) how can we account for the generalized language change occurring in all Romance languages simultaneously?

3. Previous Accounts on PPA in Modern Romance Languages

One of the first attempts to answer these questions was focussed on the grammaticalization of the auxiliary verb, i.e., from the Latin full verb HABEO to the Romance

auxiliary verb (Macpherson 1967; Smith 1995; Carmack 1996; Berta 2015). According to this, the past participle building up a small clause with the object in Latin (6a) is reanalyzed as forming a constituent with the full verb instead (6b), opening the possibility of leaving the participle without agreement (6c). Unfortunately, nothing is said about how the last step comes to be.

(6) (a) [LITTERAM SCRIPTAM] HABEO ("I have a letter written")
(b) LITTERAM [SCRIPTAM HABEO] / [HABEO SCRIPTAM] LITTERAM
(c) LITTERAM [SCRIPTUM HABEO] / [HABEO SCRIPTUM] LITTERAM

Lois (1990) and Muxí (1996) observe that there seems to be a correlation between the possibility of choosing alternating auxiliaries (BE vs. HAVE) and having PPA. They show that some languages have both auxiliary alternation and agreement (e.g., French, Italian, Occitan), whereas other languages have neither of these phenomena (e.g., Spanish, Portuguese, Romanian, Walloon). However, some languages do not exhibit this clustering of properties. For instance, Piedmontese (and spoken French) has no PPA although it does have auxiliary selection. The same problem appears when looking at Catalan, with PPA but no auxiliary alternation.

The most extended account, however, relies on the syntactic position of the object. Kayne (1989) first suggested that there is a "dedicated" functional projection for object agreement, AgrO, paralleling subject-verb agreement in AgrS. Morphological agreement succeeds under a local relation, i.e., if object and participle stay in a Spec-Head relation. Within the minimalist framework, there are several proposals trying to delimit the range of such locality restriction, mainly building on the notion of phases (e.g., Cortés 1993; D'Alessandro and Roberts 2008). Under this view, object movement is essential: it is a pre-condition for PPA. It is often assumed that object movement is case-driven (cf. Cortés 1993 and Kempchinsky 2000). However, the position and semantic import of AgrO is still subject to debate (cf. Belletti 2006). At least since Chomsky (1995), purely formal projections are avoided since they are illegible to LF. Furthermore, these accounts encounter problems explaining synchronic variation: How is case involved in PPA? Are caseless objects possible (cf. Diercks 2012)? Are other features involved? What is the relation between case, agreement and overt morphology? (cf. Lefebvre 1988; Sigurðsson 2000; Kempchinsky 2000; Sigurðsson and Holmberg 2008). Additionally, post-verbal agreeing DOs in Old Romance remain unexplained, and the trigger for this change, quite homogeneous across Romance, is not provided.

A closer look at French suggests that different agreement patterns give rise to different interpretations: the object in (7a), with PPA, is interpreted as specific and/or D-linked, whereas the object in (7b), with default marking on the participle, is non-specific and/or non-D-linked (cf. Obenauer 1992; Déprez 1998).

(7) (a) Combien de fautes a-t-elle **faites**?
how many of mistake-F.PL have-3SG-t-she make-PP.F.PL
"How many mistakes did she make?" (Belletti 2006, 508)

(b) Combien de fautes a-t-elle **fait**?
how many of mistake-F.PL have-3SG-t-she make-PP.DEF
"How many mistakes did she make?" (Belletti 2006, 508)

Belletti (2006), thus, identifies AgrO with an aspectual projection. Crucially, there is a relationship between aspect on the verb and specificity of the DO (e.g., Krifka 1989; Leiss 2000; Ritter and Rosen 2001; Fischer 2005). Depending on the specificity feature of the object in (8), some temporal modifications are infelicitous. This shows that the aspectual interpretation of a sentence correlates with the specificity of the DO.[2]

(8) (a) Cortó la leña en una hora (#toda la tarde).
cut-PAST-3SG the wood in one hour (all the afternoon)
"(S)he cut the wood in an hour (#the whole afternoon)."

(b) Cortó leña toda la tarde (#en una hora).
cut-PAST-3SG wood all the afternoon in one hour
"(S)he cut wood the whole afternoon (#in an hour)."

Summing up, although auxiliary selection seems to be related to the phenomenon of PPA in some way, I have shown that other kinds of factors need to be taken into account in order to explain all agreement patterns. Since accounts relying exclusively on object movement are not able to explain the diachronic data, I claim that other factors must be involved in PPA. The specificity of the DO seems to affect the aspectual interpretation of the whole clause, and the location of this feature is supposed to be the same as AgrO. In the remainder of this paper, I will argue that specificity is a crucial factor in explaining the change of PPA in Catalan, and probably in other Romance languages as well. Additionally, specificity has already been identified as having a relevant role in other phenomena affecting DOs, namely scrambling, differential object marking (DOM)

2 There is some confusion about the definition and criteria for the identification of specificity. It is not easy to tear specificity apart from definiteness (see Aissen 2003 and von Heusinger 2011 for discussion). The testing criteria are even more complicated when dealing with historical data, since there is no possibility of manipulating the utterances. Hence, I am going to treat them indistinctly. In fact, there is a tendency for definiteness to coincide with specificity (and indefiniteness with non-specificity). Thus, taking definiteness into the analysis can provide a reasonably good approximation to the phenomenon, with some specific indefinites and some non-specific definites blurring the picture.

and clitic doubling (CLD). An approach in terms of specificity, thus, may cover a wider range of phenomena. The interaction of these phenomena, already established by several studies, follows automatically from such a proposal.

4. PPA in a Broader Context: Scrambling, DOM, and CLD

Aspect and specificity play an important role in several phenomena concerning the DO. I will now review some of the proposals trying to unify their analyses.

CLD is a structure in which an argumental DP is doubled by a clitic within the same clause (thus excluding clitic dislocations). The DP and the clitic share one case and one thematic role. Suñer (1988) already suggested that CLD in Buenos Aires Spanish is only possible if the interpretation of the DO is specific, as in (9). In the same sentence, we see that the accusative object is marked by a preposition-like element. This phenomenon is known as differential object marking (DOM). According to von Heusinger and Kaiser (2005) and Leonetti (2004), this marker can only be inserted when the DO is specific (or definite). Additionally, Torrego (1998) shows cases of DOM where aspect (e.g., telicity) seems to be the decisive factor.

(9)	Diariamente	la escuchaba	a	una mujer
	daily	CL.ACC.3.F.SG-hear-PAST.IMPF.1/3SG	DOM	a woman-F.SG
	que	cantaba	tangos.	
	REL	sing-PAST.IMPF.3SG	tangos	
	"I/he/she used to hear a woman singing tangos daily."			

Scrambling in Germanic languages is supposed to obey similar restrictions, i.e., objects that move outside the VP get a quantificational specific reading (10a), whereas objects that remain within the VP get a cardinality non-specific reading (10b) (Diesing 1992; de Hoop 1992). Whether the DO is inside or outside the VP is seen by the relative position to certain particles like *ja doch* in German (10), placed at the VP boundary.

(10)	(a)	Weil	zwei Cellisten	**ja doch**	in diesem Hotel	abgestiegen	sind.
		since	two cellists	PRT	in this hotel	lodge-PP	be-3PL
		"Since two cellists lodged in this hotel."					
	(b)	Weil	**ja doch**	zwei Cellisten	in diesem Hotel	abgestiegen	sind.
		since	PRT	two cellists	in this hotel	lodge-PP	be-3PL
		"Since two cellists lodged in this hotel."					

Due to the similar conditions of all these phenomena, it is easy to find proposals connecting two or more of them. So, Jaeggli (1986) formulated what is known as Kayne's Generalization, i.e., the claim that CLD is dependent on DOM. Sportiche (1995), as well

as Alexiadou and Anagnostopoulou (1997) and Anagnostopoulou (2016) propose that CLD and scrambling are in complementary distribution. More relevant to the present paper, Tsakali and Anagnostopoulou (2008) suggest that clitic languages fall into one of the two categories: (1) languages with PPA, which have split-checking of the φ-features of the DO ([Gender] and [Number] are checked in AgrO; [Person] is checked higher, perhaps in CliticVoice, ClP, following Sportiche [1995]); and (2) languages with CLD and bundle-checking of the φ-features of the DO (AgrO is not projected).[3] However, this account is still insufficient to explain cases of optional PPA, neither can it explain why PPA is obligatory in Old Romance, since in their view it is movement that triggers agreement, although in Old Romance agreement seems to be independent of object movement. Furthermore, there are several languages that do not fall into any of these categories, e.g., Modern Catalan, allowing both CLD and PPA. I claim that although the main lines of their analysis are correct, their predictions are too strong. In the next sections, I will argue that diachronic data can shed some light on the relationship between DOM/CLD on the one side and PPA/Scrambling on the other. It is not necessarily an excluding relation, but what one finds is a gradual substitution. Furthermore, I suggest that this process is driven by changes in the kind of features encoded by the relevant structure (e.g., AgrO and ClP or AspP). Hence, the loss of PPA and the emergence of CLD can be tracked back to different feature configurations. The most natural candidate is, thus, specificity, as already suggested in Section 3.

5. Old Catalan PPA

In order to better understand how PPA gets lost from Old to Modern Catalan, I gathered Old Catalan data from the 11th to the 16th centuries.[4] For each century, the first 100 pages of two or three prose texts were analyzed (excepted for the 11th and 12th centuries: the written records are too scarce). All sentences with past participles in verbal function, e.g., in compound verb tenses, were excerpted, excluding passives, which have always auxiliary BE and obligatory agreement, and masculine singular objects, indistinguishable from the default form of agreement. 1,091 sentences were found, distributed along the centuries as

3 Franco (1994) connects PPA to the categorial status of clitics: Old Catalan clitics, being XP, can enter into a Spec-Head relation with the participle, triggering PPA. When they reduce to the category X°, they act as agreement markers and cannot trigger PPA anymore. The grammaticalization path of clitics has been assumed to be an important factor in the explanation of the emergence of CLD as well (Fontana 1993, but also Vega Vilanova et al., forthcoming).

4 The following texts were used: *Llibre de meravelles* (1288) by Ramon Llull; *Crònica* (1299) by Bernat Desclot; *Contes i faules* (1392) by Francesc Eiximenis; *Lo somni* (1399) by Bernat Metge; *La fi del comte d'Urgell* (1433), anonym; *Curial e Güelfa* (1468), anonym; *Col·loquis de la insigne Ciutat de Tortosa* (1557) by Cristòfor Despuig; *Epistolaris d'Hipòlita Roís de Liori i d'Estefania de Requesens* (16th century).

shown in Table 1. As can be seen, the rates of PPA until the 15th century are very high. The 16th century seems to be a point of inflection: almost half of the tokens lack agreement.

	Auxiliary HAVE [+Agreement]	**Auxiliary HAVE [−Agreement]**
11th/12th	12	2
13th	294	16 (~5.1%)
14th	297	58 (~16.3%)
15th	196	18 (~8.4%)
16th	107	91 (~46.0%)

Table 1. General rates of PPA in Old Catalan

5.1 Specificity and PPA

Since PPA seems to be related to the same functional position where aspectual information is encoded, and aspect and specificity are themselves interconnected, I first examined how specificity affects the realization of PPA in Old Catalan (see footnote 2 for the problematic of defining specificity in diachronic data).

Lack of agreement is found in Old Catalan in all kinds of contexts since the very first documents. However, it is especially frequent in the following cases:

- With [−Def] objects:

(11) car **oït** he moltes coses **que** . . .
since hear-PP-DEF have-1SG many thing-F.PL REL
"Since I heard many things that . . ." (14th century)

- In relative clauses, with objects in form of operators, which cannot be considered *sensu stricto* definite DPs:

(12) ladronices que havia **fait**
theft-M.PL REL have-IMPF.3SG do-PP.DEF
"robberies that he had done" (12th century)

- With inherent accusatives (length and time measures);
- In unaccusative verbs when they are used with the auxiliary HAVE:

(13) Han **seguit** guerres injustes
have-3PL follow-PP.DEF war-F.PL unfair-F.PL
"Unfair wars happened afterwards" (14th century)

The tendency to lack agreement increases in the 16th century. Table 2 sums up the results and shows that until the 15th century only around 15% of the indefinite objects lack agreement. In the 16th century, it is almost 70%. Table 3 shows that in all periods less than half of the non-agreeing participles have definite objects, although definite objects are much more frequent in the corpus.

	Total [−Def]	**+PPA**	**−PPA**
13th	72	62	10 (~13.9%)
14th	132	107	25 (~18.9%)
15th	44	39	5 (~11.3%)
16th	26	8	18 (~69.2%)

Table 2. Rates of PPA with indefinite objects in Old Catalan

	Total [−Agr]	**+Def**	**−Def**	**Relative clauses**	**Unacc. with HAVE**
13th	16	4 (~25.0%)	9	1	2
14th	58	21 (~36.2%)	26	7	4
15th	18	9 (~50.0%)	5	3	1
16th	89	25 (~28.1%)	45	15	6

Table 3. Types of DOs with non-agreeing past participles in Old Catalan

In a nutshell, indefinite DPs can be assumed to be the first context where PPA disappears. Definite DPs, especially when placed to the left of the verb, still trigger agreement.

5.2 DOM and PPA

Since the preposition-like marker of DOM objects already assigns case to the DP (e.g., Jaeggli 1986), it is expected that DOM blocks the agreement relation between participle and DO. DOM is not yet attested in the oldest Catalan texts. However, when it first appears, DOM and PPA do not usually appear simultaneously in the same clause (14), although there are occasional exceptions (15).

(14) (a) he aja **perdonat** a tots aquells qui . . .
and have-SUBJ.1SG forgive-PP.DEF DOM all-M.PL that-M.PL REL
“and I had forgiven all those that . . .” (14th century)

(b) los geògrafos que han **descrit** a Espanya
the geographer-M.PL REL have-3PL describe-PP.DEF DOM Spain-F.SG
“The geographers that described Spain . . .” (16th century)

(15) que havie **spolsada** una vegada
that have-PAST.3SG expel-PP.F.SG once
a la dita na Grahullana
DOM the mentioned-F.SG na Grahullana-F.SG
"that he had once expelled na Grahullana already mentioned"
(Farreny Sistac 2004, 344; 16th century)

In fact, the same pattern is found in Modern Catalan. PPA with 3rd person clitic objects is commonly accepted by all speakers, but when the clitic is doubling a full object DP with DOM (16a), PPA is rejected. In this case, there are two possible repair strategies: either the full DP must be dislocated (16b) or the participle must show default agreement (16c). In sum, both the synchronic and diachronic data suggest that DOM and PPA stand in a complementary distribution.

(16) (a) *Les ha **vistes** a elles.
CL.ACC.3.F.PL have-3SG see-PP.F.PL DOM they-F.PL
"He/she has seen them."

(b) Les ha **vistes**, a elles.
CL.ACC.3.F.PL have-3SG see-PP.F.PL DOM they-F.PL

(c) Les ha **vist** a elles.
CL.ACC.3.F.PL have-3SG see-PP.DEF DOM they-F.PL

5.3 CLD and PPA

Evidence for a correlation between CLD and PPA is elusive in Old Catalan. Only a few instances of accusative CLD are attested until the 15th century, but these do not show up in the relevant contexts. However, clauses where PPA co-appears with CLD are not attested at all. This indirectly confirms the idea that languages allow either CLD or PPA.

5.4 Interim Summary

In this section, I have shown that the loss of PPA in Catalan does not occur randomly, but rather follows a clear path. The 16th century is a turning point: PPA decreases, especially if the DO is indefinite. This leads us to look at other related phenomena like CLD and DOM, which emerge during the exact same period. The presented data suggest that the latter are gradually substituting PPA. All these phenomena show alternation during the 16th century but do not coexist within the same clause. As for the trigger of this change, specificity has been shown to have an effect on PPA and can plausibly be identified as the main trigger for PPA. Hence, the following five stages can be claimed to depict the language change with respect to PPA in Catalan:

- Obligatory PPA (12th to 15th centuries)
- PPA connected to specificity (16th century)
- PPA depending on syntactic placement of the DO (Modern Italian and French)
- Optional PPA (spoken French, Modern Catalan)
- Loss of PPA altogether (Spanish, Portuguese, and Romanian).

6. A Tentative Account: Formal Features, Agree, and Language Change

The next question is how the observed pattern can be motivated within a principled theory of language change. In order to do this, I will combine recent insights on syntactic agreement with the classic concept of grammaticalization. Since grammaticalization can be conceived as the loss of functional material, i.e., the change concerning the realization of functional material (cf. Fischer 2002 and 2010; Roberts and Roussou 2003; van Gelderen 2004), I propose that grammaticalization can also be captured as an ongoing change from semantic features to formal features (interpretable and uninterpretable), which in turn, due to economy reasons, end up disappearing, allowing a new formalization of the semantic features. This is shown schematically in (17). As for the phenomenon under study, this assumption means that the loss of PPA is not due to the grammaticalization of the auxiliary and past participle in the structure, but rather to the grammaticalization and formalization of certain features that lead to differences in the implementation of the operation Agree between DO and V.

(17) Semantic features > formal features > ø

In 6.1, I provide some theoretical background about syntactic operations in a minimalist framework. In 6.2, I expose how Agree works for subject-verb-agreement. Finally, I show how the same schema can be applied to object-verb-agreement as well.

6.1 Theoretical Background

According to the minimalist program as proposed since Chomsky (2000), the only operations in syntax are *Merge* and *Agree* (*Move* being understood as *Internal Merge*). *Agree* is a checking operation for formal features inserted in the syntax. Following Pesetsky and Torrego (2007), formal features come into the derivation as interpretable or uninterpretable features, valued or unvalued, i.e., we have all combinations listed in (18), where iF stands for interpretable features, uF for uninterpretable ones, *val* represents a specific value for a feature, and a low line means an unvalued feature. At least one occurrence of every feature in a clause must be interpretable in order to be legible to the interface of LF.

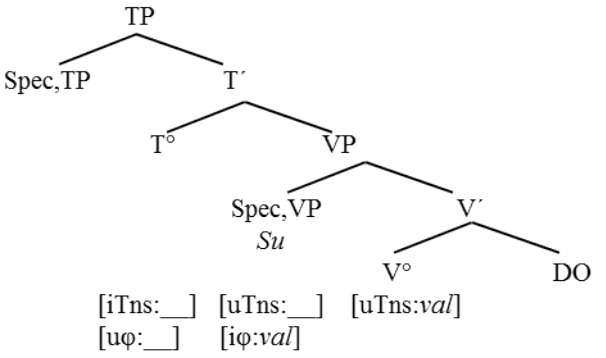
(18) (a) iF:*val*
(b) iF:___
(c) uF:*val*
(d) uF:___

Thus, checking and valuation are two independent processes. Since interpretability does not play a role until LF, it is rather the need of valuation that triggers Agree. However, there is no consensus about its directionality. The most widespread view (e.g., Epstein and Seely 2006) is upwards-valuation, i.e., the uninterpretable feature must search an interpretable counterpart within their c-command domain. Zeijlstra (2012) and Wurmbrand (2012) argue that downward-valuation, with the interpretable feature c-commanding the uninterpretable one, accounts for a wider range of phenomena. Again, there are also proposals of variable directionality (e.g., Baker 2008; Carstens 2016). In this paper, I adopt the model of downwards-valuation. I also postulate a strict similarity between object-verb and subject-verb agreement, as already claimed by Kayne (1989): the same processes of case assignment and person/number agreement that apply to the subject in AgrS are also found in a parallel functional projection, AgrO, where object case and agreement occur.

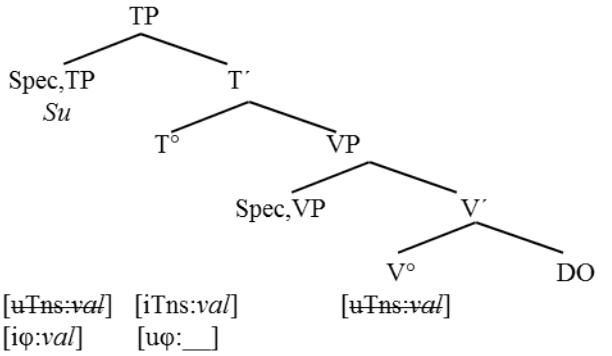
6.2 Formalizing Semantic Features: Language Change on Subject-Verb-Agreement

According to Pesetsky and Torrego (2004) and (2007), nominative Case is a reflex of the uninterpretable tense feature (uTns) on the DP. Thus, [uCase] of the subject is checked against [iTns], since nominative case can be understood as [uTns]. There is also a valued occurrence of [uTns] in V°, so [iTns] on T° c-commands all instances of [uTns], but gets its value from [uTns] on V°. This is shown in (19a). But T° still has uninterpretable φ-features [uφ] lacking any value. This triggers movement of the subject DP with [iφ:*val*] to a c-commanding position. This step is sketched in (19b), after valuation of [Tns]. Such a configuration leads to SVO word-order:

(19) (a)

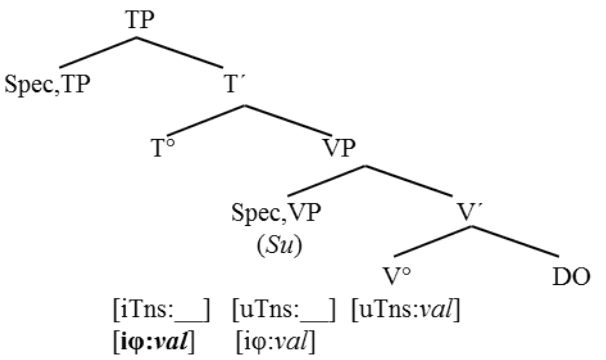


(b)



From a diachronic perspective, however, it is very common that pro-drop languages with relatively free subject placement develop into strict SVO languages (e.g., Givón 1979). Postverbal subjects triggering Long Distance Agree (LDA) could be characterized as follows: the subject DP still has [uTns:__] and [iφ:*val*], but T° comes into the derivation with an already valued [iφ] (cf. Alexiadou and Anagnostopoulou 1998). Therefore, movement is not required. This is shown in (20). I further claim that these sets of φ-features, in fact, are semantic features and can have an autonomous reference, in a doubling-like structure. Language change, thus, transforms these semantic features into a set of formal features ([iφ] on the subject, [uφ] on T).

(20)

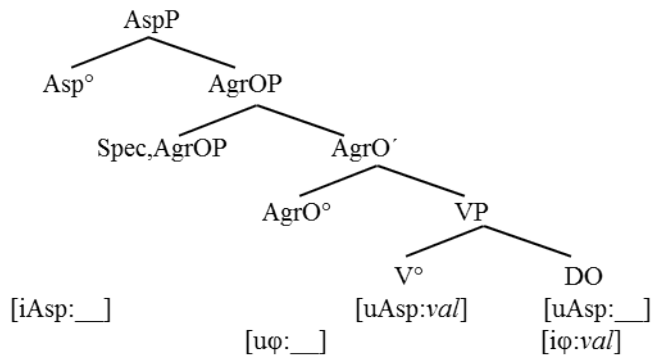


6.3 Diachronic Explanation of Object-Verb Agreement in Catalan

How does the preceding analysis account for the diachronic patterns of object-verb agreement expressed by PPA in Catalan? I assume that [uAsp] on the DO is the counterpart of [uTns] on the subject, i.e., accusative case depends directly on Asp (see references in Section 3 above). Additionally, I combine this idea with Tsakali and Anagnostopoulou's (2008) split-checking analysis for PPA. Thus, although similar in some way, object-verb agreement presents further complications.

In the first stage (12th–15th centuries), with obligatory PPA and free placement of the DO (cf. Fischer [2010] on word-order restrictions in Old Catalan), the structure one finds is such as in (20). The φ-features of the DO are checked low in the structure. Object movement to Spec,AgrO is obligatory to check the [uφ] of AgrO° (afterwards, the DO may move further for independent reasons). This head can be considered an "argumental position," providing referential values to the semantic role of the DO (cf. Koeneman and Zeijlstra 2014). Accusative case is checked against Asp, so PPA does not show any semantic restrictions (21).

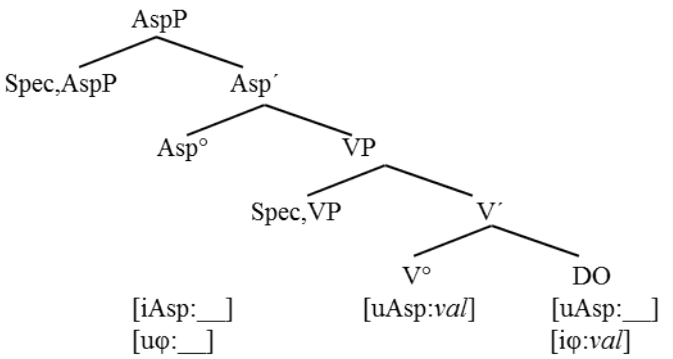
(21) (a)



(b) [$_{\text{AspP}}$ *edificades*$_{\text{i}}$ [$_{\text{AgrOP}}$ *ciutats notables*$_{\text{j}}$ *t*$_{\text{i}}$ [$_{\text{VP}}$ [~~*edificades*~~$_{\text{i}}$ ~~*ciutats notables*~~$_{\text{j}}$]]]]

When PPA is conditioned by specificity (2nd stage, from the 16th century), this can be taken as evidence that AspP and AgrOP are conflating. Only when Asp° is endowed with perfective aspect, the φ-features are active and trigger movement of the DO to Spec,AspP (22). Non-specific objects are able to check case within the VP through processes like "pseudo-incorporation" (López 2012). If the indefinite object is placed pre-verbally, PPA is still the preferred option (the rate of PPA with post-verbal indefinite objects is lower). Possible correlations between pre-verbal objects and any kind of special semantic or pragmatic reading are left for further research. VO word-order with moved objects is obtained through further verb movement to a projection over AspP (cf. Poletto 2014 for a proposal).

(22) (a)



(b) [*oït*$_i$ [$_{AspP}$ *t*$_i$ [$_{VP}$ [~~*oït*~~$_i$ moltes coses]]]]

(c) [*dites*$_i$ [$_{AspP}$ *aquestes paraules*$_j$ *t*$_i$ [$_{VP}$ [~~*dites*~~$_i$ ~~*aquestes paraules*~~$_j$]]]]

At some point, the semantic patterns for PPA are released by a mere positional criterion, i.e., PPA depends exclusively on the position where the object is placed. It is not the [uφ] on Asp° that triggers movement; rather object movement is independently motivated (cliticization, wh-movement, etc.). Agreement is also independent from [Asp].

Finally, optionality arises. It is commonly assumed that optionality is an intermediate stage in ongoing language changes. Formal φ-features of Asp° become superfluous: they are present in more than one place in the structure, although they do not trigger syntactic operations such as movement. Thus, only the DO preserves the semantic φ-features. Once PPA gets fully lost, other elements (e.g., clitics) may restart the whole process: when clitic pronouns are attached to Asp/Voice or T°, they introduce a new occurrence of φ-features, giving rise to doubling structures (CLD).

7. Some Conclusions and Open Issues

In this paper, I have provided evidence showing that specificity is a crucial factor to explain the diachronic evolution of PPA in Catalan. PPA, under this view, is connected to other phenomena like CLD, DOM and scrambling, which are all affected by the specificity feature of the DO. After the analysis of a sample of over 1,000 sentences, I was able to identify five stages in the development of PPA: (i) obligatory agreement; (ii) agreement depending on aspect/specificity; (iii) positional agreement; (iv) optionality; and (v) loss of agreement. I have shown that PPA and CLD/DOM are not mutually excluding but stand in an inverse relation: the loss of PPA correlates with the rise of CLD/DOM. To account for these patterns, I have linked the concept of grammaticalization to the theory of Agree within the minimalist framework. Just like subject-verb agreement, object-verb agreement seems to confirm the assumption that semantic features may formalize across

time, but that afterwards resulting redundancy may be reduced by economy pressure. Once PPA becomes superfluous, it becomes optional and gradually disappears. CLD, taken as another means of satisfying features in the extended verb projection, would restart the cycle reintroducing sets of semantic φ-features, which are subsequently grammaticalized as formal features.

If this analysis is on the right track, case assignment would only indirectly play a role in PPA. However, there are several open issues for further research, e.g., what kind of connection there is between [Asp] and φ-features in stage (ii) (when PPA is tight to perfective aspect). Furthermore, possible intervention effects with the indirect object should be controlled, since some datives (e.g., in Italian) may enter an agree relation with the past participle. It would be also interesting to determine if PPA in other languages, as well as other unrelated phenomena, obey the pattern described in (17). For the moment, it seems to be a promising perspective to deal with long-standing concepts of language change theorizing within modern syntactic models.

Funding Acknowledgement

This paper was partly supported by the Deutsche Forschungsgemeinschaft DFG (German Research Foundation), grant number FI 875/2-1, within the research project "Clitic Doubling Across Romance" conducted at the University of Hamburg.

Works Cited

Aissen, Judith. 2003. "Differential Object Marking: Iconicity vs. Economy." *Natural Language and Linguistic Theory* 21: 435–83.

Alexiadou, Artemis, and Elena Anagnostopoulou. 1997. "Toward a Uniform Account of Scrambling and Clitic Doubling." In *German: Syntactic Problems—Problematic Syntax*, edited by Werner Abraham and Elly van Gelderen, 142–61. Tübingen: Niemeyer.

Alexiadou, Artemis, and Elena Anagnostopoulou. 1998. "Parametrizing AGR: Word Order, V-movement, and EPP-checking." *Natural Language and Linguistic Theory* 16: 491–540.

Anagnostopoulou, Elena. 2016. "Clitic Doubling and Object Agreement." In *Proceedings of the VII Nereus International Workshop*, edited by Susann Fischer and Mario Navarro, 11–42. University of Hamburg.

Badia i Margarit, Antoni M. 1981. *Gramàtica històrica catalana.* València: Tres i Quatre.

Baker, Mark. 2008. *The Syntax of Agreement and Concord*. Cambridge: Cambridge University Press.

Belletti, Adriana. 2006. "(Past) Participle Agreement." In *The Blackwell Companion to Syntax*, edited by Martin Everaert and Henk van Riemsdijk, 493–521. Malden, MA: Blackwell.

Berta, Tibor. 2015. "On the Lack of Agreement of the Participle of Compound Tenses in Old Non-Literary Catalan Texts." *Studia Romanica Posnaniensia* 42 (5): 23–41.

Carmack, Stanford. 1996. “Patterns of Object-Participle Agreement in Eastern Ibero-Romance.” PhD diss., University of California.

Carstens, Vicki. 2016. “Delayed Valuation: A Reanalysis of Goal Features, ‘Upwards’ Complementizer Agreement, and the Mechanics of Case.” *Syntax* 19 (1): 1–42.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam. 2000. “Minimalist Inquiries.” In *Step by Step*, edited by Roger Andrews Martin, David Michaels, and Juan Uriagereka, 89–155. Cambridge, MA: MIT Press.

Corbett, Greville G. 2006. *Agreement*. Cambridge: Cambridge University Press.

Cortés, Corinne. 1993. “Catalan Participle Agreement, Auxiliary Selection and the Government Transparency Corollary.” *Probus* 5: 193–240.

D’Alessandro, Roberta, and Ian Roberts. 2008. “Movement and Agreement in Italian Past Participles and Defective Phases.” *Linguistic Inquiry* 39 (3): 477–91.

Déprez, Viviane. 1998. “Semantic Effects of Agreement: The Case of French Past Participle Agreement.” *Probus* 10: 1–65. Diercks, Michael. 2012. “Parameterizing Case: Evidence from Bantu.” *Syntax* 15 (3): 253–86.

Diesing, Molly. 1992. *Indefinites*. Cambridge, MA: MIT Press.

É. Kiss, Katalin. 2002. *The Syntax of Hungarian*. Cambridge: Cambridge University Press.

Epstein, Samuel David, and T. Daniel Seely. 2006. *Derivations in Minimalism*. Cambridge: Cambridge University Press.

Fabra, Pompeu. 1919. *Gramàtica catalana*. Barcelona: Institut d’Estudis Catalans.

Farreny Sistac, M. Dolors. 2004. *La llengua dels processos de crims a la Lleida del segle XVI*. Barcelona: Institut d’Estudis Catalans.

Fischer, Susann. 2002. *The Catalan Clitic System: A Diachronic Perspective on Its Syntax and Phonology*. Berlin: Mouton de Gruyter.

Fischer, Susann. 2005. “The Interplay between Aspect and Reference.” In *Proceedings of the Workshop ‘Specificity and the Evolution/Emergence of Nominal Determinations Systems in Romance’*, edited by Klaus von Heusinger, Georg Kaiser, and Elisabeth Stark, 1–18. Konstanz: Fachbereich Sprachwissenschaft der Universität Konstanz.

Fischer, Susann. 2010. *Word Order Change as a Source of Grammaticalization*. Amsterdam: John Benjamins Publishing Company.

Fontana, Josep M. 1993. *Phrase Structure and the Syntax of Clitics in the History of Spanish*. PhD diss., University of Pennsylvania.

Franco, Jon. 1994. “Conditions on Clitic Doubling: The Agreement Hypothesis.” *ASJU Journal of Basque Linguistics and Philology* XXVII (1): 285–98.

Gelderen, Elly van. 2004. *Grammaticalization as Economy*. Amsterdam: John Benjamins.

Givón, Talmy. 1979. “From Discourse to Syntax: Grammar as a Processing Strategy.” In *Syntax and Semantics*, vol. 1 of *Discourse and Syntax*, edited by Talmy Givón, 81–112. New York: Academic Press.

Heusinger, Klaus von. 2011. “Specificity.” In *Semantics: An International Handbook of Natural Language Meaning*, vol. 2, edited by Claudia Maienborn, Klaus von Heusinger, and Paul Portner, 1025–57. Berlin: de Gruyter.

Heusinger, Klaus von, and Georg A. Kaiser. 2005. “The Evolution of Differential Object Marking in Spanish.” In *Specificity and the Evolution/Emergence of Nominal Determination Systems in Romance*, edited by Elisabeth Stark, Klaus von Heusinger, and Georg A. Kaiser, 33–69. Konstanz: Fachbereich Sprachwissenschaft der Universität Konstanz.

Hoop, Helen de. 1992. *Case Configuration and Noun Phrase Interpretation*. PhD diss., University of Groningen.

Jaeggli, Osvaldo. 1986. “Three Issues in the Theory of Clitics: Case, Doubled NPs, and Extraction. In *The Syntax of Pronominal Clitics*, edited by Hagit Borer, 15–42. New York: Academic Press.

Kayne, Richard. 1989. “Facets of Romance Past Participle Agreement.” In *Dialect Variation on the Theory of Grammar*, edited by Paola Benincà, 85–104. Dordrecht: Foris.

Kempchinsky, Paula. 2000. “Aspect Projections and Predicate Type.” In *Hispanic Linguistics at the Turn of the Millenium*, edited by Héctor Campos et al., 171–87. Somerville, MA: Cascadilla.

Koeneman, Olaf, and Hedde Zeijlstra. 2014. “The Rich Agreement Hypothesis Rehabilited.” *Linguistic Inquiry* 45: 571–615.

Krifka, Manfred. 1989. “Nominal Reference, Temporal Constitution and Quantification in Event Semantics.” In *Semantics and Contextual Expressions*, edited by Renate Bartsch, Johan van Benthem, and Peter von Emde Boas, 75–115. Dordrecht: Foris.

Lefebvre, Claire. 1988. “Past Participle Agreement in French: Agreement = Case.” In *Advances in Romance Linguistics*, edited by David Birdsong and Jean-Pierre Montreuil, 233–53. Dordrecht: Foris.

Leiss, Elisabeth. 2000. *Artikel und Aspekt. Die grammatischen Muster von Definitheit*. Berlin: Mouton de Gruyter.

Leonetti, Manuel. 2004. “Specificity and Differential Object Marking in Spanish.” *Catalan Journal of Linguistics* 3: 75–114.

Lois, Ximena. 1990. “Auxiliary Selection and Past Participle Agreement in Romance.” *Probus* 2 (2): 233–55.

López, Luis. 2012. *Indefinite Objects: Scrambling, Choice Functions, and Differential Marking*. Cambridge, MA: MIT Press.

Macpherson, Ian. 1967. “Past Participle Agreement in Old Spanish: Transitive Verbs.” *Bulletin of Hispanic Studies* 44 (4): 241–54.

Moll, Francesc de Borja. 1952. *Gramàtica històrica catalana*. Madrid: Gredos.

Muxí, Isabel. 1996. “Object Participle Agreement with Direct Object Clitics in Catalan.” *Catalan Working Papers in Linguistics* 5 (1): 127–45.

Obenauer, Hans-Georg. 1992. "L'interpretation des structures *wh* et l'accord du participe passé." In *Structure de la phrase et théorie du liage*, edited by Hans-Georg Obenauer and Anne Zribi-Hertz, 169–93. Paris: Presses Universitaires de Vincennes.

Par, Anfós. 1928. *Curial e Güelfa: Notes lingüístiques y d'estil*. Barcelona: Biblioteca Balmes.

Pesetsky, David, and Esther Torrego. 2004. "Tense, Case, and the Nature of Syntactic Categories." In *The Syntax of Time*, edited by Jacqueline Guéron and Jacqueline Lecarme, 495–539. Cambridge, MA: MIT Press.

Pesetsky, David, and Esther Torrego. 2007. "The Syntax of Valuation and the Interpretability of Features." In *Phrasal and Clausal Architecture: Syntactic Derivation and Interpretation*, edited by Simin Karimi, Vida Samiian, and Wendy K. Wilkins, 262–94. Amsterdam: John Benjamins Publishing Company.

Poletto, Cecilia. 2014. *Word Order in Old Italian*. Oxford: Oxford University Press.

Ritter, Elizabeth, and Sara Thomas Rosen. 2001. "The Interpretive Value of Object Splits." *Language Sciences* 23: 425–51.

Roberts, Ian, and Anna Roussou. 2003. *Syntax Change. A Minimalist Approach to Grammaticalization*. Cambridge: Cambridge University Press.

Sigurðsson, Halldór. 2000. "The Locus of Case and Agreement." *Working Papers in Scandinavian Syntax* 65: 65–108.

Sigurðsson, Halldór, and Anders Holmberg. 2008. "Icelandic Dative Intervention: Person and Number Are Separate Probes." In *Agreement Restrictions*, edited by Roberta D'Alessandro, Susann Fischer, and Gunnar Hrafn Hrafnbjargarson, 251–80. Berlin: Mouton de Gruyter.

Smith, John Charles. 1995. "Agreement between Past Participle and Direct Object in Catalan: The Hypothesis of Castilian Influence Revisited." In *Linguistic Change under Contact Conditions*, edited by Jacek Fisiak, 271–89. Berlin: Mouton de Gruyter.

Solà, Joan. 1972. *Estudis de sintaxi catalana*. Barcelona: Edicions 62.

Sportiche, Dominique. 1995. "Clitic Constructions." In *Phrase Structure and the Lexicon*, edited by Laurie Zaring and Johan Rooryck, 213–76. Dordrecht: Kluwer Academic Publishers.

Suñer, Margarita. 1988. "The Role of Agreement in Clitic Doubled Constructions." *Natural Language and Linguistic Theory* 6: 391–434.

Taraldsen, Tarald. 1987. "Clitic/Participle Agreement and Auxiliary Alternation in Romance." In *Studies in Romance Languages*, edited by Carol Neidle and Rafael A. Gómez Cedeño, 263–81. Washington, DC: Georgetown UP.

Torrego, Esther. 1998. *The Dependencies of Objects*. Cambridge, MA: MIT Press.

Trask, Robert L. 1981. "Basque Verbal Morphology." In *Iker-1: Euskalarien nazioarteko jardunaldiak* [*Proceedings of the International Congress of Vasconists*], 285–304. Bilbao: Euskaltzaindia.

Tsakali, Vina, and Elena Anagnostopoulou. 2008. "Rethinking the Clitic Doubling Parameter: The Inverse Correlation between Clitic Doubling and Participle Agreement." In *Clitic Doubling in the Balkan Languages*, edited by Dalina Kalluli and Liliane Tasmowski, 321–57. Amsterdam: John Benjamins Publishing Company.

Vega Vilanova, Jorge, Mario Navarro, and Susann Fischer. Forthcoming. "The Clitic Doubling Cycle: A Diachronic Reconstruction." In *Romance Syntax. Comparative and Diachronic Perspectives*, edited by Rodica Zafiu et al. Newcastle: Cambridge Scholars Publishing.

Wurmbrand, Susanne. 2012. "The Syntax of Valuation in Auxiliary-Participle Constructions." In *Coyote Working Papers: Proceedings of the 29th Annual West Coast Conference on Formal Linguistics (WCCFL)*, edited by Jaehoon Choi et al., 154–62. Tucson: University of Arizona.

Zeijlstra, Hedde. 2012. "There Is Only One Way to Agree." *The Linguistic Review* 29: 491–53.

Definiteness Agreement in Hungarian Multiple Infinitival Constructions

Krisztina Szécsényi[a] and Tibor Szécsényi[b]

[a]Eötvös Loránd University; Research Institute for Linguistics of the Hungarian Academy of Sciences, Budapest, Hungary;
[b]University of Szeged, Szeged, Hungary

[a]kszecsenyi@gmail.com; [b]szecsenyi@hung.u-szeged.hu

Abstract: Based on evidence from multiple infinitival constructions and their parallels with preverb climbing data, the paper argues for a cyclic account of definiteness agreement in Hungarian as opposed to earlier long distance agreement proposals. Though in sentences taking only one infinitival complement it is sensible to assume that the agreeing finite verb agrees with the object of its infinitive, multiple infinitival constructions unambiguously show that, in spite of the lack of a morphological marker for definiteness on the infinitives themselves, the properties of the infinitive also play a role in definiteness agreement: infinitives covertly agree with their objects in definiteness and the finite verb agrees with the more local definiteness feature of its infinitival complement.

Keywords: Hungarian; infinitive; object agreement; definiteness; locality

1. Definiteness Agreement with the Object

1.1 The Data

As observed among others by É. Kiss (1989; 2002), Hungarian verbs show what is called definiteness agreement with the object: if the object is definite, it is the definite conjugation of the verb that appears (2), and we have indefinite conjugation when the object is indefinite (1). The form of the indefinite conjugation is the same form that appears when the verb has no object (3).[1] The two paradigms can be seen in Table 1.

1 Though the kind of agreement discussed in this paper is usually called definiteness agreement we would like to emphasize that there is no 100 percent correlation with the definiteness of the object, so the underlying feature must be something else (see, e.g., the possessive examples in [4]). For more subtle details on the nature of the object and the form of the verb and a proposal concerning the nature of the underlying feature see Bárány (2015).

	Intransitive *fut* "run"	Transitive *lát* "see"	
		indefinite	definite
1sg	*fut-ok*	*lát-ok*	*lát-om*
2sg	*fut-sz*	*lát-sz*	*lát-od*
3sg	*fut-Ø*	*lát-Ø*	*lát-ja*
1pl	*fut-unk*	*lát-unk*	*lát-juk*
2pl	*fut-tok*	*lát-tok*	*lát-játok*
3pl	*fut-nak*	*lát-nak*	*lát-ják*

Table 1. Hungarian verbal conjugations

(1) Anna lát/*lát-ja egy könyv-et.
Anna.NOM see.INDEF/see-DEF a book-ACC
"Anna sees a book."

(2) Anna *lát/lát-ja a könyv-et.
Anna.NOM see.INDEF/see-DEF the book-ACC
"Anna sees the book."

(3) Anna fut/*fut-ja.
Anna.NOM run.INDEF/run-DEF
"Anna runs."

Our research questions concern why we end up with the same form when there is no object and when there is an indefinite object, the exact nature of the trigger, and what agrees with what in the case of definite agreement. Whether the indefinite agreement pattern is the result of no agreement or default indefinite agreement is hard to decide in light of the data above. More complex structures can say more about the nature of agreement, as shown by studies on possessive DP objects (Bárány 2015; Bartos 1999; 2000). The present paper is a further contribution along these lines focusing on infinitival complement clauses. The importance of these constructions lies in the fact that in these cases there is no direct syntactic relationship between the agreeing verb and the object of the infinitive, this way providing an optimal testing ground for the common assumption that what the finite verb agrees with is the object itself.

1.2 Previous Accounts

Concerning the exact nature of the trigger, Bartos (1999; 2000) proposes a structure based account: assuming that nominal expressions come in (at least) two types, DP and NumP, and that only definite nominals project a full-fledged DP, the necessary and sufficient condition for object agreement can be identified to be the presence of a DP projection,

whereas indefinite nominals, which project only a NumP, surface with the indefinite form of the verb. The claim is based among others on the observation that possessive DPs can be both definite and indefinite, but in spite of this, possessive nominals trigger definite agreement (4). This is easy to explain with the help of a structural account: Possessive nominals are DPs and, as such, trigger definite agreement, irrespective of whether they themselves have a definite or indefinite interpretation.[2]

(4) (a) Anna lát-ja Mari-nak a könyv-é-t.
Anna.NOM see-DEF Mari-DAT the book-POSS-ACC
"Anna can see Mari's book."

(b) Anna lát-ja Mari-nak egy könyv-é-t.
Anna.NOM see-DEF Mari-DAT a book-POSS-ACC
"Anna can see a book of Mari's."

Bartos's account is made more subtle by Bárány (2015) based on, e.g., possessor extraction data cited from Szabolcsi (1994) where the verb appears in its indefinite form due to the extraction of the dative possessor (5).

(5) Chomsky-nak nem olvas-t-ál vers-é-t. (Szabolcsi 1994, 227)
Chomsky-DAT not read-PST-INDEF poem-POSS-ACC
"You haven't read any poem of Chomsky's."

Bárány's (2015) account is a structural and feature based hybrid claiming that "[o]bject agreement is only triggered by a D head that is specified for person features. If D lacks person features, it does not trigger agreement" (Bárány 2015, 75). The person feature in Hungarian is argued to grammaticalize referentiality in the D head position.

Though the presence of definite agreement clearly depends on the properties of the object, the exact nature of the features concerned is immaterial to the purposes of the present paper, which focuses on the locality issues of definiteness agreement. Whatever the exact nature of the feature on the constituent triggering agreement turns

2 Pronominal definiteness agreement sensitive to person and number (with the definite agreement form surfacing only in third person and indefinite agreement forms appearing in first and second person) poses further problems not discussed in the present paper.

out to be, it is indicated as a [±DEF] feature in our paper. For the data discussed here the conclusions of Bárány (2015) can be assumed to carry over without modification.[3]

2. Definiteness Agreement in Infinitival Constructions

2.1 The Data

When an infinitive takes a definite or indefinite object there is no difference in the form of the infinitive; infinitives do not show overt agreement with their objects, the ending is always the same *-ni* infinitival morpheme:

(6) Anna készül olvas-ni egy/a könyv-et.
Anna.NOM prepare.INDEF read-INF a/the book-ACC
"Anna is preparing to read a/the book."

In certain well definable cases, however, the finite verb taking the infinitival clause as its argument shows definiteness agreement with the object of its infinitive. Based on their behaviour concerning definiteness agreement in the presence of an infinitival complement clause, finite verbs follow two patterns (É. Kiss 1989; Kálmán C. et al. 1989): they can be *non-agreeing*, when the form of the finite verb consistently follows the indefinite paradigm irrespective of the definiteness of the object of its infinitival complement (6), or *agreeing*, when the form of the finite verb is contingent on the presence/form of the object of its infinitive: when the infinitive takes a definite object, it appears in the definite form (7), otherwise it is indefinite (8a–b).

(7) Anna *akar/akar-ja olvas-ni a könyv-et.
Anna.NOM want.INDEF/want-DEF read-INF the book-ACC
"Anna wants to read the book."

(8) (a) Anna akar/*akar-ja olvas-ni egy könyv-et.
Anna.NOM want.INDEF/want-DEF read-INF a book-ACC
"Anna wants to read a book."

3 It is not necessarily true for the second person agreement marker *-lak*/*-lek* in (i).

(i) (Én) lát-lak (téged).
I.NOM see-1SG>2SG you-ACC
"I can see you."

For more discussion see K. Szécsényi (2017).

(b) Anna akar/*akar-ja fut-ni.
Anna.NOM want.INDEF/want-DEF run-ACC
"Anna wants to run."

Agreeing verbs are usually but not exclusively transitive verbs that can also take a (definite or indefinite) DP complement, thus having the definite agreement paradigm independently of the infinitival pattern: *akar* "want," *utál* "hate" (9a) (Kálmán C. et al. 1989). Some of the exceptions are the auxiliaries *fog* "will," *szokott* "usually does," *talál* "happen to" and the auxiliary-like element *kezd* "begin."

Non-agreeing verbs that take infinitival complements are fewer in number and include verbs like *készül* "prepare," *fél* "be afraid," *igyekszik* "eager, be in a hurry," *segít* "help."[4] In case they have an argument of their own it is not in accusative case (9b).

(9) (a) Péter akar egy bicikli-t.
Péter.NOM want.INDEF a bicycle-ACC
"Péter wants a bicycle."

(b) Anna készül a verseny-re.
Anna.NOM prepare.INDEF the race-SUB
"Anna is preparing for the race."

Apart from auxiliaries and auxiliary-like elements verbs taking infinitival complements can also take finite complement clauses introduced by the complementizer *hogy* "that." In this case an expletive pronoun associated with the subordinate clause can also appear in the finite clause[5] in the case form required by the selecting verb. Agreeing verbs have an accusative marked pronoun *azt* "that.ACC" indicating that the finite clause is an object of these verbs (10a). Non-agreeing verbs have the pronoun in the oblique case form they require (*arra* "that.SUB" in [10b]). Crucially, when an agreeing verb takes a finite clause as its complement it always surfaces in its definite paradigm (10a).

(10) (a) Anna az-t akar-ja,
Anna.NOM that-ACC want-DEF
hogy Péter el-olvas-son egy könyv-et.
that Péter.NOM PV-read-SUBJ.INDEF a book-ACC
"Anna wants Peter to read a book."

4 Hungarian control constructions are fewer than those cross-linguistically. In a lot of cases embedded finite clauses (indicative or subjunctive) are used instead. For further details see K. Szécsényi (2016).

5 For more details concerning when the presence of the expletive pronoun is optional, obligatory, or banned see Kenesei (1994, 310–18).

(b) Anna ar-ra készül,
Anna.NOM that-SUB prepare.INDEF
hogy el-olvas-sa a könyv-et.
that PV-read-DEF the book-ACC
"Anna is preparing to read the book."

In light of the data in (6)–(8), infinitival constructions lead to further questions regarding definiteness agreement. Besides our original research questions (How exactly does agreement take place? What triggers agreement?), there emerge some more subtle issues to deal with. The logical assumption is that agreement is either triggered by the object or the finite verb (or potentially both). However, if the trigger is the object it is hard to explain why there is no agreement in (6), and if it is assumed to be the finite verb, answers for questions like what the agreeing verb agrees with (if anything at all) in (8) are far from straightforward.

2.2 Previous Accounts

The data introduced in this section are usually accounted for by assuming clause union (É. Kiss 1989; Den Dikken 2004) or Long Distance Agreement (LDA) taking place between the finite verb and the object of the infinitive (É. Kiss 2002). Both approaches assume that agreement is between the finite verb and the object DP. When the infinitive has no object the finite verb selecting the infinitive shows the indefinite agreement pattern. The predictions these approaches make is that the only factor to consider is the definiteness of the object (however long distance) and that other intervening constituents do not play a role. In the next section we show that this is not supported by the data.

Considering clause union, our problems are twofold: on the one hand clause union does not necessarily have to be assumed in the definiteness agreement cases: under traditional approaches no clause union is assumed to take place in the finite clauses under discussion, which trigger the definite agreement paradigm. On the other hand, assuming that non-agreeing verbs fail to participate in clause union with their infinitival complement fails to capture that these verbs actually show other clause union effects, such as scrambling. In example (11) the subject of the matrix verb, Anna, is scrambled with the constituents of the infinitival clause, but the matrix verb itself does not show definiteness agreement due to its non-agreeing nature.

(11) Holnap készül el-olvas-ni Anna a könyv-et.
tomorrow prepare.INDEF PV-read-INF Anna.NOM the book-ACC
"Anna is preparing to read the book tomorrow."

3. Multiple Infinitives

3.1 New Data

As stated before, our research question concerns the exact nature of agreement: its trigger and what exactly agrees with what. We cannot say that agreement depends on the argument structure of verbs: it is not only verbs also taking DP objects that can agree, the auxiliaries of Hungarian and some auxiliary-like elements also show agreement. Agreement does not exclusively depend on the presence of a definite object either: there are verbs that fail to agree with it. The data in (12) (first described in T. Szécsényi [2009], and extensively discussed in T. Szécsényi and K. Szécsényi [2016]) can shed some light on the agreement patterns observed. In (12a–b) containing *akar* "want," a verb that also shows definiteness agreement when finite, the verb *fog* "will," also an agreeing verb, shows agreement for definiteness. In (12c–d) the verb *fél* "be afraid" is one not showing definiteness agreement. As a result, *fog* "will" cannot show definiteness agreement with the object of the infinitive embedded into the non-agreeing infinitival clause either: the presence of the non-agreeing verb blocks agreement.

(12) (a) Péter fog/*fogja akarni nézni egy filmet.
Peter will.INDEF/will.DEF to.want to.watch a film.ACC
"Peter will want to watch a film."

(b) Péter *fog/fogja akarni nézni a filmet.
Peter will.INDEF/will.DEF to.want to.watch the film.ACC
"Peter will want to watch the film."

(c) Péter fog/*fogja félni nézni egy filmet.
Peter will.INDEF/will.DEF to.be.afraid to.watch a film.ACC
"Peter will be afraid to watch a film."

(d) Péter fog/*fogja félni nézni a filmet.
Peter will.INDEF/will.DEF to.be.afraid to.watch the film.ACC
"Peter will be afraid to watch the film."

These data indicate that agreement is not the result of LDA between the finite verb and the object of the infinitive, and clause union does not necessarily have to be assumed either (at least in order to account for agreement). Definiteness agreement seems to have a cyclic nature: the type of the infinitive also has an effect on the availability of definiteness agreement in the main clause. Based on the data in (12) we arrive at the conclusion that the infinitive also agrees with the object covertly, and that the verb selecting the infinitive agrees not with the object itself, but with the definiteness feature of the infinitive selecting it, if the infinitive has one.

The problem LDA runs into is the result of the seemingly non-unidirectional nature of definiteness agreement. In simple infinitival constructions an LDA account seems to be feasible, but multiple infinitives draw attention to the fact that definiteness agreement is the result of a more complex interaction between the clauses concerned. Definiteness agreement depends on the properties of both constituents participating in definiteness agreement: the finite verb in one clause and the object potentially appearing in the embedded clause. As we have already seen earlier in (6)–(8), if the trigger is the object it is not clear why there is no agreement in (6), and if it is assumed to be the finite verb, answers to questions like what the agreeing verb agrees with (if anything at all) in (8) are far from straightforward. It is even more highlighted in the multiple infinitival constructions in (12), which shows that the properties of the intervening verbs also play a role.

3.2 Preverb Climbing

Patterns similar to what we have just identified in the multiple infinitival constructions in (12) can be observed in the case of multiple instances of preverb climbing (for detailed discussions of preverb climbing see É. Kiss [1999]; Koopman and Szabolcsi [1999; 2000] and É. Kiss and Van Riemsdijk [2004]). Whether preverb climbing takes place depends on a property independent of definiteness agreement and hence the group of verbs participating in it differs from the division of verbs into agreeing and non-agreeing groups along the definiteness agreement property. Certain verbs identified as stress avoiding verbs (e.g., the agreeing verbs *fog* "will," and *akar* "want," and the non-agreeing verb *készül* "prepare") trigger preverb climbing which leads to the patterns shown in (13). The preverb (PV) *be* "in" belongs to the infinitive which is reflected in the translation of the sentence as well. However, due to the stress avoiding property of these verbs, in neutral sentences the preverb appears in a position preceding the stress avoiding verb.[6]

6 If the infinitival complement has no preverb of its own, it is another dependent (i) or the infinitive itself (ii) that appears in the position preceding the stress avoiding verb. Another strategy for avoiding stress is with the help of a focused constituent (iii) the position of which in Hungarian is the specifier of a FP projection directly preceding the verb. In the presence of a focused constituent (in the capital letters) bearing focus stress the verb automatically sits in a position with no stress.

(i) Péter könyv-et akar olvas-ni.
Péter.NOM book-ACC want read-INF
"Peter wants to read a book."

(ii) Péter olvas-ni akar.
Péter.NOM read-INF want
"Peter wants to read."

(13) Anna be akar ~~be~~ men-ni a szobá-ba.

Anna.NOM PV want go-INF the room-INE
"Anna wants to go into the room."

Non stress avoiding verbs (e.g., the agreeing verbs *utál* "hate" and *imád* "adore," and the non-agreeing verb *fél* "be afraid") trigger no preverb climbing, the preverb appears together with its infinitive (14a), the order with preverb climbing is ungrammatical (14b).

(14) (a) Anna utál be men-ni a szobá-ba.
Anna.NOM hate PV go-INF the room-INE
"Anna hates to go into the room."

(b) *Anna be utál ~~be~~ men-ni a szobá-ba.

Anna.NOM hate PV go-INF the room-INE
"Anna hates to go into the room."

Turning to preverb climbing in multiple infinitives we find a pattern similar to what we saw in the case of definiteness agreement: the presence of a non stress avoiding verb blocks preverb climbing. When the embedded infinitives are all stress avoiding, the preverb can end up in the position preceding the highest stress avoiding verb (15). What we see in (16) is that a non stress avoiding verb, *utál* "hate" appears between two stress avoiding ones. In this case the preverb of the most deeply embedded infinitive, *be* "in" cannot save the highest stress avoiding verb from appearing in a position associated with stress due to the blocking effect of the non stress avoiding verb (16a). Due to the lack of trigger for movement the preverb cannot end up in a position from where it could move on to the position preceding the stress avoiding verb *fog* "will." Since the preverb cannot undergo the required movement step, the resulting sentence is ungrammatical (16b) or alternative ways of avoiding stress are needed, such as focusing ([16c], see also footnote 6).

(15) Anna be fog ~~be~~ akar-ni ~~be~~ men-ni a szobá-ba.

Anna.NOM PV will want-INF go-INF the room-INE
"Anna will want to go into the room."

(iii) PÉTER akar olvas-ni.
Péter.NOM want read-INF
"It is Peter who wants to read."

(16) (a) *Anna be fog ~~be~~ utál-ni ~~be~~ men-ni a szobá-ba.
Anna.NOM will hate-INF PV go-INF the room-INE
"Anna will hate to go into the room."

(b) *Anna fog utál-ni be men-ni a szobá-ba.
Anna.NOM will hate-INF PV go-INF the room-INE
"Anna will hate to go into the room."

(c) ANNA fog utál-ni be men-ni a szobá-ba.
Anna.NOM will hate-INF PV go-INF the room-INE
"Anna will hate to go into the room."

What we see in the preverb climbing data is that the properties of the intervening verbs influence whether preverb climbing takes place or not, and whether it does is also subject to strict locality requirements. Accordingly, analyses of preverb climbing propose a cyclic account where the properties of the intervening verbs are also taken into consideration.

If the accounts of preverb climbing are on the right track, we need a similar description for definiteness agreement as well due to the number of relevant parallels: the locality restrictions observed in definiteness agreement, namely that the agreeing finite verb agrees with the object only if all the intervening verbs are agreeing as well, indicating that agreement is not directly between the finite verb and an embedded infinitival object. The multiple infinitival constructions discussed above show that definiteness agreement is cyclic, taking place from clause to clause.

4. Potential Implementation

Based on the observations of the present paper the description of the properties of definiteness agreement in infinitival constructions needs the following components:

- The most embedded infinitive agrees with its object covertly, it is just a morphological property of Hungarian infinitives that they do not show overt definiteness agreement with their objects.[7]
- In order to account for the blocking effect of non-agreeing verbs we have to assume that what the agreeing verb agrees with is the definiteness feature of its own infinitival complement. Non-agreeing verbs have a lexically defined indefinite feature.

7 Hungarian infinitives do show person and number agreement with their subjects under certain conditions, for further details see Tóth (2000).

- Agreeing verbs always need to agree with something, therefore agreeing with an objectless infinitive is the result of default indefiniteness. The same can be stated of intransitive constructions: the lack of a definite object results in a default indefinite feature leading to the same agreement paradigm in intransitives and transitive verbs appearing with an indefinite object. The claim that clauses can also have a definiteness feature is supported by the data in (10a) as well: we have seen that finite agreeing verbs taking finite clauses as objects appear in the definite paradigm, whereas the same verbs surface in their indefinite form when they take an infinitive. This indicates that both finite and non-finite clauses have to be specified for the definiteness feature.

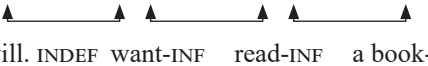
In what follows we consider the properties of the different potential patterns one by one. The two-headed arrows indicate definiteness agreement between the constituents and in this respect are to be distinguished from the arrows in the preverb climbing data, which indicate movement. Of course the parallels observed still obtain and suggest a cyclic process in both construction types.

Definiteness agreement with agreeing verbs: in the presence of a definite object and no non-agreeing verb the definiteness feature can reach the finite verb as a result of cyclic agreement (17). If the embedded object is indefinite, it is the indefiniteness feature that spreads from clause to clause (18).

(17) Anna fog-ja akar-ni olvas-ni a könyv-et.
[+DEF] [+DEF] [+DEF] [+DEF]

Anna.NOM will-DEF want-INF read-INF the book-ACC
"Anna will want to read the book."

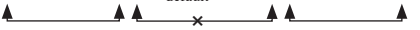
(18) Anna fog akar-ni olvas-ni egy könyv-et.
[–DEF] [–DEF] [–DEF] [–DEF]



Anna.NOM will. INDEF want-INF read-INF a book-ACC
"Anna will want to read a book."

Definiteness agreement with a non-agreeing verb (default indefinite): when a verb takes a non-agreeing infinitive (like in the case of *készül* "prepare") the result is default indefinite agreement (19). The lexically determined default indefinite feature blocks definiteness agreement with the object of its infinitival complement, but agreement with the default indefinite feature is still possible and the result is a grammatical sentence.

(19) Anna fog készül-ni olvas-ni a könyv-et.
[–DEF] [–DEF]$_{default}$ [+DEF] [+DEF]



Anna.NOM will.INDEF prepare-INF read-INF the book-ACC
"Anna will prepare to read the book."

Definiteness agreement with an objectless infinitive (default indefinite): this proposal also accounts for those patterns where the infinitive has no object: it can be argued to have a default indefinite feature; this is what the agreeing finite verb in need of a definiteness feature agrees with (20).

(20) Anna fog akar-ni fut-ni.
[–DEF] [–DEF] [–DEF]$_{default}$

Anna.NOM will.INDEF want-INF run-INF
"Anna will want to run."

Objectless finite verbs are also claimed to be the result of default indefinite agreement for the following reasons: when they appear in more complex constructions, such as infinitival complements of a verb (e.g., [20]), their properties with respect to definiteness agreement also play a role, they can be agreeing or non-agreeing. If they are agreeing, the process of agreement takes place according to the patterns we saw in (17) or (18). Non-agreeing verbs follow the pattern in (19). When they are the finite verb in a sentence and appear with no object their form can only be the result of default indefinite agreement. If they are agreeing verbs they need a feature to agree with, which can only be a default indefinite feature in the construction under discussion. If they are non-agreeing they have the default indefinite feature assigned to them in the lexicon.

(21) Anna fut.
[–DEF]$_{default}$
Anna.NOM run.INDEF
"Anna runs."

The implementation of all these slightly different constructions can be a feature-based analysis operating with default features as seen above, or, following Bartos (1999), a structural difference between definite and indefinite constituents can be assumed. Bárány (2015) offers a combination of the two for definiteness agreement in simple sentences along the following lines: in order for object agreement to arise, *v* has to be valued by a person feature via Agree with a DP direct object; and when there is no person feature a default value is assigned. Turning to infinitival complement clauses we can propose a

similar account: the finite *v* probes for a formal feature, but this time one appearing on the infinitive. As opposed to earlier accounts Agree does not have to target the nominal (DP or NumP) object, but the definiteness feature of its infinitive. The next infinitive in turn also probes for a definiteness feature it can agree with. Ultimately, each and every *v* has to be valued either via Agree with a DP or a default indefinite feature. The account proposed by Bárány has to be complemented by a proposal for non-agreeing verbs, which, similarly to those cases that lack a person feature, can be described in terms of a default indefinite feature.

An important difference between Bárány's account and ours is in the treatment of the formal feature participating in definiteness agreement. Since Bárány discusses differential object marking in simple sentences his conclusion on a person feature appearing on the D head is straightforward. When we turn to definiteness agreement in infinitival clauses we have two alternatives: either claiming that the formal feature in question is something different from the person feature or that it is the person feature we can see in Bárány (2015) but it does not have to be associated with the D head. Since the patterns observed in infinitives are parallel with the data discussed in Bárány, arguing for a different feature would lead to the loss of important generalizations. However, the multiple infinitive data suggest that the person feature is not (or not only) a D head, it can be associated with the C head as well, or, alternatively, an independent PersonP could be assumed following, e.g., Cornilescu (2016). This is nicely in line with Bartos's (1999; 2000) proposal, which argues that definiteness agreement takes place in a (then) Agr_OP, which is projected only when a DP (as opposed to a NumP) appears as the object of the verb. Our PersonP, however, should always be projected as seen in the default agreement cases. We leave the question whether the person feature is associated with a C head or it has its own projection for future research.

5. Conclusion

In this paper we have presented evidence for the cyclic nature of definiteness agreement. With the help of new, hitherto not systematically studied data showing definiteness agreement in multiple infinitival clauses and their comparison with preverb climbing data we can make the following claims:

- Definiteness agreement is more local than previously assumed;
- Properties of the intervening infinitives also play a role;
- Agreement takes place not between the matrix verb and the object of the most embedded infinitive but cyclically from infinitival clause to infinitival clause;
- Intransitive verbs and verbs with no definite object have default indefinite agreement.

Our proposal is a further addition to the growing number of proposals according to which long distance agreement is not that long distance after all.

Funding Acknowledgement

This research was funded by the Hungarian Scientific Research Funds OTKA NK 100804 and OTKA NF 84217 and by the EU-funded Hungarian grant EFOP-3.6.1-16-2016-00008.

Works Cited

Bárány, András. 2015. "Differential Object Marking in Hungarian and the Morphosyntax of Case and Agreement." PhD diss., University of Cambridge.

Bartos, Huba. 1999. "Morfoszintaxis és interpretáció: A magyar inflexiós jelenségek szintaktikai háttere." PhD diss., Eötvös Loránd University.

Bartos, Huba. 2000. "Az alanyi és a tárgyas ragozásról." In *A mai magyar nyelv leírásának újabb módszerei* IV., edited by László Büky and Márta Maleczki, 153–70. Szeged: SZTE Általános Nyelvészeti Tanszék.

Cornilescu, Alexandra. 2016. "On Dative Clitics and Nominalizations in Romanian." Paper presented at the Olomouc Linguistics Colloquium (Olinco), Olomouc, Czech Republic, June 9–11.

Dikken, Marcel den. 2004. "Agreement and 'Clause Union.'" In *Verb Clusters: A Study of Hungarian, German and Dutch*, vol. 69 of *Linguistik Aktuell*, edited by Katalin É. Kiss and Henk C. van Riemsdijk, 445–98. Amsterdam: John Benjamins.

É. Kiss, Katalin. 1989. "Egy főnévi igeneves szerkezetről." In *Általános Nyelvészeti Tanulmányok* XVII, edited by Ferenc Kiefer: 153–69. Budapest: Akadémiai Kiadó.

É. Kiss, Katalin. 1999. "Strategies of Complex Predicate Formation and the Hungarian Verbal Complex." In *Crossing Boundaries*, edited by István Kenesei, 91–114. Amsterdam: John Benjamins.

É. Kiss, Katalin. 2002. *The Syntax of Hungarian*. Cambridge: Cambridge University Press.

É. Kiss, Katalin, and Henk C. van Riemsdijk, eds. 2004. *Verb Clusters: A Study of Hungarian, German and Dutch*, vol. 69 of *Linguistik Aktuell*. Amsterdam: John Benjamins.

Kálmán C., György, László Kálmán, Ádám Nádasdy, and Gábor Prószéky. 1989. "A magyar segédigék rendszere." In *Általános Nyelvészeti Tanulmányok* XVII., edited by Ferenc Kiefer, 49–103. Budapest: Akadémiai Kiadó.

Kenesei, István. 1994. "Subordinate Clauses." In *The Syntactic Structure of Hungarian* (*Syntax and Semantics* 27), edited by Ferenc Kiefer and Katalin É. Kiss, 275–354. New York: Academic Press.

Koopman, Hilda, and Anna Szabolcsi. 1999. "Hungarian Complex Verbs and XP-Movement." In *Crossing Boundaries*, edited by István Kenesei, 115–34. Amsterdam: John Benjamins.

Koopman, Hilda, and Anna Szabolcsi. 2000. *Verbal Complexes*. Cambridge, MA: MIT Press.

Szabolcsi, Anna. 1994. "The Noun Phrase." In *The Syntactic Structure of Hungarian* (*Syntax and Semantics* 27), edited by Ferenc Kiefer and Katalin É. Kiss, 179–274. New York: Academic Press.

Szécsényi, Krisztina. 2016. “Overt and Covert Subjects in Hungarian Infinitival Clauses.” Paper presented at The Ninth conference on Syntax, Phonology and Language Analysis (SinFonIJA 9), Brno, Czech Republic, September 15–17.

Szécsényi, Krisztina. 2017. “Object Agreement and Locality.” Talk given at the Research Institute for Linguistics of the Hungarian Academy of Sciences, Budapest, Hungary, January 24.

Szécsényi, Tibor. 2009. “Lokalitás és argumentumöröklés: A magyar infinitívuszi szerkezetek leírása HPSG keretben.” PhD diss., University of Szeged.

Szécsényi, Tibor, and Krisztina Szécsényi. 2016. “A tárgyi egyeztetés és a főnévi igeneves szerkezetek.” In *“Szavad ne feledd!” Tanulmányok Bánréti Zoltán tiszteletére*, edited by Bence Kas, 117–27. Budapest: Linguistics Institute of the Hungarian Academy of Sciences.

Tóth, Ildikó. 2000. “Inflected Infinitives in Hungarian.” PhD diss., University of Tilburg.

Minimal Pronouns

Anders Holmberg[a] and On-Usa Phimsawat[b]

[a]Newcastle University, Newcastle upon Tyne, and University of Cambridge, Cambridge, UK; [b]Burapha University, Chon Buri, Thailand

[a]anders.holmberg@newcastle.ac.uk; [b]onusa@buu.ac.th

Abstract: This paper examines the properties of inclusive generic constructions, focusing on languages where the inclusive generic pronoun is a null category. We investigate empirical data from a set of languages with and without agreement to test Phimsawat's (2011) hypothesis that the inclusive generic pronoun lacks all phi-features, and therefore has the least restricted reading, due to there being no restriction on the reference. We show that this hypothesis cannot hold true universally, as phi-features trigger agreement in inflecting languages. We show that there is a correlation between presence of agreement and restriction to human reference for null inclusive generic pronouns, based on comparison of a set of languages without agreement (Thai, Korean, Vietnamese, and Sinhala) with a set of languages with agreement (Finnish, Brazilian Portuguese, Hebrew, Basque, and Tamil). An explanation in terms of feature architecture is proposed for this correlation.

Keywords: inclusive generic pronoun; phi-features; humanness

1. Introduction

The following sentences exemplify the so called inclusive generic pronoun, overt in (1), covert in (2) and (3).

(1) One shouldn't be afraid of making mistakes.
[English]

(2) Tämän koneen voi hoitaa yhdellä kädellä.
this machine can.3SG operate with one hand
"One can operate this machine with one hand."
[Finnish]

(3) díawníi ŋaan hăa yâak mâak thâa mây
nowadays job seek difficult very if NEG
còb trii.
finish BA
"To seek a job is difficult nowadays if one hasn't finished a B.A."
[Thai]

It is called inclusive because the generic reference includes the speaker, the addressee, and other people. It is, thereby, the most general of pronouns, semantically. The question we will address is how this property is encoded in the feature make-up of the pronoun. There are basically two hypotheses. One is that it is the most richly specified pronoun, specified for first, second, and third person. The other is that it is the least specified one, therefore the least restricted one, allowing reference to the speaker, the hearer, and other people. We will explore a version of the latter hypothesis, following Phimsawat (2011). The question is, what features does this minimally specified pronoun still have? A restriction that the inclusive generic pronoun has, at least in some languages, is that it can only include humans in its reference. We will show that this is true of some, but not all languages. Focusing on languages where the inclusive generic pronoun is a null category, we will demonstrate that there is a correlation between having subject agreement and having the reference of the inclusive generic subject pronoun restricted to humans. The task undertaken is to explain this correlation.

2. Inclusive, Quasi-inclusive and Exclusive

The inclusive generic pronoun can be contrasted with the quasi-inclusive generic pronoun "we," as in (4), and the exclusive generic pronoun "they" as in (5).

(4) We like smoked fish in Finland.

(5) They died young in the Middle Ages.

Generic "we" is called quasi-inclusive because it includes the speaker but not necessarily the addressee. (4) would typically be uttered by a Finn to a foreigner. It can be paraphrased as "people in general in Finland, of which I am one." Generic "they" is exclusive in that it excludes the speaker and the hearer. The pronoun in (5) can be paraphrased as "people in general in the Middle Ages." The quasi-inclusive and exclusive generic pronouns both typically require the specification of a domain, either geographical or temporal, where the temporal domain typically denotes a historical period; see Holmberg and Phimsawat (2015).

In Thai, a radical pro-drop language, the quasi-inclusive pronoun has to be overt, in an out of the blue situation, as shown by (6).

(6)	**raw**	kin	cee	nay	dɯan	tùlaakhɔm.
	we	have	vegetarian food	in	month	October

"We have vegetarian food in October."

With a null subject (6) would either be interpreted as inclusive generic ("One has vegetarian food . . .") or as having a referential 1st person subject ("I have vegetarian food . . ."). The quasi-inclusive pronoun can be null if it is bound or controlled by an overt one; see Holmberg and Phimsawat (2015).

(7)	**raw**	kin	cee	nay	dɯan	tùlaakhɔm
	we	have	vegetarian food	in	month	October
	lăŋ	**Ø**	thamboonsàjbàat			
	after		offer food to monk			

"We have vegetarian food in October after offering food to monks."

The exclusive pronoun can be overt or covert; see Holmberg and Phimsawat (2015) for more details.

(8)	bon	kɔ̀	níi	sùanyài	(**khăw**)	plùuk	chaa	khăay
	on	island	DEM	mostly	they	grow	tea	sell

"On this island they grow and sell tea."

In this the exclusive and quasi-inclusive pronouns contrast with the inclusive pronoun, in Thai, as the inclusive pronoun can be null in out of the blue sentences, in fact must be, as there is no overt counterpart.

The present paper will focus on the inclusive generic pronoun. The quasi-inclusive and exclusive pronouns are mentioned here to show that they can be clearly distinguished empirically from the inclusive one.

3. The Inclusive Generic Pronoun in Thai Has No Phi-features

What features does an inclusive generic pronoun have? The meaning is "people in general, including me and you." It has, thereby, the most general reference of all pronouns. There are two hypotheses how to encode this property as phi-features: one is that it is the most richly specified pronoun, specified for first, second, and third person, however this is formally expressed; see Hoekstra (2010). The other is that it is the least specified one, therefore allowing reference to the speaker, the hearer, and everyone else. A version of the latter hypothesis is proposed in Nevins (2007), where impersonal pronouns have an underspecified person feature; see Fenger (2016) for discussion. We will assume another version of the latter hypothesis, according to which the inclusive generic pronoun has no phi-features in some languages, namely language

without agreement, including Thai, while it has minimal phi-features in languages with agreement. Phimsawat (2011) argues, for Thai, that personal pronouns have the featural make-up (9) while the inclusive generic pronoun has (10);[1] see Déchaine and Wiltschko (2002), Holmberg (2005; 2010a, b).

(9) [uD, [φ [N]]]

(10) [uD, [[N]]

The feature uD ("unvalued D") is a referential feature, which is valued either by a referential index, which may be assigned freely or under anaphoric binding, or else by quantificational binding. In generic pronouns, and generic expressions more generally, the feature is bound by a generic operator, an adverbial operator GEN_x (= "It is generally true of x") in the C-domain (following Moltmann 2006). The phi-features include person, number, and in some languages, gender or class. We will discuss the properties of the feature/head N below. We will take this theory as a starting point. As we shall see, it cannot be the case universally that the generic inclusive pronoun is phi-featureless, because in some languages it triggers agreement.

As argued by Phimsawat, the absence of phi-feature specification explains why the inclusive generic pronoun is obligatorily null, in Thai: Having no phi-features means that there are no features to spell out, on the assumption that the uD feature and the categorial N-feature are, or at least can be, not associated with any phonological features.

This analysis of the inclusive generic pronoun is part of a theory, articulated in Phimsawat (2011), according to which arguments in Thai can be null if and only if (a) they have an antecedent which is sufficiently local, from which they can inherit a referential index, or (b) they have no phi-features but are bound by a generic operator.

An observation which can be explained immediately within this theory is that the quasi-inclusive pronoun cannot be null in Thai, in an out of the blue context. This follows since (a) the pronoun has the phi-feature value 1PL (excluding the addressee), and (b) being generic has no antecedent (see Holmberg and Phimsawat 2015). Since the value [1PL] cannot be deleted without irretrievable loss of information, it must be spelled out.

4. Inclusive Generic Pronouns and Reference to Humans

We have said, and illustrated with examples, the claim that the inclusive generic pronoun includes the speaker, the addressee, and other people in its reference. What about inanimate things and non-human animals? Can they be included as well? Is it an integral property of the inclusive generic pronoun, or possibly generic pronouns more generally,

1 In Phimsawat's (2011) notation the D-feature is R, for "referential."

that they only include humans in their reference, or is it just a consequence of the choice of predicates, so far? Predicates like "be afraid of making mistakes," "operate with one hand" and "seek a job" select a human subject. It is clearly not the case that generic reference in general is restricted to humans: *Tigers are dangerous*, *Cars are expensive* are examples of non-human generic subjects.

If it turns out that inclusive generic pronouns are restricted to human reference, this should be encoded by some feature or features, following the logic of Phimsawat (2011). We could then not maintain the explanation that the inclusive generic pronoun is null because it has no restricting features.

We will start by considering what the inclusive generic pronouns look like in some other languages.

(11)

English:	*one*, *you*
Tamil:	*oruvan* [also "one (person)"], Ø (with 3SG agreement)
Sinhala	*kenek* [also "one (person)"], Ø
Swedish:	*man* [also "man"], *du* "you"
Turkish:	*insan* [also "human"], Ø (with 3SG agreement)
Japanese:	*hito* [also "human"], Ø
Italian	*si* REFL, *tu* ("you")
Finnish:	Ø (with 3SG agreement), *sä* ("you")
Brazilian Portuguese:	Ø (with 3SG agreement), *você* ("you")
Basque	Ø (with detransitivized verb)
Thai:	Ø
Central Kurdish:	*hamu kas* ("any person")
Vietnamese	*chung ta* ["you + me + others"], Ø

English is a representative of languages where the pronoun is a cognate of the numeral "one." Other languages in this category include Tamil, where the commonest form of the overt generic inclusive pronoun is *oruvan*, which is the masculine form of the numeral "one," which can also refer to women but not to non-persons. In Sinhala, too, the inclusive generic pronoun is *kenek* ("one [person]"). Swedish, Turkish, and Japanese represent languages where the overt form of the inclusive generic pronoun is a cognate of the noun "human" or, as in Swedish, "man." Italian represents languages (including most Romance and Slavic languages) where a reflexive clitic *si* (or a cognate thereof) is used to express inclusive genericity.

(12) (a)

Si	lavora	sempre	troppo.
SI	work.3SG	always	too, much

"One always works too much."
[Italian]

(b)	W	tym	domu	umiera	się	spokojnie.
	in	this	house	die.3SG	SIĘ	peacefully

"In this house one dies peacefully."
[Polish: Krzek (2013a)]

It is debatable whether the reflexive pronoun itself is the generic pronoun, or whether it is a voice-related, detransitivizing category which serves to license a null generic pronoun; see Cinque (1988), d'Alessandro (2008), Krzek (2013a; 2013b). There are also languages where the passive is systematically used to express inclusive generic meaning. An example is Standard Arabic; see Fassi Fehri (2009). Basque, which is included in (11), also represents languages where the generic reading is marked by a special, impersonal verb form.

Finnish, Brazilian Portuguese, Basque, and Thai represent languages where the commonest form, which may be the only form, of the inclusive generic pronoun is null. Central Kurdish represents languages where there is no designated inclusive generic pronoun, but where a quantificational expression meaning something like "everyone," "anyone," or "whoever" is used. Vietnamese represents a possibly less common form of the inclusive pronoun. *Ta* means "you + me" and *chung* is a pronominal associative plural marker. This is, thus, quite explicitly an inclusive pronoun.

Many languages, but not all, have the 2SG pronoun as an alternative inclusive generic form, overt or null with 2SG agreement. Interesting though it is, we will put aside the 2SG generic pronoun in this paper; see Gruber (2013).

In some languages, the generic pronoun can be overt or null. This is the case in Japanese, for example. This is not a matter of optionality: in some contexts, the pronoun must be pronounced, in other contexts it can be null, even when not bound by another generic pronoun (Seiko Ayano, pers. comm.). It is at present unclear what determines the distribution of overt and covert inclusive generic pronouns. We leave this issue for future research.

The list in (11) indicates that humanness is common as a feature of the inclusive generic noun/pronoun, as several of the pronouns are etymologically derived from a noun meaning "human" or "man." In Tamil, the masculine inflection restricts reference to humans (Tamil has "semantic gender marking" where masculine and feminine can only refer to male and female persons, respectively). In Vietnamese, the associative plural of *ta* "you + me" can only refer to persons. It is not necessarily the case that a generic pronoun which is derived from a noun meaning "human" would be restricted to human reference, though, since it may have been grammaticalized as an even more generic pronoun, including also non-human referents.

To test whether the human restriction is endemic to inclusive genericity we need to employ a predicate which can be applied to a human as well as a non-human subject. Since the inclusive generic pronoun always includes the speaker and the addressee (or it

would not be inclusive), the predicate must be compatible with human reference. But for the purposes of this test, it must also be compatible with non-human reference. One such predicate is "grow." Humans can grow, but so can animals and plants. It is conceivable that the word for growth in humans and plants might not be the same in all languages. However, in the languages we have looked at so far, the same verb can be applied to all living beings. The test sentence we will use is a version of (13):

(13) One grows well, if one gets good care and a lot of nutrition.

The context would be a person proudly showing his garden to a visitor, offering the sentence as an explanation why the garden is so lush. The sentence is meant to be a generalisation over humans, animals, and plants. In English, (12) cannot be used in this way: the generic pronoun *one* can only refer to humans (which shows, incidentally, that the etymological link to a noun meaning "human" is not a crucial factor).

In this paper, we will, however, only consider inclusive generic constructions with a null subject. This is to test Phimsawat's (2011) hypothesis that inclusive generic pronouns are null because they have no phi-features. See Fenger (2012) for discussion of the features of overt generic pronouns.

Consider the following list of examples. The extension, humans only or humans and plants, is indicated. The sentences are meant to be uttered "out of the blue," i.e., the subject should not be anaphoric.

(14) thâa dâayráb khwaamrák khwaamʔawcaysày
if get love care
kôo cá too rew.
then FUT grow fast
"If one gets love and care, one will grow faster." [humans and plants]
[Thai]

(15) rúguǒ néng huòdé gèng duō de yíngyǎng,
if can get more much DE nutrition
nàme huì zhǎng de gèng kuài.
then be.likely grow DE more fast
"If one gets love and care, one will grow up faster." [humans and plants]
[Mandarin Chinese]

(16) yeongyangpwun -ul seopchwiha-myeon, ppali calaṇ-ta.
nutrition -ACC take -if quickly grow-PRES DECL
"If one gets more nutrition, one will grow faster." [humans and plants]
[Korean]

(17) vadi poshana labuvuth honthata hadai
more nutrition get.PTCP.CON well grow.PRS
"If one gets more nutrition, one will grow faster." [humans and plants]
[Sinhala]

(18) Nếu hấp-thụ được nhiều chất dinh -dưỡng, thì
if receive obtain many CLF nutrition COND
sẽ phát- triển nhanh.
FUT grow fast
"If one gets much nutrition, one will grow fast." [humans and plants]
[Vietnamese]

(19) Sitä kasvaa nopeammin jos saa paljon ravintoa.
EXPL grow.3SG quicker if gets much nutrition
"One grows quicker if one gets much nutrition." [humans only]
[Finnish]

(20) im meqablim harbe ahava ve
if receive.3PL much love and
maym az gdelim maher.
water then grow.3PL faster
"If one gets much love and water, one will grow faster." [humans only]
[Hebrew]

(21) Com boa alimentacão cresce mais rápido.
with good nutrition grow.3SG more quick
"One grows faster with good nutrition." [humans only]
[Brazilian Portuguese]

(22) Behar bezala zainduz gero, hemen
appropriately take.care.IMP after here
ongi hazitzen da.
well grow.HAB is
"If one is treated appropriately, one grows well here" [humans only]
[Basque]

According to our informants, the Thai, Mandarin Chinese, Korean, Sinhala, and Vietnamese examples may well be said about plants as well as animals and (necessarily) humans. The Finnish and the Hebrew examples cannot include plants. The Brazilian Portuguese example is not acceptable for all speakers (some speakers want an overt

pronoun here, which would be *você* [“you”] to convey the inclusive reading), but for those who accept it, it can only refer to humans.[2] The Basque example also cannot include plants.

One salient property that distinguishes Korean, Sinhala, Vietnamese, and Thai from Finnish, Hebrew, Brazilian Portuguese and Basque is that the former set lacks subject-verb agreement.[3]

Tamil provides some interesting evidence that agreement is, or at least can be, crucial.

(23) (a) kooda satthu kidaithaal, nalla valarum
more nutrition get.PRTC.CON well grow.FUT.**3N**
“If they get more nutrition they will grow well.” [plants, not humans]

(b) kooda satthu kidaithaal, nalla Valaruvan
more nutrition get.PTCP.CON well grow.FUT.**3SG.M**
“If one gets more nutrition, one will grow well.” [humans only]

(c) kooda satthu kidaithaal, nalla Valaramudium
more nutrition get.PTCP.CON well grow.INF.**can**
“If one gets more nutrition, one will grow well.” [humans and plants]

The null subject in (23a) can only refer to plants and animals because the gender agreement on the verb is incompatible with human reference. The null subject in (23b) can

2 Marcello Modesto (pers. comm.) has provided the following example from Brazilian Portuguese as a case where a null generic pronoun can refer to plants and animals as well as humans.

(i) Se está vivo, um dia morre.
if is alive one day dies
“Whoever/whatever is alive, will die one day.”

This means that Brazilian Portuguese and Finnish are not exactly alike in relevant respects, and suggests that the correlation between agreement and human reference is not universal. We will leave this case for future research.

3 Three other languages which have a null inclusive generic pronoun and agreement, and are reported to allow reference to humans only are Bengali (Wim van der Wurff, pers. comm.), Assamese (Hemanga Dutta, pers. comm.), and Icelandic (Halldór Sigurðsson, pers. comm.). For various reasons we don’t have examples from these languages directly comparable with the nutrition examples in (15)–(20).

only refer to humans, because the gender agreement on the verb is incompatible with non-human reference. In (23c), the head of the predicate is a modal auxiliary which does not show agreement. Now the null generic subject can refer to humans as well as animals and plants.

Why would agreement make a difference to generic reference in languages which do not show the kind of gender agreement on T that Tamil does, though?

The following is a possible hypothesis, which can, however, be rejected: In the languages without agreement the null subject in (14)–(18) is ambiguous between an inclusive generic pronoun referring to humans in general and an exclusive generic pronoun referring to plants (or non-humans) in general. This hypothesis can be rejected, at least in the case of Thai, on the grounds that there is no exclusive generic pronoun, null or overt, which would refer to plants/non-humans.

(24) thîi kɔ̀ níi yùudiikindi
at island this live well
"They live well on this island."

This sentence cannot be taken to be an exclusive generic statement about plants or animals, only about people; see Holmberg and Phimsawat (2015). To refer to plants and/or animals, the subject would have to be overt.

The following is another possible hypothesis, which can also be rejected. The subject in (14)–(18) is not a generic pronoun at all, but a multiply ambiguous referential pronoun: "I," "you," "he," "it," "they," etc., covering all people, animals, and plants. This can be rejected because referential pronouns other than first person and in some circumstances second person cannot be null in out of the blue sentences; they need a topic antecedent in the immediate discourse context (Phimsawat 2011; Holmberg and Phimsawat 2015). A first person, and in certain cases, a second person subject, can be null in out of the blue sentences because, in informal terms, the speaker and the addressee provide contextual antecedents for the null subjects. In more formal terms, the null subject can be bound by a "speaker feature" or "addressee feature," a syntactic representation of the speaker and the hearer in the C-domain (Sigurðsson 2004, 2015; Holmberg and Phimsawat 2015).

5. Inclusive Reference in Languages with Agreement

We assume a Chomskyan theory of agreement (Chomsky 2001). Subject–verb agreement is formally a set of unvalued phi-features of T, person, number, and in Hebrew also gender. These features need to be assigned a value in the course of the syntactic derivation. They are assigned a value by the subject DP, being the closest DP which is "active," not having been assigned a Case by some independent means. The valued phi-features of T are spelled out as an inflection on the finite verb or auxiliary, in the

languages under discussion here. If the unvalued phi-features are not assigned a value, the derivation will crash at PF, as they, and thereby the finite verb, cannot be spelled out.

This means that there must be a null generic subject in the structure, which has inherently valued phi-features. The agreement in the Hebrew example shows that it has 3PL.M. In Finnish and Brazilian Portuguese it has 3SG.

Could the 3SG in Finnish and/or Brazilian Portuguese generic sentences be default agreement, though? Default agreement is well known from many languages, employed when, for some reason, the phi-features of T (in the case of subject agreement) cannot be valued by the subject DP. This could be because the subject DP is assigned Case independently, and is thereby deactivated, or because there is no subject DP. Default agreement is typically 3SG. This can be seen in the Finnish sentence (25):

(25)	Minun	pitää	ostaa	uusi	auto.
	I.GEN	should.3SG	buy	new.NOM	car.NOM
	"I should buy a new car."				

Some predicates assign genitive case to the subject, in which case it cannot assign phi-feature values to T. In that case, the phi-features of T get the default value 3SG (Laitinen and Vilkuna 1993). This suggests that the 3SG agreement in construction with the inclusive null generic subject could be default agreement. The same could then be true of Brazilian Portuguese. However, as demonstrated in Holmberg (2010b), the default agreement analysis is not right for Finnish. The argument is based on the fact that default agreement and "true" agreement, including 3SG agreement, have clearly different effects elsewhere in the clause: If the subject of a transitive verb does not trigger agreement the object will get nominative case, as in (25). If the subject does trigger agreement, which entails that the subject gets nominative case, the object will get accusative case, as in (26).

(26)	Minä	voin	ostaa	uuden	auton.
	I.NOM	can.1SG	buy	new.ACC	car.ACC
	"I can buy a new car."				

As shown in (27), sentences with a null inclusive generic subject show the same variation as sentences with an overt subject, which is to say that the null subject triggers agreement just like an overt subject. In (27a) the predicate assigns genitive case to the (null) subject, hence it does not trigger agreement, and the object has nominative case. The verb has the default 3SG form.

(27)	(a)	Nyt	pitää	ostaa	uusi	auto.
		now	should.3SG	buy	new.NOM	car.NOM
		"One should buy a new car now."				

(b) Nyt voi ostaa uuden auton.
now can.3SG buy new.ACC car.ACC
"One can buy a new car now."

In (27b) the subject triggers agreement, which is 3SG because the generic subject is 3SG. In return, the subject gets nominative, and the object consequently gets accusative.

Under the present theory of agreement, the existence of subject agreement marking on the verb which can be shown not to be default agreement, is evidence that there is a subject, even though nothing is spelled out (in the case of Finnish there is no overt form of a 3SG inclusive generic subject), and shows what phi-features it has, while tests such as the nutrition sentence test, can be used to show what other restricting features it has. We take it that we have established that it has the feature [+Hum] (we will later provide a reason for taking it to be the value of a binary feature rather than a privative feature). There are other tests which can be employed to establish whether an understood, but covert subject is actually syntactically represented. Such tests have been applied to the Finnish inclusive generic pronoun, and have showed consistently that there is a syntactically represented subject (Hakulinen and Karttunen 1973; Laitinen 2006; Vainikka 1989; Vainikka and Levy 1999; Holmberg 2010b). This covert subject can bind anaphora, control a PRO subject in a purpose clause, and license agentive adverbials. See Holmberg (2010b) for examples, with details. There is consensus among the linguists who have worked on the inclusive generic construction in Finnish that it has a syntactically represented subject.

We can explain why there has to be a subject with phi-features in the languages with subject agreement. We have not explained why that subject must be restricted to human reference.

6. Explaining the Relation between Inclusive Reference, Phi-features and Humanness

First, what we call Human in grammar would be more appropriately termed something like Conscious Being, to also include talking animals and extraterrestrials and other such imaginary entities which have crucial human properties. With this proviso, we will continue to use the label Human or [±Hum].

There are various ways to integrate the feature Human in the structure of pronouns. One is that this feature is a component of N, the nominal "base" of nominal expressions, perhaps appropriately seen as the root of a pronoun, a minimal root. *He* and *she* would have the root feature Human, or [+Hum], non-human-referring pronouns like English *it* would have a [−Hum] root. We may want to make a distinction between pronouns that get their interpretation from an antecedent and pronouns that do not. In the former case the component N, the root component of the pronoun, may be taken to be a copy of the NP of the antecedent, deleted under identity with this antecedent (see Panagiotidis 2002; Elbourne 2008 for different versions of this idea). In the case of the generic pronoun, there

is no antecedent. Therefore, it needs a root of its own. The [+Hum] feature would provide this. The fact that the inclusive generic reading includes, by definition, the speaker and the addressee in the extension of the pronoun means that in the case of this pronoun, the feature [−Hum] is not an option.

But what is the connection with agreement? What about all the languages where the generic pronoun is so inclusive that it can include plants along with humans and animals? In this case the pronominal root would seem to be unspecified for humanness, [±Hum], allowing reference to entities of any kind. The generalisation that we want to express, though, suggested by our data, is that a pronoun cannot have phi-features without specification of the feature [±Hum].

The following is an alternative. First, the minimal root of a pronoun is, universally, [ENTITY]. Second, there are two ways that a pronoun can refer to everything and/or everybody: one is not to have any phi-features, hence no restriction. The other is to have minimal phi-features, just enough to satisfy the requirements of agreement, yet allowing reference to the speaker, the addressee, and a maximally general set of "non-participants." The feature [participant], widely assumed as part of pronominal systems, following Harley and Ritter (2002), distinguishes between speaker and addressee on the one hand, and everyone/everything else on the other hand. In Harley and Ritter (2002) all the features are privative. Third person is when the feature [participant] is absent, i.e., "third person is no person" (see Nevins 2008 for discussion). Such a system does not allow for a pronoun with phi-features which allow reference to the speaker, the addressee, and everyone/everything else. The system must include a feature which can be underspecified for person: [±participant] (see Nevins [2008] for other arguments that this device is needed). On its own, this feature will not exclude reference to non-human entities, and therefore must be supplemented by at least one more feature.

Assume that the phi-feature set of a pronoun has to include at least one specified feature. The pronominal phi-features are person, number, and class (Harley and Ritter 2002). The inclusive generic pronoun, although formally singular is not semantically singular. Arguably this rules out the use of a pronoun specified for singular number as an inclusive generic pronoun. Assume that the first division among the class features is between human and non-human, as seen in the many pronominal systems which make a distinction between human and nonhuman third person pronouns. The inclusive generic pronoun cannot be specified [–Hum], as it must allow inclusion of the speaker and addressee. But it can be specified as [+Hum]. The minimal feature make-up of a pronoun with phi-features which will allow inclusive, generic reference will therefore be [±Participant, +Human].[4]

4 Hebrew is a language with a null inclusive pronoun which triggers plural agreement, an option which would appear to be consistent with the semantics of inclusivity. The idea that one specified feature is enough would then seem to predict that the inclusive pronoun in Hebrew could remain unspecified for [Hum]. The data we have indicates that this is a false prediction.

This presupposes that the unvalued phi-features of T are, or at least can be, formally valued by this minimal phi-feature set, where the spell-out of the so valued T is the third person singular suffix on the finite verb (in most but not all of the relevant languages; in Hebrew it is plural). That is to say, the third person singular form that the finite verb has in Finnish, discussed in Section 5, would be a form of default agreement after all, in that the subject valuing the features of T would not be specified for person or number, but only for class (the [+Hum] value), which, however, has no morphological effect in Finnish.[5]

In languages without agreement, there is no reason why a generic pronoun would have to have any phi-features. All it needs is the root feature [ENTITY] and merged with it, the [uD]-feature. When the D-feature is bound by the generic operator this results in a reading which can be rendered as "entities in general including the speaker and the addressee," the minimally specified DP giving the maximally inclusive reading.

7. Conclusions

The starting point is the hypothesis, articulated in Phimsawat (2011), that the inclusive generic pronoun is the least specified nominal category, which therefore has the most general reference, including the speaker, the hearer, and everyone else. The observation is that there is cross-linguistic variation as to whether the pronoun is or is not restricted to humans. Focusing on languages which have a null inclusive generic pronoun in finite clauses, we have found that the null inclusive generic pronoun is restricted to human reference in some of them, but not all. The generalisation, based on data from primarily ten languages, five without agreement, four with subject–verb agreement, and one [Tamil] with or without agreement) is that the pronoun is restricted to human reference in the languages that have subject–verb agreement in finite clauses. The explanation proposed is (a) in languages with subject agreement, i.e., unvalued phi-features in T, the inclusive generic pronoun has to have at least one specified phi-feature, to value the phi-features of T; (b) if the pronoun is to be inclusive, it cannot be specified for number, which entails that it must be specified for class; (c) if the pronoun is to be inclusive, i.e., include the speaker and the addressee, it must be specified [+Hum].

5 According to the theory of null subjects in Holmberg (2010a, b), Roberts (2010b), based on the theory in Roberts (2010a), null subjects in languages with agreement are derived by copy deletion. The valued phi-features of T and the subject pronoun form a chain of two copies, where one, the subject, is deleted, provided its features are a subset of the phi-features of T. Since the subject, if it is third person, is valued for gender (i.e., class) in many languages, T must be valued for gender as well, for the subject to be deletable, even when this is not morphologically realised, as is the case in many languages. The notion that T has, or may have, an invisible class feature in languages with phi-features in T thus has independent motivation.

Funding Acknowledgement

Anders Holmberg's research for this paper was funded by the European Research Council Advanced Grant No. 269752 "Rethinking Comparative Syntax" (ReCoS). Thanks to the organisers and the audience at Encontro Intermediário do GT-TG at Universidade Federal de Minas Gerais, 2015, and especially to Fábio Bonfim Duarte for inviting Anders Holmberg. Also thanks to the support of the Faculty of Humanities and Social Sciences, Burapha University for funding On-Usa Phimsawat's trip to Olinco 2016. We wish to thank the following people for having contributed with data and discussion of the data: Seiki Ayano, Pauli Brattico, Sonia Cyrino, Maia Duguine, Ricardo Etxepare, Yujia Han, Saara Huhmarniemi, Rebeen Kareem, Sakorn Phimsawat, Shin-Sook Kim, Tawee Kueakoolkiat, Kadri Kuram, Marcello Modesto, Makiko Mukai, Trang Phan, Michelle Sheehan, Ur Shlonsky, Halldor Sigurðsson, Salinee Somtopcharoenkul, Harold Thampoe, Hofa Meng Jung Wu. Thanks also to the ReCoS team: Ian Roberts, Theresa Biberauer, Jenneke van der Wal, Sam Wolfe, Georg Höhn.

Works Cited

Cinque, Guglielmo. 1988. "On *si* Constructions and the Theory of *ARB*." *Linguistic Inquiry* 19: 521–81.

D'Alessandro, Roberta. 2007. *Impersonal si Constructions.* Berlin/New York: Mouton de Gruyter.

Elbourne, Paul. 2008. The Interpretation of Pronouns. *Language and Linguistics Compass*: 119–50.

Déchaine, Rose-Marie, and Martina Wiltschko. 2002. "Decomposing Pronouns." *Linguistic Inquiry* 33 (3): 409–42.

Fassi Fehri, Abdelkader. 2009. "Arabic Silent Pronouns, Person and Voice." *Brill's Journal of Afroasiatic Languages and Linguistics* 1: 1–38.

Fenger, Paula. 2015. "How Impersonal Does One Get? A Study of 'Man'-Pronouns in Germanic. LingBuzz. Accessed January 15, 2016. http://ling.auf.net/lingbuzz/002802.

Gruber, Bettina. 2013. "The Spatiotemporal Dimensions of Person. A Morphosyntactic Account of Indexical Pronouns." PhD diss., LOT, Utrecht University.

Hakulinen, Auli, and Lauri Karttunen. 1973. "Missing Persons: On Generic Sentences in Finnish." In *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society*, edited by Claudia W. Corum, Thomas Cedric Smith-Stark, and Ann Weiser 157–71. Chicago: Chicago Linguistic Society.

Harley, Heidi, and Elizabeth Ritter. 2002. "Person and Number in Pronouns: A Feature-Geometric Analysis." *Language* 78: 482–526.

Hoekstra, Jarich. 2010. "On the Impersonal Pronoun *Men* in Modern West Frisian." *Journal of Comparative Germanic Linguistics* 13: 31–59.

Holmberg, Anders. 2005. "Is There a Little Pro? Evidence from Finnish." *Linguistic Inquiry* 36: 533–64.

Holmberg, Anders. 2010a. "Null Subject Parameter." In *Parametric Variation: Null Subjects in Minimalist Theory*, edited by Theresa Biberauer et al., 88–124. Cambridge, UK: Cambridge University Press.

Holmberg, Anders. 2010b. "The Null Generic Subject Pronoun in Finnish: A Case of Incorporation in T." In *Parametric Variation: Null Subjects in Minimalist Theory* edited by Theresa Biberauer et al., 200–30. Cambridge, UK: Cambridge University Press.

Holmberg, Anders, and On-Usa Phimsawat. 2015. "Generic Pronouns and Phi-Features: Evidence from Thai." In *Newcastle and Northumbria Working Papers in Linguistics*. Accessed January 6, 2016. http://www.ncl.ac.uk/linguistics/research/workingpapers/Volume21.1.htm.

Krzek, Małgorzata. 2013a. "Interpretation and Voice in Polish SIĘ and -NO/-TO Constructions." In *Current Studies in Slavic Linguistics*, edited by Irina Kor Chahine, 185–98. Amsterdam: John Benjamins.

Krzek, Małgorzata. 2013b. "Generic Subjects and Voice in Polish Impersonal Constructions." In *Microvariation, Minority Languages, Minimalism and Meaning: Proceedings of the Irish Network in Formal Linguistics*, edited by Catrin S. Rhys, Pavel Iosad, and Alison Henry, 186–206. Cambridge, UK: Cambridge Scholars.

Laitinen, Lea. 2006. "Zero Person in Finnish." In *Grammar from the Human Perspective*, edited by Marja-Liisa Helasvuo and Lyle Campbell, 210–31. Amsterdam: John Benjamins.

Laitinen, Lea, and Maria Vilkuna. 1993. "Case Marking in Necessive Constructions and Split Intransitivity." In *Case and Other Functional Categories in Finnish*, edited by Anders Holmberg and Urpo Nikanne, 23–48. Berlin: Mouton de Gruyter.

Moltmann, Friederike. 2006. "Generic One, Arbitrary PRO, and the First Person." *Natural Language Semantics* 14: 257–81.

Nevins, Andrew. 2007. "The Representation of Third Person and its Consequences for Person-Case Effects." *Natural Language & Linguistic Theory* 25 (2): 273–313.

Panagiotidis, E. Phoevos. 2002. *Pronouns, Clitics and Empty Nouns*. Amsterdam: John Benjamins.

Phimsawat, On-Usa. 2011. "The Syntax of Pro-Drop in Thai." Ph.D. diss., Newcastle University.

Roberts, Ian. 2010a. *Agreement and Head Movement: Clitics, Incorporation, and Defective Goals.* Cambridge, MA: MIT Press.

Roberts, Ian. 2010b. "A Deletion Analysis of Null Subjects." In *Parametric Variation: Null Subjects in Minimalist Theory*, edited by Theresa Biberauer et al., 58–87. Cambridge, UK: Cambridge University Press.

Sigurðsson, Halldor Armann. 2004. "The Syntax of Person, Tense, and Speech Features." *Italian Journal of Linguistics* 16: 219–51.

Sigurðsson, Halldor Armann. 2015. "About Pronouns." Accessed January 8, 2016. http://ling.auf.net/lingbuzz/001593.

Vainikka, Anne. 1989. "Deriving Syntactic Representations in Finnish." Ph.D. diss., University of Massachusetts, Amherst.

Vainikka, Anne, and Yonata Levy. 1999. "Empty Subjects in Finnish and Hebrew." *Natural Language and Linguistic Theory* 17: 613–71.

Formal Lexical Entries for French Clitics: PF Dissociations of Single Marked Features

Joseph Emonds

Palacký University, Olomouc, Czech Republic

jeemonds@hotmail.com

Abstract: Systems of pronominal clitics for arguments and adverbial adjuncts of verbs in Romance languages have several regular properties that widely accepted grammatical models have yet to account for. The analysis here accounts for (i) orderings among French clitics that do not reflect syntactic phrasal ordering; (ii) the limited "structural distance" between clitics and the interpreted phrases they replace; (iii) why clitics frequently have the same form as strong pronouns; and (iv) the extent to which language-particular clitic paradigms conform to Borer's Conjecture. This system uses no clitic movements, and expresses all generalizations in terms of formalized, constrained lexical entries. Taken together, clitic properties suggest that groups of clitics are single lexical entries inserted in Phonological Form, with allomorphs specified by the parentheses/brace notations, and a "dissociation" principle called Alternative Realization.[1]

1 Twenty-five years ago, Henk van Riemsdijk organized an invitation from the Netherlands Science Foundation for a year of research at Tilburg University. He suggested as a topic for my course that I try to use Alternative Realization to account for Romance clitics. The main results were published in Emonds (1999; 2001). Some puzzles remained, e.g. the Person-Case Constraint, the ordering of *le/la/les,* and what I call here "Missing Exponents." This essay combines and unifies proposals for all these aspects of French clitics.

I want to belatedly express my gratitude to the patient, critical and yet encouraging participants in the seminar: to Henk, who created it, Hap Kolb, who helped me formulate Alternative Realization as it appears here, Riny Huybrechts, Angeliek van Hout, Bart Hollendbrandse, and several others. Independent of my own efforts, it can be said that that those years represented Tilburg's finest hour. I also thank Markéta Janebová and Monika Pitnerová for help with editing.

Keywords: Borer's Conjecture; clitic ordering; French clitics; lexical notation; person-case constraint; clause-mate condition

1. Basic Distribution of French Clitics

French clitics are pronouns and "pro-adverbs" that occur in sequence inside and at the left edge of lexical verbs (Kayne 1975, Chap. 2), even though French VPs are uniformly head-initial. Such clitic sequences are underlined in (1).[2]

(1) (a) Jean, (souvent), [$_{VP}$il ne me (*souvent) [$_{V}$dit] pas qu'il rentre].
John (often) he not me (often) tells not that he comes home
"John (often) doesn't tell me (often) that he is coming home."

(b) Marie veut [$_{VP}$ les leur distribuer pendant la réunion].
Mary wants them to them distribute during the meeting
"Mary wants to distribute them to them during the meeting."

Many sources since Perlmutter (1971) give the allowed sequences as in (2). "Person clitics" refer to 1st and 2nd person clitics and 3rd person reciprocal/reflexive *se* (Kayne 1994). More recently, French clitic ordering has been summarized in Veselovská and Vos (1999, 970), who exemplify their categories with the exponents in the third line:

(2) Ordering of French pro-clitics on verbs, plural person forms omitted:

Subj clitics –	neg –	person clitics –	dir obj 3rd per –	ind obj 3rd per –	"there" –	"thereof"
Il, *je*, *tu*, *on*	*ne*	*me*, *te*, *se*	*le*, *la*, *les*	*lui*, *leur*	*y*	*en*

In positions 3 through 7, sequences of more than two clitics are marginal, and four seem excluded. In (3), any two clitics are acceptable, but three of these (non-subject) clitics together are strange at best.

(3) (a) ?Marie te l'y expliquera lundi.
Mary you it there will explain Monday
"Mary will explain it to you there Monday."

(b) Je me l'y mènerai, . . .
I me him there bring-will
"I will bring him there just for me, . . ." (Veselovská and Vos 1999, 958)

2 I wish to thank Henri-José Deulofeu for judgments and discussion of many of this essay's examples. Errors are of course my own.

In only one construction, affirmative imperatives, do these the clitic sequences appear as enclitics, with a very few differences in ordering. I do not focus on these differences in this essay.

1.1 Person Restrictions

There is only one slot for object clitics marked for Person, as defined above (see Section 5). If both direct and indirect objects are +Person, it is often said informally that the sole clitic must be the direct object. However, an undoubled direct object pronoun can appear post-verbally in focus with *que* "only," leaving the indirect object as a sole pro-clitic:

(4) Finalement, Marie ne m'a présenté que vous (et votre femme).
"In the end, Mary to me introduced only you (and your wife)."

1.2 The Person-Case Constraint

The last 15 years have seen much discussion of a "Person-Case Constraint" (PCC), mostly in terms of combinations of pronoun objects.

(5) ***Person-Case Constraint***. French Person clitics cannot occur with 3rd person "dative" clitics.

(6) (a) *Jean me leur a présenté.
"John me to them introduced."

(b) *Marie se lut est décrite.
"Mary self to him described."

(c) Marie s'est décrite à lui.
"Mary described herself to him."

This essay will return to and explain the Person-Case Constraint in Section 6.

2. Theoretical Advantages of a Lexical Entry/PF Approach

There are four very general and empirically well justified principles that underlie this study's approach. All have played important explanatory roles in empirical descriptions, though frequently the appeals to them have been rather implicit. It is of course the aim of formal grammar to unequivocally spell out such putative universals of linguistic theory.

The first principle simply names a recognized desideratum, namely that individual morphemes should be paired, at least optimally, with single morpho-syntactic features.

(7) **Single Feature Exponents.** Optimally, phonological forms ("exponents") of grammatical items are paired in lexical entries with at most a *single marked feature.*[3]

This principle also suggests, almost implies, that paradigms in lexical representations are an illusion, if by paradigm is meant some kind of matrix whose entries spell out feature complexes of equal status. Perhaps a good way to understand (7) is to consider a form which is *not* optimal; for example, the German dative plural suffix *-en,* which seems to spell out two marked features, and thus does not conform to (7).

As a reviewer points out, the implication of Single Feature Exponents is that languages should follow what Anderson (1982) names (and also distances himself from), the "Agglutinative Ideal." While this lexical property may conflict with an (a priori) notion of economy of representation (i.e. that multi-morphemic agglutinative sequences are "less economic" than compact mono-morphemic inflections),[4] the actual role of (7) is to enhance *lexical economy*, i.e., there are fewer entries overall, and optimally each entry is simple.

(8) **Parentheses and brace notation.** The linguistically significant generalizations about single exponents are expressed in lexical entries by extensive and crucial use of *parentheses and disjunctive braces.*

Analysts long accustomed to these, structuralists as well as generativists, may overlook the fact that these two notations represent highly contentful claims about the human language faculty (Chomsky 1967; Chomsky and Halle 1968). Here we will see how parentheses and braces in lexical entries elegantly express co-occurrence properties of clitics.

(9) **Borer's Conjecture**. Natural languages differ only in their *lexical entries of grammatical items*, i.e., items which have no purely semantic features (for detail, see Ouhalla [1991] and Emonds [2000, Chap. 3 and 4]).

Actually, the source of this working hypothesis for particular grammars (Borer 1984, 29) is phrased in terms of inflections. However, its current widely accepted interpretation is

3 Thus, a 2nd person singular pronoun is specified as +2nd, but not Singular, this being an unmarked value. Similarly, 1st and 2nd person pronouns, and probably pronouns in general, need not be lexically specified as +Definite or +Animate.

4 The idea that inflection is more "compact" or "economic" than agglutination is from Humboldt (1822), who argued that the Indo-Aryan languages were superior to agglutinative languages such as Malay for developing advanced intellectual reasoning. While this dubious consequence has been discredited, his sense of what motivates inflection has not been. Veselovská and Emonds (2016) propose to reconcile the two tendencies by locating the Agglutinative Ideal at LF, thus limiting inflectional economy to PF.

as in (9). Thus, when the not dissimilar verbal pro-clitics in different Romance languages are formally expressed, Borer's Conjecture predicts that any significant differences will be best represented by their lexical entries. In fact, we will see that of the four lexical entries needed for 20 French clitics, Italian lacks counterparts for one of them and Spanish for two. Moreover, one entry that they largely share (for the Person clitics) is not the same in French as in the other two. In sum, the expectations of Borer's Conjecture are borne out; the lexical entries involving clitics precisely express the differences of each of three Romance languages.

(10) **Clause-Mate Constraint on clitics.** A clitic on a V/I can be related to only those phrases *that are immediately dominated by a projection of that same V/I.*

This claim contradicts a vast literature on "clitic movement/climbing/raising," which dates from Kayne's (1975) classic analysis of French causative constructions. These analyses eventually came to include four sub-types of clitic raising.[5] These cases are all challenged with counter-analyses in Emonds (1999); see Section 7.3 for an outline of the reasoning and sources.

Alternative Realization ("AR") is the formal centerpiece of such clause-mate analyses. It is a general structural principle which limits the structural distance allowed between the Logical Form ("LF") and Phonological Form ("PF") positions of a single item, i.e., it formalizes the "dissociation" discussed in Embick and Noyer (2001). The formulation of AR used here is cross-linguistically justified for many other syntactic constructions (especially for "inflections"; see Emonds [2000, Chap. 4]). AR thus dispenses with all syntactic movement either to or from clitic positions. Consequently, I claim that French clitics provide no justification for syntax-based feature attraction/probing or agreement, nor for heads or specifiers of clitic-based functional projections. Additionally*, no case features* are needed in the analysis of French. Only PF *allomorphs* of person and place morphemes are required (cf. Parrott 2009).

Throughout, then, this study will show that appropriate lexical entries for clitics have these ***SF–PB–AR–BC*** properties.

(11) (a) SF = exemplifies Single Feature Exponents,
(b) PB = uses Parentheses and Brace notation,
(c) AR = instantiates Alternative Realization,
(d) BC = confirms Borer's Conjecture.

5 In addition to Romance causatives, raising analyses were later proposed for complements of (i) restructuring and auxiliary verbs, (ii) adjectives, and (iii) indefinite direct objects.

3. The Position of the French V and Its Subject Clitics

3.1 Singular Subject Clitics

French finite verbs and present participles move to the functional head I (also known as T), while infinitives remain inside verbal projections VP/vP. This difference is motivated by the contrasting placements of negation and adverbials (Emonds 1978). The V moved to I includes any pro-clitic sequence of object and adverbial clitics, including possibly the negative proclitic *ne*. Subject clitics, which never occur with non-finite verbs, are not part of this operation.

This syntactic movement creates [$_{I}$V]. Subject clitics, exemplified in (13), are then attached to this I.[6] They consist of a reference feature D, which also has either marked Person features or an indefinite value +HUMAN in the third person. If a D is not marked for Person (i.e. it is 3rd person and possibly –HUMAN), it can still be overtly marked as ±FEMININE.

(12) ***Entry for Subject clitics (preliminary version):***
+___I, D, { PER { 1st, *je* / 2nd, *tu* / *on*}, / FEM, *elle* } / *il*

(13) (a) Jean et moi, on [$_{i}$ [$_{V}$ ne va]] jamais au cinéma.
John and me, one not goes ever to the movies
"John and I never go to the movies."

(b) Cette femme, est-ce qu'elle t'aime?
that woman is it that she you loves
"That woman, does she love you?"

The entry (12) is to be read thus: "D can be a prefix on I, with possibly a marked feature of either Human (PERSON) or Feminine." Since a D with such features is uninterpretable in the I position, it must "alternatively realize" these features of the nearest interpretable D. In particular, such clitics "double" a separate full DP in subject position (De Cat 2002). The doubled features of clitics, unlike French agreement features, suffice to license null subject DPs. That is, if a clause has a subject clitic on I, its lexical DP subject can be null, as in fact exemplified in (15) below.[7]

As for the SF–PB–AR–BC properties: Entry (12) crucially uses the brace notation (PB). This entry in the French grammatical lexicon also conforms to Borer's Conjecture, because e.g. Standard Spanish and Italian lack subject clitics, making the entry language-particular.

6 There is no need in the system of this essay to specify the nature of this attachment.

7 Note that the alternatively realized subject clitics involve *no movement* and *no case feature*.

In accord with Single Feature Exponents (SF), every exponent in (12) is listed with at most one feature not shared by less marked morphemes. Previous treatments take for granted that the French subject clitic paradigm is unpredictably skewed and asymmetric. E.g. the second person clitic has no special feminine form. But the formulation of entry (12) shows rather that the subject clitic system is not skewed, but is rather a perfect example of Single Feature Exponents (7).

I now introduce the general formal statement of AR.

(14) **Alternative Realization**. A feature F of an interpreted closed class item α can also be phonologically realized under a γ^0 outside α^0, provided that some projections of α and γ are sisters.

In the AR for subject clitics, (i) the F are the person and gender features of D in (12), (ii) α^0 must be a D head of a DP so that the F can be interpreted, and (iii) γ^0 = I. The subject DP and I' are the only projections of α^0 and γ^0 that satisfy the condition that both are sisters.[8]

A French V can move over the subject clitic in questions, since V itself can move from I to C (Roberts 2010, Chap. 3), as follows:

(15) (a) Tu – $[_I[_V$ ne – la – vois]] jamais.
you – not – her – see never
"You never see her."

(b) $[_C[_V$ Ne – la – vois]] – tu jamais?
not – her – see – you never
"Don't you ever see her?"

In order for this I to C movement to work, pre-verbal object clitics should be left sisters of V, and then grouped with V as a derived V, as seen in the bracketing in (15). The subject pro clitics do not move with V to C.[9]

8 A reviewer insists that e.g. clitics are "interpreted," and in a pre-theoretical sense they are. The claim here is that their LF interpretation arises formally because of their link to canonical positions. Thus, *la* "her" in (15a) is interpreted as a direct object exactly like *her* in *I see her.* Only their PF positions are different.

9 The fact that French clitics spell out both under I and under V does not mean that cliticization is "two processes," but only that it occurs in positions specified differently in two lexical entries. The difference is motivated by I to C movement, ordering relative to *ne* "not," and the ability of the types to freely co-occur.

3.2 Missing Exponents in Context Features

The preliminary Subject clitic entry (12) does not specify any exponents for the plural subject clitics *nous* "we" and *vous* "you." The French grammatical lexicon avoids such separate specifications by having non-clitic "strong forms" of Plural Person pronouns do double-duty as bound forms. This can be formally expressed in terms of a general convention that increases the Economy of lexical entries:

(16) **Missing Exponents.** If a lexical entry for bound morphemes lacks an exponent for a given set of specified syntactic features, their exponent of the free form is used.

To reflect this convention, the entry for subject clitics requires revision. This revision provides no exponents for the sets [D, PER, 1st/2nd, PL], because these can be (and are) spelled out with their free morpheme exponents *nous* and *vous*. We can also use (16) to dispense with spelling out *elle* "she" in two different entries.

(17) ***Entry for all Subject clitics***. (Recall, A/B means "A or B," and not both.)

$$+\_\_\_\text{I, D, } \left\{ \begin{array}{l} \text{PER, } \{ \text{ 1st, } \{ \textit{je} \text{ / PL } \} \text{ / 2nd, } \{ \textit{tu} \text{ / PL} \} \text{ / } \textit{on} \} \\ \qquad\qquad\qquad \textit{il} \text{ / FEM} \end{array} \right\}$$

This entry is thus piggy-backing on the separate general entry for plural Person pronouns that are free forms:

(18) ***Plural Person Pronouns:*** D, PER, PL, {1st, *nou-* / 2nd, *vou-* }

I do not include the final segment *-s* in this entry because *all* French plural Ds, including clitics, are followed by a morpheme *-s*, more precisely { *-z-,* +___vowel/ *Ø* elsewhere }. I take it that this morpheme, so-called liaison, is a separate formative in representations of both strong and clitic pronouns, possessives, demonstratives, etc. I do not specify its full lexical entry here.[10]

4. The Pro-PP Clitics of French (and Italian)

The most basic Prepositions of French are:

(19) (a) *à* "to/at," interpreted as: Static Location / towards a Goal / or semantically empty.

(b) *de* "of/from," interpreted as: Possession / from a Source / or semantically empty.

10 The only exception, a more specific form which indeed blocks the appearance of *-s*, is the 3rd person indirect object plural *leur*, which we return to in Section 6.

The minimal general interpretation of the category P is "Location," simply a semantically flavored name for the interpretation of this category. *De* differs from *à* by the marked feature +SOURCE. The common uses of these most basic Ps seem to indicate that if the directional Goal component of meaning is removed, what semantically remains is Static Location in Space/Time, and if the directional Source content is removed, what remains is Possession.[11]

4.1 Clitic Placement of *en* and *y*

French PPs composed of *à* + pronoun and *de* + pronoun can often be replaced by verbal clitics. These adverbial or PP clitics (Kayne 1975, Chap. 2) are *y* "(to) there" and *en* "from/of there." That is, *y* replaces a minimal [P, (GOAL)], and *en* replaces the features [P, SOURCE]. These replacements take place whether or not the P has LF content; sometimes these P serve only to assign case.

Distributionally*, the P pro-clitics immediately precede the verb* and must follow any pronominal clitics. Here are some examples from Veselovská and Vos (1999, 925):

(20) (a) Il y pense souvent.
he to-it thinks often
"He often thinks of it."

(b) Il en a déjà parlé.
he of-it has already spoken
"He has already spoken about it."

Beyond this basic point, several analyses have discussed whether and when *y* and *en* co-occur on one V. The two do readily co-occur in the impersonal existential construction *il y a* "there is," as in (21a). On the other hand, many speakers do not accept non-idiomatic combinations such as (21b).

(21) (a) De bons vins, il y en aura peu ce soir.
of good wines it there thereof have-will few this evening
"Of good wines, there will be few this evening."

11 French and probably Universal Grammar (UG) contain many configurations where P and P, SOURCE lack any locational sense. Thus, unmarked P *à* and *de* often indicate pure possession, and P, SOURCE is used for purely syntactic linking in partitive and pseudo-partitive constructions. In all these uses, the non-locational Ps possibly have no role other than to assign oblique case.

(b) ??Marie y en a déjà parlé.
Mary there thereof has already spoken
"Mary has already spoken of that there."

For capturing the productive usage in lexical terms, I leave aside the idiom *il y a* and propose a single disjunctive entry for these two clitics:

(22) ***Entry for PP clitics.*** +___V, P, { SOURCE, *en / y* }

Let us see how well this entry conforms to the formal characteristics expected in (11):

- SF: Entry (22) clearly conforms to the *Single Feature Exponents* Principle (7);
- PB: Lexical entry (22) uses the *brace formalism* to express a disjunction;
- BC: (22) confirms *Borer's Conjecture*: while Italian has a close counterpart, Spanish lacks PP clitics altogether.
- AR: According to AR, lexical entries specify clitics in their surface positions, and thus express the fact that the PP clitics have no (non-idiomatic) interpretations attributable to their PF positions.

I next go into some detail to compare this PF account of French *en* and *y* in terms of AR (14) with proposals to derive them from some kind of movement.

4.2 *En* and *y* as Phrase-Mates of V

First, I consider whether movement of these P-clitics might be motivated by some non-local "distance" (greater than that allowed by AR) between these clitics and their base or interpreted PP positions. For example in (23), the clitic *en* is linked to empty categories *e* that seem to be *inside* indefinite object DP sisters of its verbal host.

(23) (a) Il en prendra [six litres e] pour sa famille (, de ce vin excellent).
he of it will take six liters for his family (of that excellent wine)

(b) Il en veut [deux e] tout de suite (, de litres de vin).
he of it wants two right away (of liters of wine)

(c) Il en prendra [(beaucoup) e] plus tard (, de votre vin blanc).
he of it will take (a lot) later (of your wine white)

However, long held conclusions that these *en* must "move over" intervening heads of object DPs, i.e., that *en* is not a clause mate of its PP source, are simply wrong. Emonds (2001) shows that the uses of French *en* in (23) (similarly for Italian *ne*) depend

on these languages *independently allowing extraposition* of *de*-phrases to the end of VP. Such overt extraposed PPs are shown in parentheses in (23), making the *e* inside DP the traces of this rightward movement. *Such extraposition is totally excluded in Spanish*. Consequently, Spanish has no counterpart to these PP-clitics, as the locality imposed by AR correctly predicts. Emonds (2001) concludes that French *en* and Italian *ne* never directly "climb" out of direct object projections into the higher VPs. Rather, PP complements of N or Q must first extrapose, and then become available for clause-mate AR by adverbial clitics.

There is also much literature on the possible raising of *y* out of infinitival complements of some grammatical causative verbs, studying for example the contrast in Veselovská and Vos (1999, 1005) between embedded adjunct interpretation as in (24a) and excluded embedded complement interpretation in (24b).

(24) (a) Cela y fera aller Jean.
that there make-will go Jean
"That will make John go there."

(b) *Jean y fera comparer cette sonatine à Paul.
Jean to it make-will compare that sonata to Paul
cf: "Jean will make Paul compare that sonata to it."

Section 7.3 will review argumentation in favor of AR and against any "climbing" in such constructions.

4.2 Comparison of AR with Mechanisms of Movement

In current movement accounts, separate sets of probe features, as in Adger and Harbour (2007, Sect. 4), Roberts (2010, Chap. 3), or Preminger (2014, Chap. 4), are located on functional heads that are always empty, i.e., both phonetically and semantically unrealized (Kayne 1994, 42–46). This in itself would seem to be a formalized expression of "ad hoc" or "redundant," but yet this doubling of features that insure movement (that is, "attraction to probes") is currently so widely accepted that pointing out this problem will probably do little to remedy it.

So let us next consider which feature(s) on PP sources of *en* and *y* might trigger movement (toward functional category "probes" on or above V). These attracted features must be the most basic features of P, as in (19), such as P itself, and/or GOAL, and/or SOURCE. Now the fact is, *other full PPs* e.g. of location *also have these same features*, and yet these are never "attracted" to the probe(s), even when the PPs are single words,

e.g. *dehors* "outside," *là* "there." Movement accounts are silent on how this more general movement is prevented.[12]

Even if theoretical elaboration might circumvent these two general problems, the attracted feature(s) of the P sources of *en* and *y* still cannot be plausibly specified. In the probe-attract framework, the attracted features, unlike those of the probe, are taken to be "interpretable," that is, they have some recognizable content. For example, if a "moved" clitic *en* is attracted to a probing functional head just above V, the only non-ad-hoc candidate for the attracted feature on *en* would be SOURCE. But the problem is, *en* often lacks any SOURCE interpretation, and so cannot be considered to have an "interpretable feature."

Perhaps then some feature other than SOURCE could be assigned to *en* and attracted to its pre-verbal position. But no unified interpretive content, however vague, can be associated with the various uses of *en*:

- *En* can stand for complements of verbs and adjectives introduced by an empty *de* "of": *fier de* "proud of," *loin de* "far from," *remercier de* "thank for," *parler de* "talk about."
- *En* can stand for adjuncts of "place from" with verbs such as *arriver de* "arrive from," *revenir de* "come back from," and *descendre de* "come down from."
- A meaningless *en* can be linked to extraposed complements of (underlined) N, Q, and V, as in (23).
- *En* can be used to indicate existence, in the impersonal idiom *il y a* "there is" (25).

(25) Existe-t-il des bébés qui ne pleurent pas? Oui, il y en a.
"Are there babies that don't cry? Yes, there are."

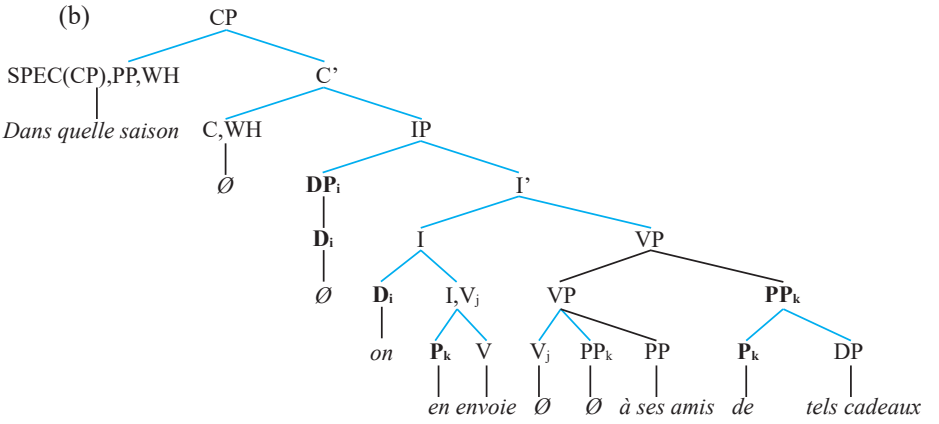
In sum, the great *advantage of AR* for P clitics is that it expresses their compatibility with *any clause mates of V introduced by the Ps à and de.* They are correctly predicted to be unrelated to syntactic or semantic differences among these phrases; the construct of "(un)interpretable" plays no role.

I conclude this section with an example of AR for a subject clitic and a P-clitic, the two types I have so far discussed and written lexical entries for. The pairs of AR morphemes and their sources in LF are in bold. P_k satisfies AR *before* V_j raises to I.

12 Of course, the same problem arises with any "movement" limited to any subsets of closed class items, such as English finite copula raising, "affix movement," or any variant of "clitic placement." This is why AR rather than movement should be used to account for all of them (Emonds 2000, Chap. 4).

(26) (a) Dans quelle saison, on en envoie à ses amis, de tels cadeaux?
In which season, one thereof sends to one's friends, of such presents
"In which season does one send to one's friends such presents?"

(b)



For justifying V_j movement to finite I, see Emonds (1978).

5. Object Clitics Expressing Person

The French clitic system treats the following pronouns as a special group: the 1st and 2nd person pronouns, a reflexive *se* and an indefinite subject clitic *on.* To express this, I have adopted Kayne's (2000) proposal that they all realize a marked (but perhaps unvalued) feature PER not shared by non-reflexive 3rd person pronouns. Any pronoun specified for PER is always +HUMAN.

A fundamental fact about the person proclitics is that they always precede object clitics and the P clitics analyzed in Section 4.[13] For the latter combination, see again (3).

(27) (a) Elle nous l'expliquera en français.
she us it explain-will in French
"She will explain it to us in French."

13 Curiously, the ordering in (27) is reversed in Standard French affirmative imperative enclitics:

(i) Explique le nous en français!
"Explain it to us in French!"

(ii) Ces vers, répétez les vous chaque fois que vous pensez à lui.
"These verses, repeat them to yourselves each time that you think of him."

(b) Ces vers, vous devriez vous les répéter chaque fois que vous pensez à lui.
these verses you should you them repeat each time that you think of him
"These verses, you should repeat them to yourself each time that you think of him."

(c) Anne s'en disait très fière.
Anne herself thereof said very proud
"Anne said herself to be very proud of it."

The Person clitics are notably unspecified with any kind of case-like feature which might indicate their grammatical relation to the verb, though their non-subject status is indicated by their context feature +___V rather than +___I. The following preliminary entry specifies their singular forms.

(28) ***Object clitics of Person*** (preliminary)**.** +___(P) V, PER, { 1st, *me* / 2nd, *te*, / *se* }

The feature content of *se*, namely the single feature [PER], can be taken as "unvalued for a specific person." I assume that Universal Grammar requires that it have a clause-mate antecedent if it can (i.e. *se* is an alternatively realized bound anaphor). Analogously, the same feature PER characterizes the subject clitic *on.* In support, note that *se* is the reflexive object required by a subject *on*, and hence must share its features: *Dans cette famille, on se critique rarement.* "In that family, one criticizes oneself rarely." Not accidentally, in Italian and Spanish, which lack subject clitics, this spelling of PER in the context +___V is uniformly *si/se,* even when this *si/se* translates the French indefinite subject *on.*

Like subject clitics, the plural PER clitics have the same exponents as their free morpheme counterparts, namely *nous* "us" and *vous* "you." As with plural subject clitics, Missing Exponents (16) exempts these forms from being repeated in the lexical entry for the clitics in the context +___V. The strong forms *nous*/*vous* occur equally well as free forms, bound subject clitics, and bound object clitics.

(29) ***Object clitics of Person*** (final)**.** +___(P) V, PER, $\left\{ \begin{array}{ll} \text{1st,} & \{me \text{ / PL}\} \\ \text{2nd,} & \{te \text{ / PL}\} \\ & se \end{array} \right\}$

Entry (29) alternatively realizes *any direct or indirect object pronoun with the feature* PER as a verbal proclitic. Since the single feature PER can be spelled out only as *a single clitic*, the well-known ban on two co-occurring non-subject Person clitics follows.[14]

14 When both direct and indirect objects are +PER, the indirect object must usually be in a PP: thus, **Marie m'a présenté vous.* "Mary introduced you to me." As observed in Section 1.1, however, under certain conditions these positions can be reversed.

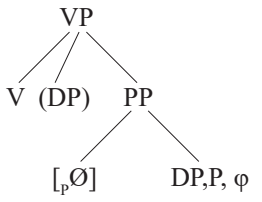
As French traditional grammar has long recognized, the clitic ordering PER – P – V does not reflect or derive from any "direct – indirect" object order or from the fact that adjuncts follow complements in phrasal syntax. When the clitic *en* represents a partitive or indefinite direct object, it still follows indirect object and "dative of interest" adjunct clitics.

6. Third Person Indirect Object Clitics

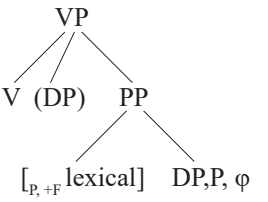
6.1 Alternatively Realized Indirect Objects

Indirect objects are DPs case-marked by a P that has *no other feature*. These P are empty at Spell Out and so allow their object to formally "be a" sister of V.[15] As a result, indirect object closed class DPs (pronouns) can potentially be alternatively realized under V (30a). If, however, a P has additionally a content feature F, the P is spelled out rather than null and hence visible in Logical Form, so that its own object DP is not the only spelled out daughter of PP (30b). This DP thus fails to satisfy the defining sisterhood condition on AR in (14), and as a result this DP sister of a lexical P cannot be alternatively realized as a clitic on V, a well-known generalization.

(30) (a)



(b)



Rephrasing: the P on the DPs is abstract oblique Case. In (30a), the oblique DP "is a" sister of V, so its features [P,φ] can be alternatively realized under V, as a verbal clitic. In (30b), because of the lexical P, the oblique DP is *not* a sister of V, so its features cannot be alternatively realized under V.

French indirect object DPs include PER clitics. Besides these, which lack case features, indirect objects can also be the 3rd person non-reflexive clitics *lui* (singular) and *leur* (plural), both unmarked for Gender.[16] As Definite (and otherwise unmarked) appears with all indirect object pronouns, the *single lexical feature* of these exponents, in conformity with (7), is an alternatively realized "case-feature" P.

15 This formalism is from Chomsky's early work, and is further explained in Emonds (1999). The main idea is that A "is a" B if all the phonological material under B is also under A.

16 Though both these forms are homophones with free pronouns (*lui* = free strong form "him"; *leur* "their"), I take only the former as related to the lexical entry of the clitics The clitic *lui*, like the strong and clitic homophones *nous*, *vous*, and *elle*(*s*), is the same as the free form, except for lacking +FEMININE.

Like other object clities, *lui* and *leur* must *precede* the P-clitics *y* and *en,* as in (32a, c). The optional P in (31) allows these P-clitics between *lui/leur* and the latter's verbal host. The opposite ordering in (32b, d) is excluded by the immediate adjacency of P clitics to V mandated by (22).

(31) ***Indirect object clitics*** (tentative). +___(P)V, DEF, P, {PL, *leur* / *lui*}

(32) (a) Luc lui en a parlé.
Luke to-him of-it has spoken

(b) *Luc en lui a parlé.

(c) Elle leur y donne des sous souvent.
she them there gives money often

(d) *Elle y leur donne des sous souvent.

Another often puzzling property of the indirect object clitics is captured by the parentheses notation in (31). In particular, when P clitics are present, the *longest insertion context* ___*P-V in* (31) *must* be chosen, as argued in Chomsky and Halle (1968). This yields e.g. *leur–y–V* and *lui–en–V*, while correctly excluding **y–leur–V* and **en–lui–V.*

Because of the Missing Exponents Convention (16), the entry for indirect object clitics need not spell out *lui*, since *lui* also serves as the free form for unmarked (masculine) third person pronouns. Moreover, while the strong form pronoun has a marked Feminine counterpart (*elle*), this feature is not alternatively realized on the clitic. We thus arrive at the following revision:

(33) ***Indirect object clitics.*** +___(P)V, DEF, P, (PL, *leur*)

If the parenthesized Plural is not chosen, then a singular "dative" definite pronoun *lui* can be alternatively realized in the context ___(P)V, i.e., on a verb with or without an adverbial proclitic.

6.2 The Person-Case Constraint (PCC)

A much discussed restriction of French grammar, noted as a problem in Perlmutter (1971), is that a direct object *Person clitic cannot co-occur with indirect object clitics such as lui/leur.* Other Romance languages exhibit similar restrictions, though language-particular details differ. Rivero (2004, 498) provides this Spanish example:

(34) Ella se le entregó cuerpo y alma.
she herself him gave body and soul
"She gave herself to him body and soul."

Due to the PCC, the French counterpart is ungrammatical: **Elle se lui est livrée corps et âme*. Cardinaletti (2008, Sect. 3.2 and 4.3) also provides Italian examples in which both a verb's direct and indirect objects are Person proclitics, whose exact French counterparts (with *me te V*) are ungrammatical.[17]

Several studies since Anagnostopoulou (2003) and Béjar and Řezáč (2003) have accounted for the PCC in terms of some restriction on probe features originating on Verbs or functional heads adjacent to V. For example, Adger and Harbour's account of the PCC (2007, Sect. 5) is a construction-particular restriction on how uninterpretable features on empty functional heads fh^0 can search for their interpretable counterparts (for them, fh^0 is the Applicative Phrase head that unites direct and indirect objects in a single constituent). It is hard to imagine other grammatical phenomena that might serve to independently justify such a highly particularized restriction (or help a child to learn it). So I remain unconvinced by these attempts to use UG to account for language-particular restrictions, which at the same time leave aside the lexical statements required to make Borer's Conjecture into something more than a vague statement of belief.

Instead, I propose that a single lexical entry with braces expresses the French *complementary distribution* of Person clitics with a (3rd person) indirect object clitic (33). This is a crucial and yet maximally simple use of the brace notation, the main formal device that expresses "A or B but not A and B."

(35) ***Person and indirect object clitics*** (automatically expresses the Person-Case Constraint).

$$+\_\_\_(P)V,\ DEF,\ \left\{ \begin{array}{l} PER,\ \{1st,\ \{me\ /\ PL\}\ /\ 2nd,\ \{te\ /\ PL\}\ /\ se\} \\ P,\ (PL,\ leur) \end{array} \right\}$$

I assume that in a given context, here ___(P)V, a lexical entry can be used only once. By virtue of the *brace notation* in this entry, it is impossible to have *simultaneous AR* both of a Person DP and an indirect object DP as verbal clitics. Entry (35) thus easily expresses the French PCC and its language-particular character (BC).[18]

17 Cardinaletti (2008, Sect. 7) presents other counterexamples to the PCC from Old Italian, which she attributes to clitic orders. French excludes a Person clitic with any 3rd person indirect object clitic.

18 The same disjunction holds for the somewhat differently ordered enclitics in affirmative imperatives.

7. Third Person Direct Object Clitics

French direct object clitics have the same form as definite articles: Fem Sg *la*, Plur *les*, and unmarked "Masc Sg" *le*. When they appear as definite pronoun object clitics, they also precede the PP clitics *en* and *y*. French direct object clitics should thus *alternatively realize* as prefixes on V the free morphemes for the feature DEF.

(36) ***Direct object clitics.*** +___ (P)V, DEF, ({FEM/ PL})

Due to the Missing Exponents Convention (16), this entry need not stipulate any bound form exponents for the clitic, because these exponents are precisely those specified for free form definite articles in their base or interpretable position.

(37) ***Definite article entry.*** DEF, {(FEM, *la*) / (PL, *les*) / *le*}

Alternatively the parenthesis notation should perhaps extend to contexts in lexical entries, which would allow us to economically combine (36) and (37):

(38) ***Definite article entry (extended).*** DEF, (+___ (P)V), {(FEM, *la*) / (PL, *les*) / *le*}

7.1 The Ordering of Third Person Direct Objects and Person Clitics

Using only the lexical entries formulated so far, these two groups would not co-occur, since insertion of either in the context ____(P)V would remove the adjacency required for the subsequent insertion of the other.

In fact, only one of these two orders is grammatical, namely PER – DEF – (P) – V:

(39) (a) Des garçons me les ont apportés hier.
some boys me those brought yesterday
"Some boys brought me those yesterday."

(b) Je vais vous la décrire.
I will you her describe
"I will describe her to you."

(c) * Des garçons les m'ont apportés hier.
* Je vais la vous décrire.

We can accommodate this ordering by treating the third person direct object clitics as infixes in the clitic sequence. If Person clitics are inserted, they become part of the "longest context" in (40), so that *le*, *la*, *les* can only be inserted on their right.

(40) ***Definite articles.*** DEF, ((PER)___ (P)V), {(FEM, *la*) / (PL, *les*) / *le*}

There is no way to generate the examples in (39c) because the alternatively realized Person clitics are not specified with a (dative) P feature to serve as a right context for the direct object clitics.

7.2 The Ordering of Third Person Object Clitics

Many previous analyses, spanning Kayne (1976) and Adger and Harbour (2007), have sidestepped specifying ordering among clitics. In contrast, the language-particular lexical entries of the present study succinctly account for clitic ordering.[19] For example, two long standing formal puzzles have been, what accounts for the ordering contrast *le–lui, la–leur* etc. vs. **lui–le, *leur–la* (41), as well as the marginality of three clitic sequences (3)? (They are OK for some, * for others.)

(41) Marie la leur donne le samedi. *Marie leur la donne le samedi.
Mary it them gives the Saturday
"Mary gives it to them on Saturday."

In fact, the content of the entries formulated so far, (33) and (36), and the lexical PB notations themselves have already answered these questions.

(i) If a direct object clitic *le*, *la*, *les* is first inserted in ___(P) V, yielding ___DEF–(P)–V, the insertion context for an indirect object clitic is no longer satisfied, so no combination can result.
(ii) If the indirect object clitic *lui*/*leur* is first inserted in ___(P) V, this yields ___P–(P)–V, into which direct object clitics can be inserted, yielding the correct *le–lui/leur–V.* (The direct object cannot appear in P___V because a longer context must always be chosen).
(iii) Some speakers may be able to interpret a sequence P-P-V as satisfying ___P–V, yielding e.g. *le leur y donne* "give it to them there." Other speakers interpret the context feature more strictly, excluding this last phrasing.

19 As a number of studies observe, clitic orders do not reveal anything about the complement vs. adjunct status of the phrases they spell out. For example, the Person clitics that precede French *le/la/les* can express either adjuncts ("datives of interest") or complements, and the adverbial clitics *y* and *en* can also realize either adjuncts or complements; for the latter, see (20) and (23).

As indicated at the end of Section 3, the plural *-s* that follows both clitics and articles is probably a sequentially separate and independent morpheme and hence not part of the entries for clitics.

This AR of P on both *lui/leur* and on P-clitics thus means that direct object clitics that lack P must precede them all, due to the requirement that a longer context feature have precedence; ___P – V is longer than ___V. These considerations together yield correct sequences and exclude sequences such as **lui–le–V* and **en–le–V*.[20]

7.3 No Clitic Climbing

Many generative studies, beginning with Kayne (1975), have proposed that clitics can raise out of phrases where they originate into higher clauses. Emonds (1999) undertakes a full critical investigation of four different types of putative "climbing" of Romance clitics:

(42) (i) obligatory raising of clitics to auxiliary verbs,
(ii) optional raising of clitics out of restructuring and causative infinitives,
(iii) clitic movement of *en/ne* out of object nominal phrases, as exemplified in (23),
(iv) clitic movement out of adjective phrases to their selecting verbs, as in (43).

(43) (a) Paul en semblait très fier.
Paul thereof seemed very proud
"Paul seemed very proud of it."

(b) Je lui suis reconnaissant.
I him am thankful
"I am thankful to him." (Veselovská and Vos 1999, 1008)

That investigation provides arguments for replacing all clitic climbing with analyses using Alternative Realization, in accord with arguments based on the empirical paradigms of these clitics.

(44) **Phrase Mate Hypothesis**. Romance clitics on V_i are related to only XP sisters to some projection V^k of V_i. (Emonds 1999, 314)

I thus claim that counter to "climbing," clitics on a verb arguably realize only phrasemates of that verb. Thus in an Italian restructuring sequence, when a clitic on a main verb, underlined below in Rizzi's example, realizes an object of the verb's infinitive complement, *no VP node intervenes* between the matrix VP and the phrasal complements of that infinitive (Rizzi's conclusion in 1978, Section 7.1). That is, *no separate VP* is

20 A reviewer asks is there is some deeper reason for why "longer contexts win out." I can only refer the reader to the rather strong defense of this convention in Chomsky (1967).

comprised of the sequence *parlare al piu presto*; the two verbs, of which the first must be in a closed class, are in a single maximal VP.[21]

(45) Questi argumenti, dei quali $\underline{t_i}$ verró a parlare $\underline{e_i}$ al piu presto, . . .
these topics of which you come-will-I to talk at most soon
"These topics, about which I will come to talk to you as soon as possible, . . ."

According to further argument in Burzio (1986), Italian and French infinitive complements of a closed class of causative and perception verbs, which exhibit similar "raised" clitics, have the same structure as Italian restructuring verbs. Thus, in at least one structure for these French constructions, *the two underlined complements of the second verb in (47) are in fact also sisters of the bold first one.*

(46) (a) Marie a vu distribuer les prix aux étudiants par le propriétaire.
Mary has seen distribute the prizes to-the students by the owner
"Mary has seen the prizes distributed to the students by the owner."

(b) Marie a fait planter des fleurs dans mon jardin.
Mary has made plant some flowers in my garden
"Mary has made someone plant some flowers in my garden."

On the basis of these structures, the Phrase Mate Hypothesis (44) predicts that all French cliticization of complements and adverbials on verbs selecting infinitives, as exemplified in bold face in (47), should be subsumed under AR. The verbs, containing AR clitics, then raise to finite I, as in Section 3.

(47) (a) Marie $[_I$ $\underline{les_i}$ a] **vu**$[_V$ distribuer] $\underline{e}$ aux étudiants par le propriétaire.
Mary to-them-has seen distribute to-the students by the owner

(b) Marie $[_I$ $\underline{y}$ a] **fait** $[_V$ planter] des fleurs $\underline{e}$.
Mary there-has made plant some flowers

(c) Béatrice $[_I$ $\underline{le}$ **fera**] $[_V$ rediger] $\underline{e}$ à l'auteur.
Beatrice it make-will edit to the author
"Beatrice will make the author edit it." (Veselovská and Vos 1999, 997)

21 This implication is part of each of Rizzi's nine arguments for this flat structure, but sometimes slips into the background in his discussions. His final cited section nonetheless makes this structural conclusion crystal clear.

The Phrase Mate Hypothesis (44) for the clitic climbing cases (42) in fact involves nothing more radical than drawing out the natural structural conclusions of the robustly empirical and converging arguments of Rizzi and Burzio.[22]

There remain the possible instances of climbing in (42iii–iv). Section 4.2 here has summarized the argument in Emonds (2001) against the movement of *en.* As for (42iv), verbal clitics can correspond to complements of adjectives (*Paul en semble fier* "Paul thereof seems proud"). This is because, in French as well as English, phrasal complements to adjectives can optionally appear as sisters of the selecting verb rather than of the adjective itself: *How proud he seemed of his medal.* As sisters to this selecting verb, these complements can be alternatively realized by this verb's phrase-mate clitic *en*. For more detailed presentations of these arguments against all the proposed sub-cases of clitic climbing, see Emonds (1999).[23]

7.4 The Default Use of the Clitic *le*

The unmarked French verbal clitic *le* in (38) has an important use whereby it can replace obligatory complements of *any extended projection* other than DP, namely IP, CP, NP, PP, and AP. It is thus a *default AR* in the context ___(P)V of any phrase XP. This *le* never varies in person, gender or number, even when its antecedent is an NP:

(48) (a) On dit que Henri est coupable, mais je ne le pense pas. (*le* replaces IP)
One says that Henry is guilty, but I-not-it-think not
(". . . , but I don't think so.")

(b) Marie m'assure qu'elle est fiable, mais je me le demande. (*le* replaces CP)
Mary assures me she is reliable, but I-me-it-wonder
(". . . , but I wonder about it.")

(c) Anne est institutrice, et Marie et Françoise le sont aussi. (*le* replaces NP)
Ann is teacher and Mary and Frances it-are too

22 These well supported conclusions squarely contradict the generative literature motivated by imposing a priori binary branching and small clause structures. The gap between this literature and the predictive power of the Phrase Mate Hypothesis should contribute to not letting theoretical preferences override the data, and to asking instead why so often the former have such a tenacious hold.

23 The system here retains transformational Head-to-Head Movement. However, it cannot duplicate AR by moving D or P to functional heads in a verbal projection. Head Movement is limited to moving (all) items of a given category under the condition that *a head β^0 can have the landing site α^0 only if α^0 and β^0 are heads in the same extended projection.*

(d) Anne est de bonne famille, et Marie et Françoise le sont aussi. (*le* replaces PP[24])
Ann is of good family, and Mary and Frances it-are too

(e) Tu es plus content que nous les femmes le sommes. (*le* replaces AP)
You are more satisfied than we the women it-are

This default usage of *le* suggests that category in (40) should be "underspecified" as in (49), where X stands for any phrasal head category D, N, I, V, A, or P.

(49) ***Definite articles*** (final)**.** X, DEF, ((PER)___ (P)V), {(FEM, *la*) / (PL, *les*) / *le*}

If any non-contextual features of the Definite Article in (49) are used when choosing this entry, they ensure that X = D, since these features are specified in this entry only in combination with DEF.[25]

8. Inventory of Final Lexical Entries and Conclusions

The analysis of French verbal proclitics in this study has exploited Lieber's (1980) word-internal subcategorization, the parentheses and brace notation (PB), and the principle of Single Feature Exponents (7). The lexical entries proposed here parsimoniously express ordering and other restrictions within clitic sequences. Moreover, the principle of Alternative Realization (14) correctly limits the structural distance between the interpreted and pronounced positions of clitics to single structural clauses, in accord with the empirically supported Phrase Mate Hypothesis. Two of the entries, for PP and subject clitics, are clearly language-particular, consistent with Borer's Conjecture. And even though the French entry (49) is basically the same as in Spanish, it is certainly language-particular (BC), being found rarely if at all outside Romance. Similarly, the verbal proclitic positions of (35) are a marked language-particular option.

(17) ***Entry for all Subject clitics***.

$$+\_\_\_\text{I, D, } \left\{ \begin{array}{l} \text{PER, } \{ \text{1st, } \{ \textit{je} / \text{PL} \} / \text{2nd, } \{ \textit{tu} / \text{PL} \} / \textit{on} \} \\ \textit{il} / \text{FEM} \end{array} \right\}$$

(22) ***Entry for PP clitics.*** +___V, P, { SOURCE, *en / y* }

24 The more specific entry for a locational pro-PP, namely +__V, P, *y,* blocks using *le* as a default pro-PP for physical location.

25 If the unmarked combination X, DEF, *le* is chosen in non-clitic position, it is in a canonical (= LF) position, and so X, DEF requires interpretation. Then, if X ≠ D, the feature DEF is undefined, rendering the structure containing [X, D] ill-formed (i.e. it violates Full Interpretation).

(35) ***Person and indirect object clitics***
(subsumes the PCC)

$$+\_\_\_(P)V,\ DEF,\ \left\{\begin{array}{l} PER,\ \{1st,\ \{me\ /\ PL\}\ /\ 2nd,\ \{te\ /\ PL\}\ /\ se\} \\ \quad\quad\quad\quad\quad\quad P,\ (PL,\ leur) \end{array}\right\}$$

(49) ***Definite articles*** (final)**.** X, DEF, ((PER)___ (P)V), {(FEM, *la*) / (PL, *les*) / *le*}

The results of this study in terms of formal grammar is that the distribution of French clitics in all constructions (in so-called causatives, auxiliaries, pseudo-partitives, etc.) reduces to these four lexical entries.

There is quite a notable difference between this essay and much other work which nominally adheres to Borer's Conjecture (the claim that the functional category lexicon is the sole source of language-particular grammars). Very few of these studies actually formulate any results in terms of the explicit language-particular lexical entries or parameters that the Conjecture calls for. Most give no hint, much less justification, of how possibly related groups of language-particular morphemes (such as the c. 20 French verbal clitics) appear in such entries; this task is left to the side as somehow not central to the generative enterprise. Here in contrast, I have formulated four such entries. These formulations eliminate syncretism and redundancy, and all conform to four plausible universal principles of Lexical Economy: (i) Missing Exponents, (ii) Alternative Realization, (iii) Parenthesis and Brace notation, and (iv) Single Feature Exponents.

This contrast results from the fact that most generative syntax has forgotten the methodological motivation for postulating Universal Grammar. Namely, the seemingly impossible task of children quickly and flawlessly learning the highly complex system of a given language was greatly simplified—they need only learn the residues of particular languages $\{R_i\}$ that are not part of UG. But after a sharpening of this proposal, namely Borer's Conjecture, a strange thing happened. Although most Chomskyans quickly accepted the Conjecture without argument, almost no contentful proposals emerged for modelling these R_i. Even as studies of UG flourished, whose purpose was to greatly simplify $\{R_i\}$, very few studies actually formalized the latter. Such has been the state of grammatical affairs that this essay has tried to remedy. Without fragments of formalized language-particular grammatical lexicons, such as the example set in Ouhalla (1991), research in Universal Grammar is losing its empirical footing.

Works Cited

Adger, David, and Daniel Harbour. 2007. "Syntax and Syncretisms of the Person Case Constraint." *Syntax* 10 (1): 2–37.

Anagnostopoulou, Elena. 2003. *The Syntax of Ditransitives: Evidence from Clitics.* Berlin: Mouton de Gruyter.

Anderson, Stephen. 1982. "Where's Morphology?" *Linguistic Inquiry* 13: 571–612.
Béjar, Susana, and Milan Řezáč. 2003. "Person Licensing and the Derivation of PCC Effects." In *Romance Linguistics*, edited by A. T. Perez-Leroux and Y. Roberge, 49–62. Amsterdam: John Benjamins.
Borer, Hagit. 1984. *Parametric Syntax: Case Studies in Semitic and Romance Languages*. Dordrecht: Foris.
Burzio, Luigi. 1986. *Italian Syntax: A Government and Binding Approach*. Dordrecht: Reidel.
Cardinaletti, Anna. 2008. "On Different Types of Clitic Clusters." In *The Bantu-Romance Connection: A Comparative Investigation of Verbal Agreement, DPs and Information Structure*, edited by Cécile de Cat and Katherine Demuth, 41–82. Amsterdam: John Benjamins.
Chomsky, Noam. 1967. "Some General Properties of Phonological Rules." *Language* 43 (1): 102–28.
Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English.* New York: Harper and Row.
De Cat, Cécile. 2002. *French Dislocation.* PhD diss., University of York.
Embick, David, and Ralph Noyer. 2001. "Movement Operations after Syntax." *Linguistic Inquiry* 32: 555–95.
Emonds, Joseph. 1999. "How Clitics License Null Phrases: A Theory of the Lexical Interface." In *Empirical Approaches to Language Typology: Clitics in the Languages of Europe*, edited by H. van Riemsdijk, 291–367. Berlin: Mouton de Gruyter.
Emonds, Joseph. 2000. *Lexicon and Syntax: The English Syntacticon.* Berlin: Mouton de Gruyter.
Emonds, Joseph. 2001. "La relation entre la Dislocation à Droite et le clitique franco-italien *en/ne*." *Journal of the Linguistic Society of Japan* 119: 1–32.
Humboldt, Wilhelm von. (1822) 1997. "On the Origin of Grammatical Forms and Their Influence on the Development of Ideas." In *Essays on Language*, edited by T. Harden and D. Farrelly. Frankfurt: Peter Lang.
Kayne, Richard. 1975. *French Syntax.* Cambridge, MA: MIT Press.
Kayne, Richard. 1994. *The Anti-symmetry of Syntax.* Cambridge, MA: MIT Press.
Kayne, Richard 2000. "Person Morphemes and Reflexives in Italian, French, and Related Languages." In *Parameters and Universals*, edited by Richard Kayne, 131–62. Oxford: Oxford University Press.
Lieber, Rochelle. 1980. *On the Organization of the Lexicon*. PhD diss., MIT.
Ouhalla, Jamal. 1991. *Functional Categories and Parametric Variation*. London: Routledge.
Parrott, Jeffrey. 2009. "Danish Vestigial Case and the Acquisition of Vocabulary in Distributed Morphology." *Biolinguistics* 3 (2–3): 270–304.
Perlmutter, David. 1971. *Deep and Surface Structure Constraints in Syntax.* New York: Holt, Rinehart and Winston.

Preminger, Omer. 2014. *Agreement and Its Failures.* Cambridge, MIT: MIT Press.
Rivero, Maria-Luis. 2004. "Spanish Quirky Subject, Person Restrictions, and the Person-Case Constraint." *Linguistic Inquiry* 35 (3): 494–502.
Rizzi, Luigi. 1978. "A Restructuring Rule in Italian Syntax." In *Recent Transformational Studies in European Languages*, edited by S. J. Keyser. Cambridge, MA: MIT Press.
Roberts, Ian. 2010. *Agreement and Head Movement: Clitics, Incorporation and Defective Goals.* Cambridge, MA: MIT Press.
Veselovská, Ludmila, and Joseph Emonds. 2016. "Morphology, Divided and Conquered?" *Linguistica Brunensia* 64 (1): 143–62.
Veselovská, Ludmila, and Riet Vos. 1999. "Clitic Questionnaire." In *Empirical Approaches to Language Typology: Clitics in the Languages of Europe*, edited by H. van Riemsdijk, 891–1009. Berlin: Mouton de Gruyter.

Syntactic Derivations

Multiple Wh-structures in Hungarian: A Late Insertion Approach

Mark Newson[a] and Márton Kucsera[b]

[a]Eötvös Loránd University, Budapest, Hungary, and the Research Institute for Linguistics, the Hungarian Academy of Sciences;
[b] Eötvös Loránd University, Budapest, Hungary

[a]newson.mark@btk.elte.hu; [b]kucseram@inf.elte.hu

Abstract: The paper examines Hungarian multiple wh-structures with a quantified interpretation. After pointing out problems with the "reinterpretation" approach of earlier analyses, it is argued that late insertion provides a more straightforward explanation for the data. It is proposed that an underlying universal quantifier may surface as a wh-element if it is required to type a clause. It is further argued that clausal typing depends on scope interpretation. It is shown that the Subset and the Superset Principles for vocabulary insertion are both insufficient in this case; therefore, the paper argues for the adoption of Targeted Underspecification to account for both the under- and overspecification in the insertion process.

Keywords: interrogatives; quantification; late vocabulary insertion; Hungarian

1. Introduction

The present paper discusses the analysis of Hungarian multiple wh-constructions that involve universal quantification.[1] After the relevant data are introduced, the analyses of É. Kiss (1993) and Lipták (2000) are presented and it is shown that their assumption that wh-phrases can be reinterpreted as universal quantifiers (the WH-assumption) faces a number of problems. We propose a late insertion approach, which assumes that the

1 The authors wish to thank the audience at Olinco 2016 for their very helpful comments. We also thank Marcel den Dikken and Krisztina Szécsényi, whose comments on the written version of this paper have helped to further refine it. Remaining mistakes are ours.

wh-phrases are universal quantifiers (the Q-assumption). This avoids all the problems facing the WH-assumption. We argue that the realization of quantifiers as wh-phrases is motivated by a clausal typing requirement which targets wide scope operators. Finally, the method of vocabulary insertion is examined; there, we argue in favor of the Targeted Underspecification approach of Newson (2014), which allows for both the under- and overspecification of input elements.

2. Hungarian Data

Hungarian multiple wh-structures are categorized into two types according to whether or not they involve quantification. First, there are cases with multiple "real" wh-phrases, each of which carries an interrogative force, as shown in (1).

(1) Ki látott mit?
who-NOM saw what-ACC
"Who saw what?"

In this example, a single answer is required. The interpretation is that the speaker knows that some person *x* saw some object *y* and wants to know the identity of this <*x*, *y*> pair. The important point here is that there is a single such pair that is relevant to the discussion; therefore, none of the wh-phrases are interpreted as a quantifier over the elements of a set.[2] In the rest of the paper, we will not be concerned with this type of multiple wh-structure as they do not involve a quantifier interpretation.

The second type of Hungarian multiple wh-structures involves some form of quantification. There are two main varieties of such constructions: pair-list interrogatives and multiple relatives.

In pair-list questions, only the last wh-phrase is interpreted as an actual interrogative; all the others function as quantifiers over the set of entities considered relevant. An example for such a construction is given in (2).

(2) Ki mit vállalt?
who-NOM what-ACC undertook
"For every person in question, what did that person undertake?"

2 É. Kiss (1993), amongst others, claims (1) to be ungrammatical. However, we have found that Hungarian speakers more readily accept them when given an appropriate context. We assume that this is because questions which request more than one piece of information are difficult to process and thus require a lot of contextual support.

For (2), the answer is an exhaustive list consisting of pairs of people and tasks undertaken.

If there are more than two wh-phrases, only the one closest to the verb is a true interrogative. All the others are quantifiers.

(3) Ki kinek mit adott?
who-NOM who-DAT what-ACC gave
"For everyone x and y, what did x give to y?"

Here, the answer is a list of <*x*, *y*, *z*> tuples where *x* is in the set of possible agents, *y* is in the set of possible beneficiaries and *z* ranges over the answers to "what did *x* give to *y*?". Therefore, both *ki* and *kinek* serve as universal quantifiers.

Multiple relatives also display the basic pattern shown in (2). In these relatives, the first relative pronoun similarly serves as a universal quantifier applying across a given set and the one closest to the verb functions as an ordinary relative element.

(4) Aki amit talált, megette.
a-who-NOM *a*-what-ACC found ate
"Everybody ate whatever they found."

The main questions posed by these data is why it is possible for wh-elements to seemingly act as universal quantifiers and how this observation can be stated in a more general form and integrated into syntactic theory.

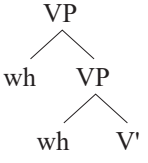
3. Earlier Analyses

3.1 É. Kiss (1993)

É. Kiss (1993) was the first to discuss multiple wh-phenomena in Hungarian as involving quantification. She takes only multiple interrogatives into consideration and argues that both wh-elements are located inside the verb phrase. Her proposal is that the wh-element with interrogative interpretation (i.e., the one closest to the verb) occupies the specifier of the VP. This position, she argues, is where focused constituents in general appear, and the wh-element is focused in single interrogatives.

As there is no upper bound on the number of additional wh-elements that may appear in front of the one with interrogative interpretation, É. Kiss argues that these are adjoined to the VP yielding a structure like (5).

(5)



While it is argued that this adjunction position is one that normal quantifiers occupy, the placement of a wh-phrase in this position does not, in itself, guarantee a quantifier interpretation.[3] Therefore, an additional interpretive rule is needed (É. Kiss 1993, 107):

(6) Interpret a Wh-operator as a distributive universal quantifier if
(i) it has a clause-mate Wh-phrase in its scope, and
(ii) it has a potentially universal force, and
(iii) it is specific.

Much of the complexity in this interpretive rule has to do with specificity effects, with which the current paper is not concerned. All that is important from our perspective is that some such rule is required on the assumption that the operator in the adjunction position is a wh-phrase.

Clearly it is a central tenet of É. Kiss's approach that quantified multiple wh-constructions contain multiple wh-operators, some of which are later reinterpreted as universal quantifiers. We will refer to this as the WH-assumption.

3.2 Lipták (2000)

Lipták (2000) identifies a possible problem with É. Kiss's (1993) analysis. She points out that universally interpreted wh-phrases are dependent on the presence of a wh-operator, which has two consequences.

First, regular universal quantifiers cannot appear preceding a wh-operator.[4]

(7) *Mindig mit vállaltál?
always what-ACC undertook-2SG
"What did you always undertake?"

This generalisation is difficult to account for with an interpretive rule that only targets wh-elements and has no effect on the distribution of actual quantifiers. Also, Lipták notes that in the absence of a wh-operator, a wh-phrase can never be interpreted as a quantifier.[5]

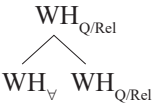
3 É. Kiss does argue that as the wh-phrases not in focus position are not marked for focus, they cannot be interpreted as true wh-phrases. But this does not shed any light on why they are interpreted as universal quantifiers.

4 We will see that this configuration is grammatical, but only if the quantifier has a contrastive topic interpretation.

5 This is stipulated in É. Kiss's interpretive rule (6), in clause (i), but no explanation is given for it.

(8) *Kit János hívott meg.
who-ACC John-NOM invited-3.S PERF
Intended meaning: “For everybody it was John that invited them.”

To account for this mutual dependence between universally interpreted wh-phrases and wh-operators Lipták proposes a different structure for these constructions; she argues that any universally interpreted wh-element is adjoined to a wh-operator as (9) shows.

(9)


As the wh-phrase that is adjoined to will be the one sitting in the relevant wh-position, it follows that this, and no other, will be interpreted as a true interrogative. However, although Lipták gives no details, it is clear that an interpretive rule, similar to that of É. Kiss’s, is also required. Thus Lipták also makes use of the WH-assumption.

3.3 Problems

In this section, we argue that the interpretive rule, necessitated by the WH-assumption, is especially problematic for two reasons.

First, from the standard lexicalist position (which both É. Kiss and Lipták appear to take), the existence of the interpretive process seems particularly odd. A lexical item is a unit with fixed syntactic, semantic and phonological properties. If lexical items are fed into syntactic processes with a certain meaning, it is hard to justify why that meaning should be replaced with something else. This is especially bizarre when the meaning it is supplied with is that which is already associated with another lexical item. If nothing else, this appears to violate basic economy principles assumed to hold of human languages.

Second, the reinterpretation rule is *ad hoc* and has little explanatory power. It gives rise to many questions which are unanswered by its mere proposal. Moreover, it is not at all clear that similar rules are needed elsewhere in the grammar. Essentially what is missing is a restrictive theory of such interpretive rules. As a consequence it gives rise to the possibility that any lexical item could be reinterpreted as any other and therefore completely undermines the lexicalist position that it is built on.

We conclude that the WH-assumption is untenable and therefore another approach is necessary to account for quantified multiple wh-structures.

4. A Different Approach

4.1 Outline of a Solution

The problems facing the WH-assumption are wiped out if we assume that the quantified multiple wh-structures do not actually contain multiple wh-phrases. Instead, we claim

that those wh-phrases with quantifier interpretations are actually universal quantifiers which are realized as wh-phrases. We call this the Q-assumption. From this perspective there are two kinds of phrases with surface "wh" realizations: real wh-phrases which realize underlying wh-operators (represented below as WH) and those wh-phrases which realize underlying universal quantifiers (Q_{WH}). This contrasts with the WH-assumption, in which there is one type of wh-phrase, though sometimes this may be interpreted as a quantifier ($WH_{\forall}$).

To demonstrate the advantages of the Q-assumption, let us consider the distribution of quantifiers and wh-elements under the two assumptions, side by side. First, let us consider the distribution of wh-phrases, as represented in (10) and (11):

(10)		Ki	látott	mit?
		who	saw	what
	WH-assumption	WH/*$WH_{\forall}$	V	WH/*$WH_{\forall}$
	Q-assumption	WH	V	WH

(11)		Ki	mit	látott?
		who	what	saw
	WH-assumption	$WH_{\forall}$/*WH	WH/*$WH_{\forall}$	V
	Q-assumption	Q_{WH}	WH	V

From the perspective of the WH-assumption, the distribution of wh-phrases is rather complex. At least one wh-phrase must immediately precede the verb; others may precede that wh-phrase or follow the verb. Of course, there are conditions on which of these get interrogative or quantifier interpretations and these add to the complexity of observations. Under the Q-assumption, once the quantifiers are factored out, we see that one wh-phrase must precede the verb and all others follow. Not only is this a comparatively simple distribution, but it is a pattern found in many of the world's languages.

Now consider the distribution of universal quantifiers under the two assumptions, as represented in (12) to (14):

(12)	(a)	Mindenkit	János	hívott	meg
		everyone-ACC	John-NOM	invited-3.S	PERF

	(b)	János	hívott	meg	mindenkit

"It was John who invited everyone."

(13)		Mit	látott	mindenki
		what	saw	everyone
	WH-assumption	WH	V	$Q_{\forall}$
	Q-assumption	WH	V	$Q_{\forall}$
		"What did everyone see?"		

(14)		Ki/*mindenki	mit	látott?
		who/everyone	what	saw
	WH-assumption	$WH_{\forall}$/*Q	WH	V
	Q-assumption	Q_{WH}/*Q	WH	V
		"For everyone, what did they see?"		

First of all, as (12) shows, universal quantifiers can generally appear before or after the verb. When they precede the verb, they also precede the focus which sits in the immediate preverbal position (12a). As these sentences contain no wh-phrase (or quantifier realized as such) there is no distinction between their treatment under either the WH- or the Q-assumption.

From the WH-assumption perspective, as shown in (13) and (14), when quantifiers appear in interrogative clauses, they must follow the verb, even though they can precede other kinds of foci. Clearly this is a complication that requires an explanation. Both É. Kiss (1993) and Lipták (2000) assume that this follows from the fact that a wh-phrase preceding a wh-phrase is interpreted as a universal quantifier and this somehow blocks the appearance of the quantifier, which would yield a structure with the same interpretation. However, as we have seen, the interpretive rule they assume has very little explanatory power and likewise it provides a poor basis for explaining why quantifiers cannot precede wh-phrases in focus position.

Once more, the Q-assumption provides a more straightforward picture. Quantifiers can appear before foci and postverbally in both interrogative and declarative contexts. The only issue to be addressed is that when they come before a wh-phrase they are realized as wh-phrases. Accounting for this turns out to be a much easier task than any account adopting the WH-assumption, as we will demonstrate in the following sections.

4.2 Late Insertion

One of our criticisms of the WH-assumption is the lack of a theory that allows one lexical item to be reinterpreted as another. It is another advantage of the Q-assumption that there already exists a theory which allows for the situation in which an element is realized differently in different contexts.

A number of current frameworks have adopted a late (vocabulary/lexical) insertion approach, such as Distributed Morphology (Halle and Marantz 1993) and Nanosyntax (Starke 2009). This moves away from the lexicalist tradition in assuming that morphemes

do not come preformed, with all properties intact, ready to be manipulated by syntactic operations. Instead such frameworks assume that the syntactic system operates on abstract sub-morphemic elements, which lack phonological properties, and builds them into larger constructs. Only once the syntax is finished are these syntactically constructed morphemes realized by associated exponents.

The main advantage of this approach is that it is not constrained by the idea that exponents and underlying syntactic/semantic elements are in a fixed relationship and therefore a given exponent can be used to realize a number of morphemes which differ in their sub-morphemic composition. This idea has proved useful in accounting for various phenomena, such as syncretism, and in simplifying the description of morphological distributions. The theory is also compatible with the idea that the same underlying construct can be realized by different exponents in different contexts, i.e., the idea behind the Q-assumption.

Typically late insertion approaches work with a simplified vocabulary/lexicon and the more limited lexical resources are made to work harder with exponents competing against each other for selection. The selected exponent is the "best fitting" one for any instance of morpheme realization. This relies on the assumption that exponents do not have to be associated with exactly the set of features that they are used to spell out. For example, under some assumptions as to what counts as the "best fit," it may be that a morpheme constructed of sub-morphemic elements [a, b] is realized by an exponent X that is associated with [a] in its lexical specification, if there is no better fitting exponent. Thus, the selected exponent does not have to be associated with exactly the set of features that it is used to spell out.

Turning to the case in hand, the situation must be as follows. The Hungarian syntactic system constructs a universal quantifier out of a set of sub-morphemic elements (e.g., [Op+∀+non-human] "everything," [Op+∀+human] "everyone," [Op+∀+place] "everywhere," etc.).[6] It also constructs a set of interrogative morphemes (e.g., [Op+WH+non-human] "what," [Op+WH+human] "who," [Op+WH+place] "where," etc.). Each of these has an associated exponent which realizes them under normal circumstances:

(15) (a) [Op+∀+non-human] ↔ *minden*
[Op+∀+human] ↔ *mindenki*
[Op+∀+place] ↔ *mindenhol*

6 It is not our concern in this paper to give the details of exactly how these morphemes are constructed. Our proposals are compatible with a number of different frameworks and so it is not important to select any particular one here.

(b) [Op+WH+non-human] ↔ *mit*
[Op+WH+ human] ↔ *ki*
[Op+WH+place] ↔ *hol*

However, for some reason to be identified, the exponents in (15a) are not the best spell out possibility for underlying universal quantifiers when they precede wh-phrases and those in (15b) turn out to be better. All that remains is to identify the reasons behind this.

4.3 Clausal Typing

Our proposal will be based on the notion of clausal typing, introduced by Cheng (1991). Clausal typing is the overt marking of a clause's interrogative status in one of a number of ways.

Generally Cheng considers there to be two types of languages with respect to clausal typing: those that mark interrogative clauses with special particles and those which do not. The latter utilize wh-fronting instead. The core of Cheng's theory is the Clausal Typing Hypothesis:

(16) Every clause needs to be typed.

For now, all that needs to be noted is that Hungarian is the kind of language which types wh-interrogative clauses via wh-fronting.[7] We have already noted that in Hungarian interrogative clauses one and only one true wh-phrase is fronted to a position immediately before the verb. This can be seen as a direct result of the application of (16).

However, another very well known aspect of the syntax of Hungarian is the fact that scope relations are overtly marked by placing operators in front of those they scope over. We claim that these two properties, clausal typing by wh-fronting and the leftmost condition on wide scope operators (henceforth LWO), are in direct conflict with each other and that it is precisely this conflict which gives rise to the realization of underlying universal quantifiers as wh-phrases. Before we can build on these claims, we will need to further investigate the properties of clausal typing.

4.4 The Semantic Basis of the Typing Requirement

One way to actualize the claim that there is a conflict between the Clausal Typing Hypothesis and the LWO would be through the idea that both are left edge conditions. This seems quite natural given that fronting obviously involves the left periphery. Thus the two conditions require different elements to be leftmost and the conflict arises in an interrogative clause with a wide scope non-interrogative operator: the typing condition

7 Hungarian uses the particle strategy for yes-no questions, though we will have nothing to say about this in the paper.

would require the wh-operator to be leftmost and the LWO would require the non-interrogative to be leftmost.

This seems to be borne out by the data we have reviewed so far, as a narrow scope quantifier coming to the right of a clausal typing wh-phrase is perfectly grammatical ([17a] with interpretation [i]), but a wide scope quantifier coming either to the left (17b) or the right ([17a] with interpretation [ii]) of a typing wh-phrase is ungrammatical:

(17) (a) Mit látott mindenki?
what-ACC saw everyone-NOM
(i) Qy ∀x [x saw y]
(ii) * ∀x Qy [x saw y]

(b) * Mindenki mit látott?

These observations might suggest that this conflict is located entirely in the syntax and has to do with a competition for the relevant leftmost position.

However, there are at least two reasons to believe that this is not the best characterization of the situation. The first is that it isn't true that universal quantifiers never precede wh-phrases. This is possible if the quantifier is interpreted as a contrastive topic:

(18) /Mindenki \mit látott?
everyone-NOM what-ACC saw
"What did /everyone \see?"

A question such as (18) can be understood in the following context. Suppose a small group of tourists visit an art gallery but decide to explore it separately, each member going to see the pieces that they are personally interested in. The question asks for the identity of the artwork that all members of the group ended up seeing, in contrast to those which only subsets of the group may have seen. Note that the question has a special intonation pattern, indicated by the rise on "Mindenki" and the fall on "mit." We see the same rise-fall pattern in the English translation too and it is a well known phenomenon in a number of other languages (for example, see Gyuris [2009] and references cited therein). The semantic effect of this intonation pattern seems universally to force a narrow scope interpretation on the quantifier. Hence, unusually for Hungarian, (18) has an inverse scope interpretation, equivalent to that in (17a), interpretation (i).

We might try to salvage the left edge character of clausal typing by defining the domain that the typing wh-element must precede as being smaller than the domain that contains the contrastive topic. However, this would not address the second reason to doubt the syntactic characterization of the conflict between clausal typing and scope marking. As the following data show, it is not only the leftmost universal quantifier preceding a

wh-phrase that must be realized as a wh-phrase, but all non-contrastive topic universal quantifiers that precede a wh-phrase:

(19) (a) Ki kinek mit adott?
who-NOM. who-DAT what-ACC gave
"Who gave what to who?"
∀x ∀y [what did x give to y]

(b) Mit adott mindenki mindenki-nek
what-ACC gave everyone-NOM everyone-DAT
Qy [everyone gave y to everyone]

(c) Ki mit adott mindenki-nek
who-NOM what-ACC gave everyone-DAT
∀x [what did x give to everyone]

In (19a) there are two universal quantifiers realized as wh-phrases (*ki* and *kinek*) and only one actual wh-phrase (*mit*). Both of the quantifiers precede the wh-phrase, but obviously only one of them is at the left edge of whatever domain is relevant for clausal typing. If clausal typing were simply a left edge phenomenon we would expect only the leftmost quantifier to be realized as a wh-phrase and the second one to be realized as a quantifier. This, however, is ungrammatical:

(20) *Ki mindenki-nek mit adott?
who-NOM everyone-DAT what-ACC gave

The rest of the data in (19) suggest a different account as to what conditions lead to a universal quantifier being realized as a wh-phrase. In (19b) both quantifiers follow the verb, and subsequently the wh-phrase. True to the nature of Hungarian, the scope interpretation of the quantifiers is narrow with respect to the wh-phrase. In (19c) however, one quantifier precedes the wh-phrase and one follows. The preceding one is realized as a wh-phrase and the following as a quantifier. The preceding quantifier has a wide scope interpretation with respect to both the wh-phrase and the other quantifier.

Thus, quantifiers that are realized as wh-phrases not only precede wh-phrases, but have wide scope interpretations with respect to them as well. In fact, associating clausal typing with a wide scope interpretation, rather than a leftmost position, handles the data more straightforwardly. Note that in (19a), where both quantifiers are realized as wh-phrases, while only one is at the left edge, both have scope over the real wh-phrase. In this situation, the quantifiers' scopes do not interact. Consequently, both are equally interpreted as having wide scope. It is only in (19c) that the two quantifiers'

scopes interact, and in this the one with wide scope is realized as a wh-phrase and the one with narrow scope is not. We therefore propose the following version of clausal typing theory:

(21) The Clausal Typing Hypothesis
Every clause needs to be typed.

(22) Clausal typing comes in two forms:
(i) typing with particles
(ii) typing by realizing wide scope operators as wh-phrases.

Obviously, for a language making use of strategy (22ii), such as Hungarian, when the wide scope operator is a wh-phrase, its realization as such is straightforward. It is only cases where interrogative clauses have wide scope non-wh-operators that we get the special realization of this operator as a wh-phrase.

In the last section of this paper we will discuss how clausal typing interacts with the process of vocabulary selection.

5. Vocabulary Selection

5.1 The Subset and Superset Principles

Different approaches to late insertion tend to adopt different strategies to determine the best exponent for spelling out underlying morphemes in those cases where there is competition. These cases tend to involve the situation in which there is no exponent associated with all the features of the morpheme to be spelled out. The two most common strategies involve whether the selected exponent is allowed to be associated with features not possessed by the morpheme (overspecification) or whether it can be allowed to not be associated with some of those features (underspecification).

Proponents of Distributed Morphology tend to favour the Subset Principle in deciding cases of selection:

(23) The Subset Principle
Select the exponent associated with the largest subset of the features of the morpheme to be spelled out.

This allows selected exponents to be underspecified, but sanctions against overspecification. Therefore, any exponent associated with a feature not present on the underlying morpheme is automatically ruled out as a possible candidate realization.

Those who work in the framework of Nanosyntax, however, have argued against adoption of the Subset Principle (see Caha [2016] for a detailed criticism) and instead propose virtually the opposite, which they call the Superset Principle:

(24) The Superset Principle
Select the exponent associated with the smallest superset of the features of the morpheme to be spelled out.

Obviously, this allows overspecification whilst sanctioning against underspecification.

It will not be our purpose to argue in favor of one or the other of these principles as the point to be made is that neither are suitable for accounting for the Hungarian data discussed here. To see this, consider the proposed situation. We start with an underlying universal quantifier which presumably has all the features compatible with universal quantifiers in other positions. However, this gets realized by a wh-phrase, which, although it is specified for certain features compatible with the underlying quantifier, is not specified for universality (or whatever the feature is that distinguishes a universal quantifier from other operators). Therefore this particular realization involves underspecification: the exponent is not associated with features present on the morpheme to be spelled out. Given that the Superset Principle does not allow underspecification, clearly it is incompatible with the observations.

In addition, the wh-phrase which actually spells out the underlying quantifier is specified for an interrogative feature, which is not present on the underlying morpheme. Hence we also have a case of overspecification. As the Subset Principle explicitly denies the possibility of overspecification, it is also not compatible with the data.

One hope to salvage the Subset Principle comes from Cheng's (1991) analysis. According to this, Hungarian wh-elements are not specified for an interrogative feature. This conclusion is reached from the fact that some of these morphemes occur in non-interrogative operators, such as universal and epistemic quantifiers:

(25) *ki* (who) *minden-ki* (everyone)
vala-ki (someone)

hol (where) *minden-hol* (everywhere)
vala-hol (somewhere)

mi (what) *vala-mi* (something)

Cheng proposes that the quantificational element in the quantifiers (*minden*, *vala*, etc.) are determiners which provide the relevant quantificational feature for the whole construction. For the interrogative pronouns, she claims that there is a null interrogative determiner providing the relevant feature:

(26) $[\varnothing_{\text{WH}}]_{\text{D}}$-*ki*
$[\varnothing_{\text{WH}}]_{\text{D}}$-*hol*
$[\varnothing_{\text{WH}}]_{\text{D}}$-*mi*

If it could be maintained that the wh-element which is used to spell out the underlying quantifier is not accompanied by the interrogative determiner, then the process of realization would not involve overspecification. Hence the Subset Principle might yet be used under these assumptions.

Unfortunately it is not possible that the wh-feature, whatever its origin, is missing when the quantifier is realized as a wh-pronoun, as this exponent is selected in order to satisfy clausal typing requirements. Presumably it is only something that is specifically marked for interrogative that can be used to type an interrogative clause. Therefore even if the "interrogative" pronouns are themselves not associated with the WH-feature, this feature must still be present when the pronouns are used to spell out quantifiers. It follows that the overspecification involved here cannot be circumnavigated and the Subset Principle cannot be salvaged.

5.2 Targeted Underspecification

There is an alternative to the Subset and Superset Principles which allows for both under- and overspecification, making it more suitable for present purposes. This was proposed in Newson (2014), which showed that in accounting for the distribution of English modal verbs across the set of modal features that they spell out, it is important to allow for overspecification. For example, it is very typical for English modals to spell out certain features under certain uses which they do not in other uses: *may* is formal in its use as a deontic, but when it is used as an epistemic there is no indication of formality. However, it is also important to allow for a limited amount of underspecification on certain "targeted" features. The fact that every modal is used to express more than one type of modality (epistemic, deontic or dynamic) demonstrates that these features are underspecified for many modals (see Newson [2014] for details).

We claim that the system of Targeted Underspecification is exactly what is needed to account for the spelling out of an underlying universal quantifier as a wh-phrase in Hungarian. The bare bones of the proposal can be summed up as follows. It is more important to satisfy a condition requiring an element to type a clause than it is to spell out the feature which identifies an operator as a universal. Thus a wh-exponent which is underspecified for the universal feature is a better selection to spell out a universal quantifier when that quantifier takes wide scope in an interrogative clause.

To add some flesh to this account, first of all let us point out a number of important facets to the proposal. The system is, like all late insertion accounts, based on the notion of competition in which the "best fit" exponent is selected. Furthermore, exponents are selected on the basis of how well they satisfy certain conditions. Most of these conditions require there to be a match between the lexical specifications of the exponent and the features it is used to spell out. For example, we might think of a condition "Match ∀," which is satisfied when the universal feature ∀ is to be spelled out and the selected exponent is specified for this feature. Under normal circumstances, such a condition

would favor the selection of *everyone* ([Op+∀+human]) over *who* ([Op+WH+human]) when the features to be spelled out are those of a universal quantifier.

The typing requirement is not a matching condition, as it does not concern the situation in which an underlying feature is to be spelled out. Instead it imposes a general condition on clauses: that they must be typed by the appearance of a typing morpheme or a wh-feature on a wide scope operator. This condition overrules the matching condition for the universal feature, though the matching condition remains operative in contexts where the typing requirement does not conflict with it.

Optimality Theory offers a framework from within which we can exactly model the situation described above. In OT a set of candidate expressions compete against each other for grammaticality and are evaluated against a set of constraints. The constraints are ranked in terms of importance: the satisfaction of highly ranked constraints is imperative, while the violation of constraints with lower ranking is possible, if such violation ensures the satisfaction of a higher ranked constraint.

We can take the exponents competing against each other to be the candidates of an optimality system. This system takes the matching and other conditions to be the constraints which evaluate the candidate set and decide which is the best. As is standard in OT, we can represent this in table form:

(27) Condition: WH > ∀

		Typing	Match ∀
☞	*ki . . . mindenki*		
	ki . . . ki		*

This table represents the spelling out of an interrogative and a universal quantifier where the quantifier has narrower scope than the interrogative. Given that the interrogative is the widest scope operator and is spelled out as an interrogative, the Typing condition is satisfied. However, only when the quantifier is spelled out by an exponent specified for the universal feature is the Match ∀ constraint satisfied. Hence the winner is the first candidate, as indicated by the pointy finger.

Table (28) shows the result when the quantifier has wide scope:

(28) Condition: ∀ > WH

		Typing	Match ∀
	mindenki . . . ki	*	
☞	*ki . . . ki*		*

In this case, as the quantifier has wide scope, the typing condition requires it to type the clause and hence is violated if the universal quantifier is realized as such. Although the realization of the quantifier as a wh-pronoun violates the Match ∀ condition, this

violation allows the satisfaction of the higher ranked typing condition and is therefore an admissible violation. The second candidate is optimal.

6. Conclusion

In this paper we have argued for three main specific points. The first is that a late insertion approach to Hungarian quantified interrogative clauses (Q-assumption) is superior to the approach which assumes certain wh-phrases are reinterpreted as quantifiers (WH-assumption). The advantage of the former is twofold. First it does not rely on a theory that must be specifically developed for the purposes of accounting for the phenomena as the WH-assumption requires. Second, the Q-assumption leads to a much simpler description of the phenomena and therefore facilitates an easier account.

The second point concerns the process of lexical/vocabulary insertion. The analysis that we propose depends crucially on semantic facts concerning whole sentences: typing is determined by the relative scopes that operators have to each other. This means that it is not just a matter of viewing underlying features to determine which exponents to select. Instead, there must be direct input to the process from the semantic representation as well. This is not a unique conclusion as virtually the same thing has been discovered, though concerning very different phenomena, in both Distributed Morphology (Marantz 1997) and Nanosyntax (Starke 2011). It seems therefore that this is turning out to be a central tenet of late insertion approaches.

Finally we have shown that, if quantified interrogatives involve the realization of an underlying universal quantifier with a wh-exponent, then the Subset and Superset Principles are not able to account for the phenomena. Such a realization involves both under- and overspecification at the same time and given that these principles sanction against one or the other of these, they must be rejected. Targeted Underspecification, on the other hand, offers a very simple account which fits the data perfectly, thus supporting this as the correct determiner of exponent selection.

Works Cited

Caha, Pavel. 2016. “Notes on Insertion in Distributed Morphology and Nanosyntax.” Lingbuzz/002855. Accessed September 9, 2016. http://ling.auf.net/lingbuzz/002855.

Cheng, Lisa. 1991. “On the Typology of Wh-Questions.” PhD diss., MIT, Cambridge, MA.

É. Kiss, Katalin. 1993. “WH-movement and Specificity.” *Natural Language & Linguistic Theory* 11 (1): 85–120.

Gyuris Beáta. 2009. *The Semantics and Pragmatics of the Contrastive Topic in Hungarian.* Budapest: The Library of the Hungarian Academy of Sciences and Lexica Ltd.

Halle, Morris, and Alec Marantz. 1993. “Distributed Morphology and the Pieces of Inflection.” In *The View from Building* 20*: Essays in Linguistics in Honor of Sylvain Bromberger*, edited by Kenneth Hale and Samuel Jay Keyser, 111–76. Cambridge, MA: MIT Press.

Lipták, Anikó. 2000. “Multiple Relatives as Relatives of Question.” In *Approaches to Hungarian* 7, edited by Gábor Alberti and István Kenesei, 153–77. Szeged: JATE Press.

Marantz, Alec. 1997. “No Escape From Syntax: Don’t Try Morphological Analysis in the Privacy of Your Own Lexicon.” *University of Pennsylvania Working Papers in Linguistics* 4 (2): Article 14.

Newson, Mark. 2014. “English Modals and Late Insertion.” In *Complex Visibles Out There: Proceedings of the Olomouc Linguistic Colloquium 2014*, edited by Ludmila Veselovská and Markéta Janebová, 253–73. Olomouc: Palacký University.

Starke, Michal. 2009. “Nanosyntax—A Short Primer to a New Approach to Language.” In *Nordlyd* 36 (1), edited by Peter Svenonius, Gillian Ramchand, Michal Starke, and Knut Tarald Taraldsen, 1–6. Tromsø: CASTL. Accessed May 27, 2016. http://septentrio.uit.no/index.php/nordlyd/article/view/213/205.

Starke, Michal. 2011. “Towards an Elegant Solution to Language Variation: Variation Reduces to the Size of Lexically Stored Trees.” Unpublished manuscript. Accessed May 27, 2016. http://ling.auf.net/lingbuzz/001183.

Prepositions and Islands: Extraction from Dative and Accusative DPs in Psych Verbs

Ángel L. Jiménez-Fernández

University of Seville, Spain

ajimfer@us.es

Abstract: In current research on the structure of DPs in Differential Object Marking and Dative Clitic Constructions, there has been an explosion of proposals suggesting that the preposition *a* present in both accusative and dative objects is not a true P in Spanish, but a morphological marker (Demonte 1995; Cuervo 2003; Ormazabal and Romero 2013a, b; among others). In this paper I analyse subextraction in the form of *wh*-movement out of both accusative and dative DP objects in psych constructions in Spanish. Experiencers have been held to be lower on the scale of resistance to different phenomena. Subextraction from experiencers introduced by *a* is a case at issue. Assuming van Riemsdijk's distinction between lexical and functional prepositions, I claim that *a* does not project into a PP but rather occupies a Kase position above DP, endowed with an Edge Feature which allows subextraction if other conditions are satisfied.

Keywords: functional prepositions; subextraction; islands; DOM/dative constructions; psych verbs

1. Introduction

In this work I explore the differences between the Spanish preposition *a* "to" in dative and Differential Object Marking (DOM) constructions, and other prepositions in terms of the relative transparency which it shows in cases of subextraction in the form of *wh*-movement. Psych verbs in Spanish may select either a dative or accusative object or both types of object.[1] The first case is illustrated in (1), the second in (2), and the third

1 One of the discriminating properties of psych verbs selecting dative or accusative is that of word order. The neutral order (all-focus) preferred for the dative construction is OVS, whereas SVO is the neutral order for the accusative construction. Any rearrangement is caused by information structure (see Fábregas et al. [forthcoming]; Jiménez-Fernández and Rozwadowska [forthcoming] for a detailed analysis).

one in (3a) for accusative object and (3b) for dative object (see Campos 1999; Marín and McNally 2011; Fábregas et al., forthcoming):

(1) A Juan le gusta ir al cine.
"John likes going to the cinema."

(2) Pedro ofendió a Juan.
"Peter offended John."

(3) (a) Marta lo molesta.
"Marta (actively) bothers/is bothering him."

(b) El humo le molesta.
"The smoke bothers him." (Marín and McNally 2011, 468)

I discuss cases of subextraction from the DP object introduced by the preposition *a* (hereafter, *a*-DP) selected by the three types of psychological verbs. I show that *a*-DPs are not islands by nature; rather their degree of islandhood depends on different factors, such as Specificity and d-linking. Subextraction with verbs which may only occur with Dative Clitic Constructions is illustrated in (4):

(4) ¿De qué edificio dices que no le han gustado tus sugerencias [a ningún vecino]?
"Of what building do you say that no neighbour has liked your suggestions?"

Subextraction from *a*-DPs selected by verbs requiring accusative objects is exemplified in (5):[2]

(5) ¿De qué edificio dices que ofendieron a un vecino?
"Of what building do you say they offended a neighbour?"

Finally, I will also take into account psych verbs which can occur either with accusative or dative *a*-DPs. Still, subextraction is possible:

(6) ¿De qué edificio dices que esos gamberros han molestado a varios vecinos?
"Of what building do you say those vandals have bothered several neighbours?"

2 I focus on varieties of Spanish where a distinction is made between accusative and dative clitics. Accusative clitics occur with direct objects, whereas dative clitics are only compatible with indirect objects. For *Leístas* such a difference is blurred (see Fernández-Ordóñez [1999]; Ormazábal and Romero [2013a, b] for an overview of clitics for both direct and indirect objects and the connection with microvariation).

(7) ¿De qué edificio dices que a ningún vecino le molesta el humo de la calle?
"Of what building do you say the street smoke hasn't bothered any neighbour?"

In the relevant literature extraction out of an *a*-DP (both dative and DOM-marked) has been claimed to yield an ungrammatical outcome (Ordóñez and Roca, forthcoming; Kayne 2005):

(8) (a) ?*[$_{CP}$ De quién$_i$ C has visitado [$_{v*P}$ pro v* [a muchos amigos t$_i$]]]?
"Who have you visited many friends of?"

(b) *[$_{CP}$ De quién$_i$ C le diste [$_{v*P}$ pro v* los libros [a los padres t$_i$]]]?
"Whose parents did you give the books?"
(Ordóñez and Roca, forthcoming, ex. 77)

In this work, I show that subextraction out of *a*-DPs is possible in Spanish, given that certain conditions are met. I contend that subextraction in (4)–(7) results in grammaticality and that the ungrammatical or unnatural cases are the consequence of the violation of one or more conditions. My analysis of cases of subextraction from *a*-DPs runs along the following claims:

1) Accusative and dative *a* are functional Ps which do not project a PP; lexical Ps do project a PP.
2) Having projected into a PP, lexical P blocks movement only if PP is an island. On the other hand, if *a* does not project into a PP, movement is expected not to be blocked.
3) Dative and Accusative P are transparent for movement since the preposition *a* is just a case-assigning element, resurrecting the suggestion that this kind of DP projects into a Kase Phrase (KP), a functional projection endowed with an Edge Feature (EF) which makes extraction possible.

The article is organised as follows. In Section 2, I show that dative and accusative P *a* is not a lexical P projecting into a PP; it is rather a functional P. In Section 3, I suggest that *a* is the head K of a KP, which is not a strong island since it is endowed with an Edge Feature, thereby facilitating subextraction. In Section 4, I present the conclusions.

2. Some Remarks on the Grammatical Status of *a*

In current research on the structure of DPs in Differential Object Marking and Dative Clitic Constructions, there has been an explosion of proposals suggesting that the preposition *a* present in both accusative and dative objects is not a true P in Spanish, but a morphological marker (see Demonte 1995; Torrego 1998; Cuervo 2003; Ormazabal and Romero 2013a, b; Pineda 2013, among others). Within this line of research, Rodríguez-

Mondoñedo (2007), López (2012) argue that the insertion of *a* is a consequence of Distributed Morphology. Zdrojewski (2013) claims that its insertion is a PF operation after the impossibility of valuing the DP's Case feature.[3]

On the other hand, other linguists such as Ordóñez and Roca (forthcoming) and Kayne (2005) consider *a* as a full preposition. As such, the preposition probes in search of its complement for reasons of Agree. If this analysis is on the right track, the prediction is that as a PP the constituent *a*-DP should be an opaque domain for extraction, given the traditional view that PPs are islands (Boeckx 2003).

Abels (2013) has an intermediate approach, proposing that PPs are phases and depending on the language and on the type of movement they can be transparent for extraction since the operator moves to the edge of the phase (an escape hatch), thereby explaining the subextraction of the PP *za kakie prestuplenija* "for which crimes" out of another PP in Russian, as illustrated in (9):

(9) Za kakie prestuplenija on otkazal-sja ot otvetsvennosti?
for which crimes he rid-REFL of responsibility?
"Which crimes did he reject responsibility for?" (Abels 2013, 216)

The interim conclusion which can be drawn so far is that, under the right circumstances (to be discussed below), subextraction out of a PP is possible. Note that the P *ot* in Russian is very similar to Spanish dative/accusative *a* in that they are primarily used for case-assignment purposes.

2.1 Two Types of Preposition: Functional and Lexical

Prepositional phrases have been claimed to exhibit functional properties (cf. Hornstein, Nunes and Grohmann 2005; Radford 1997; Rooryck 1996; Rouveret 1991, among others). More precisely, van Riemsdijk (1978, 2015) has claimed that we can distinguish two types of preposition, namely lexical and functional prepositions, providing a list of salient properties which characterize each group.

Among the properties for functional prepositions that Riemsdijk singles out is the possibility that the DP selected by P can be a controller of PRO in a complement clause, as illustrated below for English:

(10) I rely on you$_i$ [PRO$_i$ to solve the problem].

(11) I$_i$ live with a woman$_j$ [PRO$_{i/*j}$ to water my plants].

3 In line with Ormazabal and Romero (2013a, b) and Zdrojewski (2013), I assume that in dative and DOM-marked DPs the *a*-DP is exactly the same element.

A second property of functional prepositions is that it signals the oblique case in languages that do not have specific morphology for case.

In Spanish evidence for the functional status of P *a* is provided by exactly the same two properties, which Van Riemsdijk (2015) uses for independent constructions. First, datives and some accusative objects in Spanish may be selected by a preposition *a*, which primarily serves the purpose of case assignment. For this reason, Demonte (1995) analyses objects introduced by P as DPs. I agree that *a* has a case-assigning function, but I will not make a stand for the claim that it is generated DP-internally.

In addition, this *a*-DP can act as controller of PRO in a complement clause, as shown in (12):

(12) A Juan$_i$ le gusta [PRO$_i$ bañarse en el río]
"John likes having a swim in the river."

Both are properties which point to the fact that *a* is a functional preposition.

Abraham (2010) has found out that real PPs, i.e., projections of lexical P, are islands for the purposes of anaphoric relations, and offers the contrasts in (13)–(14) from English:

(13) The group$_i$ laughed [about themselves$_i$/*them$_i$].

(14) The group$_i$ sat under a big rain shelter [above them$_i$/*themselves$_i$].

The P *about* is transparent and the anaphor *themselves* in (13) is licensed in compliance with Principle A of the Binding Theory. Nevertheless, a P such as *above* is lexical and hence opaque in (14). As a consequence, the DP the group cannot bind the anaphor.

This difference between lexical and functional Ps with respect to anaphoric relations raises the question as to whether *a*-DPs exhibit this opacity. Consider examples in (15) and (16), DOM-marked and dative constructions respectively:

(15) Jimena se mira [a sí misma/*ella] en el espejo.
"Jimena is watching herself at the mirror."

(16) Ángela se dio [a sí misma/*ella] una última oportunidad.
"Angela gave herself a last chance."

As is clear from the data, *a* patterns with functional prepositions in that the object introduced by this preposition are transparent with respect to licensing anaphors. The question now is whether this functional P is also transparent for the purposes of other phenomena, an issue that I discuss in the next subsection.

2.2 May *a*-DPs be Transparent for Wh-Extraction?

Within the distinction between two types of P and the opacity/transparency of *a*, the question arises as to the island status of direct and indirect objects introduced by P *a* in terms of subextraction in Spanish. Gallego (2007, 312) has explicitly argued that "both Case marked direct objects . . . and indirect objects . . . are islands." Pineda (2014) echoes Gallego's words and gives the following examples:

(17) ¿[De qué escritor]$_i$ has comprado [$_{DO}$ dos libros t_i]?
"Of what writer did you buy two books?"

(18) (a) *¿[De quién]$_i$ has saludado [$_{DO}$ a muchos amigos t_i]?
"Who have you greeted many friends of?"

(b) *¿[De quién]$_i$ le diste los libros [$_{IO}$ al padre t_i]?
"Whose father did you give the book?"

Gallego and Uriagereka (2007) also assume the islandhood status of *a*-DPs, illustrating the degradation with sentences such as (19)–(20) (see also Torrego [1998] for a similar view). Note, nevertheless, that the extractees in all the ill-formed examples are non-Discourse-linked in the sense of Pesetsky (1987). Thus there is no previous mention the extracted material in the context.

(19) *¿[$_{CP}$ De quién$_i$ has visitado [$_{DP}$ a muchos amigos t_i]]?
"Who have you visited many friends of?"

(20) *¿[$_{CP}$ De quién$_i$ le diste los libros [$_{DP}$ a los padres t_i]]?
"Who did you give the books to the parents of?"

The problem that these authors adduce is that the presence of *a* blocks Agreement between *v* and the DP. Ordóñez and Roca (forthcoming) claim that the P *a* in overtly-cased objects is a probe, and hence it is a real P with an active role in syntax. And part of their evidence is precisely based on the ban on extraction when *a* is present, thereby predicting the difference in grammaticality of sentences (21)–(22):

(21) ¿De qué autor has leído los libros más representativos?
"Which author have you read the most representative books?"

(22) *¿De qué autor has visto a los representantes más obstinados?
"What author have you seen the most obstinate representatives?"

As is well-known since Chomsky (1973), extraction is licit only from non-specific objects. Hence, a change in terms of specificity in the extraction site will yield a perfectly grammatical output (in line with Jiménez-Fernández [2009, 2012]; Haegeman et al. [2014]):

(23) ¿De qué autor dices que has visto a varios representantes?
"What author do you say you have seen several representatives of?"

Therefore, my intuition is that the ungrammaticality detected in all previous examples must have some origin other than the presence of the P *a*. It is not the case that *a*-DPs are opaque *per se*. The degradation of the preceding examples is not due to the possible default opacity of the extraction site.

Ordóñez and Roca argue that the behaviour of the P *a* with respect to extraction is exactly the same as other prepositions, and provide examples of extraction out of a PP headed by what I have called functional preposition, such as *de* "of":

(24) (a) Me han hablado muy bien de los libros de Cortázar.
"They have talked to me very well about the Cortázar's books."

(b) *¿De quién$_i$ te han hablado muy bien [de los libros t_i]?
"Who have they talked to you very well about the books of?"

(25) (a) Le han dado el premio al hijo del vecino.
"They have given the prize to the son of the neighbor."

(b) *¿De quién le han dado un premio [al hijo t_i]?
"Who have they given a prize to the son of?"

If *a* were a full (lexical) preposition, and as such its projection were an island, how come the following examples are acceptable?

(26) ¿De qué autor han hablado hoy de varios libros?
"Of what author do you say they have talked about several books today?"

(27) ?¿De qué libro parece que le van a dar al autor el premio planeta?
"Of what book does it seem they will give the author the Planeta prize?"

(28) ¿De qué edificio dices que a varios vecinos no les ha gustado los cambios en el barrio?
"Of what building do you say several neighbours haven't liked the changes in the neighbourhood?"

(29) ¿De qué edificio dices que han aterrorizado a algunos vecinos con amenazas?
"Of what building do you say they have scared some neighbours with threats?"

In sentence (26) the verb requires the preposition *de* "of." As suggested above, this preposition is similar to accusative/dative *a* (as illustrated in [27]–[29]) in that both of them are transparent for the purposes of subextraction.

2.3 Diagnoses: *a* is Not a Lexical P

Kayne (2005) claims that preposition *à* in French causative constructions such as (30) is a real instance of P:

(30) Jean a fait manger la tarte à Paul.
"Jean has made Paul eat the cake."

He extends the same analysis to datives introduced by *à*. Ordóñez and Roca (forthcoming) also include under the very same label the preposition which occurs in Spanish DOM constructions.

Since Ordóñez and Roca base their analysis on Kayne's (2005) approach to French *à*, it will be interesting to test the properties of the French P with the behaviour of Spanish *a* in datives and accusatives. Apart from the impossibility of extraction, which is also mentioned by Kayne with respect to the opaque character of *a* as a P, let us see some other diagnoses:

i) On a par with other prepositions, a DP introduced by à cannot be extracted out of an adjunct in causative constructions, as illustrated in (31)–(32):

(31) ??L'enfant que je me suis endormi après avoir fait manger
"The child that I fell asleep after having made eat"

(32) *L'enfant à qui je me suis endormi après avoir fait manger une tarte
"The child whom I fell asleep after having made a cake"

Neither construction in Spanish is possible, extraction out of an adjunct is not licit (since Huang's [1982] Condition on Extraction Domains [CED]). Ordóñez and Roca provide the following examples:

(33) (a) ??La conferencia que yo me dormí después de haber oído
"The conference that I slept after listening to"

(b) *La persona a la que yo me dormí después de haber saludado
"The person to the that I slept after listening to"

The two sentences are ill-formed and we cannot draw the conclusion that Spanish *a* is a true P based on the basis of data which are not crystal-clear.

ii) The subject-related à-DP in French always occurs after a direct object, a typical position for PPs. The reverse order yields an unacceptable result:

(34) (a) J'ai montre la tarte à Jean.

(b) (?)J'ai montre à Jean la tarte.
"I have shown the cake to John."

(35) (a) J'ai fait manger la tarte à Jean.

(b) (?)J'ai fait manger à Jean la tarte.
"I have made John eat the cake."

In Spanish the direct object can be preceded or followed by a PP and this is constrained by Information Structure, as largely discussed in Jiménez-Fernández and Spyropoulos (2013). To be more precise, the relative ordering of a direct object and an indirect object is strongly influenced by discourse factors. The order DO+IO is preferred when the IO is information focus, whereas IO+DO is preferred when the information focus is the DO, but both alternatives are grammatical:

(36) (a) Le enseñé la tarta a Juan.

(b) Le enseñé a Juan la tarta.
"I showed the cake to John."

Exactly the same behavior can be observed with causatives:

(37) (a) Hice comerse la tarta a Juan.

(b) Hice a Juan comerse la tarta.
"I made John eat the cake."

The conclusion that can be reached here is that the position occupied by the *a*-DP in Spanish is not connected with the prepositional nature of *a* in the double object construal and in causative constructions.

How about datives and DOM-marked objects in psych constructions? Again the *a*-DP may occur in different positions, either pre- or post-verbally:

(38) (a) A Juan le gusta el caviar.

(b) El caviar le gusta a Juan.
"John likes caviar."

(39) (a) El pueblo adora muchísimo al presidente.

(b) Al presidente lo adora muchísimo el pueblo.
"The people adore the president very much."

As observed in (38)–(39), the position for the dative/accusative *a*-DP poses no problems, albeit the Information structure-based word order selected in each case (Jiménez-Fernández and Rozwadowska, forthcoming; Fábregas et al. 2015). In (38a) and (39a) the *a*-DP occurs pre-verbally and it may be the topic of the sentence. On the other hand, in (38b) and (39b) the *a*-DP is information focus when used post-verbally. As is clear the relative position of *a*-DP is not relevant for the prepositional status of *a*.

iii) In French Clitic Left Dislocation, preposing of a direct object requires the presence of a clitic, but with the subject-related *à*-DP this is not the situation (on a par with other Ps):

(40) (a) Paul, elle l'a déjà fait manger.
"Paul, she him has already made eat."

(b) A Paul elle a déjà fait manger une tarte.
"To Paul she has already made eat a pie."

The problem is again that in Spanish we do not have this contrast. In both direct objects and subject-related *a*-DPs we require the clitic when they are CLLD-ed:

(41) (a) A Juan lo vi ayer.
"John, I saw him yesterday."

(b) A Juan lo hice comer una tarta ayer.
"John, I made him eat a cake yesterday."

As we may notice, the behaviour of DOM-marked DPs is not similar to other PPs. If my hypothesis that DOM and datives are grouped together is correct, the question arises as to whether there is any positional constraints in Clitic dative constructions. CLLD-ed datives are exemplified in (42):

(42) A María Juan le dio un libro.
"To Mary John gave a book."

As is clear, the output is fully well-formed. Thus the conclusion is that Spanish *a*-DP cannot be analysed as French *à*-DP. More precisely, the data in this section indicate that Spanish *a* is not a lexical P. Recall that in Section 2.1 I discussed the properties which distinguish lexical Ps from functional Ps. Those properties alongside the tests in the present section lead us to conclude that Spanish *a* is a functional preposition (*sensu* van Riemsdijk).

2.4 Why Psych Verbs?

Haegeman et al. (2014) and Alexiadou et al. (2007) have proposed that Experiencers/Goals are less resistant to subextraction than Agents. Dealing with subject islands, Chomsky (2008, 160, fn.39) comments that "difference among theta roles might be relevant," but does not elaborate any further. The DPs analysed in Haegeman et al. and Chomsky (2008) are all subjects:

(43) (a) *Of which car did [the driver] cause a scandal?

(b) Of which books did [the authors] receive a prize? (Chomsky 2008)

For subject Experiencers of psych verbs, it is clear that subextraction yields better results in (44)–(45), thereby supporting the idea that Experiencers are lower on the scale of resistant θ-roles:

(44) How many teams do you think [supporters of] like to cause trouble at away games? (Radford, pers. comm.)

(45) ¿De qué equipo dices que se molestaron muchos fans por el resultado del partido?
"Of what team do you say many fans were bothered with the result of the party?"

As far as objects are concerned, if my view on DOM and Datives which are assigned the θ-role of Experiencer is correct, the prediction is that *a* should not block subextraction.

In (46)–(47) I show subextraction out of both accusative (and dative) DPs in psych constructions in Spanish and English, two languages which differ in the presence (Spanish) or absence (English) of P in the configuration of objects.

(46) (a) ¿De qué partido crees que ha conmocionado [a muchos votantes] la nueva normativa?
"Of what party do you think the new regulations have shocked many voters?"

(b) ¿De qué partido crees que no les ha gustado [a muchos votantes] la nueva normativa?
"Of what party do you think many voters didn't like the new regulations?"

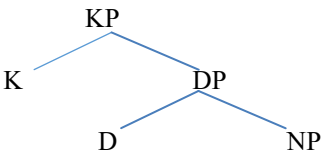
(47) Of which team did the press claim away games shocked [many fans]?

The grammaticality of (47) in English is explained straightforwardly since it involves extraction from an object, traditionally taken to be transparent (Huang [1982] and his CED), as opposed to subjects and adjuncts. However, the Spanish data are far from clear in that if a P is opaque for extraction (Boeckx 2003), we predict that the Experiencers introduced by P *a* should induce island effects, contrary to facts. Conversely, if the distinction between lexical (and hence opaque) and functional (and hence transparent) Ps is correct, the data can easily be accounted for. This is exactly the analysis that I put forth for *a*-DPs.

3. My Analysis of *a*-DPs

As stated earlier, my proposal is that the constituent *a*-DP is a Kase Phrase, which projects above DP (a functional projection independently proposed by Loebel [1994] or Lyons [1999], as a category separate from D):

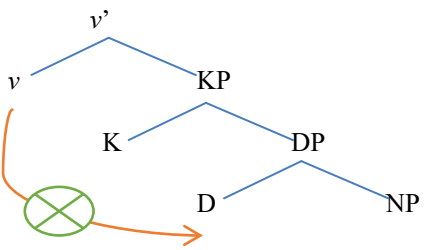
(48)



The analysis of *a*-DP as KP is proposed for dative and DOM-marked DPs in Zdrojewksi (2013), but he suggests that its insertion is a PF operation. The analysis here entails the insertion of the P in syntax, blocking Agree between *v* and the DP since K intervenes: DP agrees with the closest probe (K), hence the blocking effect proposed by Ordóñez and Roca obtains.

The generalization that can be drawn is that, when DP is an object with the featural set [+animate, +specific], K is inserted and *v* cannot probe DP since K is a closer probe, as shown in (49).

(49)



Only when The DP object is animate and specific will K be merged with DP. The intuition behind this claim is that when the DP has a different featural array, there is no intervening K and *v* probes DP so Agree is established.

This analysis poses some problems, given that only non-specific objects allow subextraction, as illustrated before, and that the functional preposition *a* selects a specific DP. The question arises then as to whether all DP objects introduced by *a* are specific.

One of the tests used for specificity is the distinction between indicative and subjunctive. The presence of *a* favours the use of indicative (Leonetti 2004):

(50) (a) Necesita a una enfermera que pasa la mañana con ella. (Indicative)
"He needs a nurse who spends the morning with her."

(b) Necesita una enfermera que pase la mañana con ella. (Subjunctive)
"He needs some nurse who might spend the morning with her."

Leonetti observes the correlation between specificity, indicative mood and their compatibility with *a*: (50a) is interpreted as about a specific nurse who will spend the morning with her, whereas (50b) is interpreted as any nurse who may spend the morning with her. The problem with this is that *a* is compatible with Quantifiers such as *algunos* "some," *varios*; " several," etc. These Quantifiers can be ambiguous and they can be interpreted as specific or non-specific (Suñer 2003). The crucial point is that the ambiguity does not vanish when they are introduced by *a*:

(51) Necesita a varias enfermeras que sepan usar el material quirúrgico.
"He needs several nurses that know how to use the surgical material."

Note that the *a*-DP is compatible with subjunctive, which is a symptom of non-specificity. The interpretation here involves several nurses who I may not know. This shows that *a* in DOM does not forcefully introduce specific objects. In Dative Clitic Constructions, where *a* is obligatory, there is no restriction about specificity (Ordóñez 1998) and Quantifiers are ambiguous:

(52) A algunas enfermeras les molestó la actitud del paciente.
"Some nurses were bothered by the patient's attitude."

Recall that *molestar* "bother" can also be used with accusative, as shown at the beginning of this work, and yet the Quantifier Phrase is still ambiguous:

(53) El paciente molestó a algunas enfermeras.
"The patient bothered some nurses."

The conclusion is then that *a* is compatible with animates regardless of whether they are specific or non-specific.

This conclusion is crucial for the purposes of subextraction, since one of the conditions that must be satisfied is precisely the non-specific nature of DPs. To illustrate, let us consider sentence (5), repeated here as (54):

(54) ¿De qué edificio dices que ofendieron a un vecino?
"Of what building do you say they offended a neighbour?"

Since the *a*-DP is a KP whose head has an Edge Feature, its transparency is predicted in our analysis. The cases where subextraction yields an ill-formed sentence are not to be explained by the opaque nature of the preposition. Rather the reasons are Specificity of the *a*-DP or the non D-linked nature of the extractee (Haegeman et al. 2014). This accounts for the ungrammaticality of (55):

(55) ¿De qué dices que ofendieron a los vecinos?
"Of what do you say they offended the neighbours?"

4. Conclusions

In this paper I have discussed cases of subextraction from DPs introduced by the P *a* (*a*-DPs). These are either accusative or dative objects of psych verbs. The conclusions arrived at follow. First, since these *a*-DPs project into a KP whose head K in endowed with an EF in Spanish, material can be extracted out of these Experiencer objects. This K accounts for the functional properties of *a*. Secondly, some verbs allow for either a dative or accusative DP and in both cases subextraction is permitted. Cases of degradation are explained if conditions other than the presence of the functional P *a* is taken into account. Two such conditions are specificity and D-Linking.

Funding Acknowledgement

The research I am presenting in this talk has been partly funded by the Research Project FFI2013-41509-P of the Spanish Government (MINECO) and by the Grant 2014/15/B/HS2/00588 of the Polish National Science Centre (NCN).

Acknowledgements

I am very thankful to Henk van Riemsdijk, Antonio Fábregas, Andrew Radford, Rafael Marín, Francisco Ordóñez and Mercedes Tubino for insightful discussion which has helped me a lot. Versions of this paper have been presented at the II GETEGRA workshop on Nominals at the University of Pernambuco (Recife, Brasil) and in OLINCO 2016 (Olomouc, Czech Republic). I thank the audiences there for their helpful comments.

Works Cited

Abels, Klaus. 2012. *Phases: An Essay on Cyclicity in Syntax*. Berlin: De Gruyter.

Alexiadou, Artemis, Liliane Haegeman, and Melita Stavrou. 2008. *Noun Phrase in the Generative Perspective*. Berlin: M. de Gruyter.

Boeckx, Cedric. 2003. *Islands and Chains. Resumption as Stranding*. Amsterdam: Benjamins.

Campos, Héctor. 1999. "Transitividad e intranstividad." In *Gramática Descriptiva de la Lengua Española,* edited by Ignacio Bosque and Violeta Demonte, 1519–74. Madrid: Espasa-Calpe.

Cuervo, María Cristina. 2003. "Datives at Large." PhD diss., MIT, Cambridge, MA.

Demonte, Violeta. 1995. "Dative alternation in Spanish." *Probus* 7: 5–30.

Fábregas, Antonio, Ángel L. Jiménez-Fernández, and Mercedes Tubino. Forthcoming. "What's Up with Datives?" In *Romance Languages and Linguistic Theory* 12. *Selected Papers from the 45th Linguistic Symposium on Romance Languages (LSRL), Campinas, Brazil*, edited by Ruth Lopes, Juanito Avelar, and Sonia Cyrino. Amsterdam: John Benjamins.

Fernández Ordóñez, Inés. 1999. "Leísmo, laísmo y loísmo." In *Gramática Descriptiva de la Lengua Española*, edited by Ignacio Bosque and Violeta Demonte, 1317–97. Madrid: Espasa Calpe.

Gallego, Ángel. 2007. "Phase Theory and Parametric Variation." PhD diss., Universitat Autònoma de Barcelona.

Gallego, Ángel, and Juan Uriagereka. 2007. "Conditions on Subextraction." In *Coreference, Modality, and Focus*, edited by Luis Eguren and Olga Fernández Soriano, 45–70. Amsterdam: John Benjamins.

Haegeman, Liliane, Ángel L. Jiménez-Fernández, and Andrew Radford. 2014. "Deconstructing the Subject Condition in terms of Cumulative Constraint Violation." *The Linguistic Review* 31(1): 73–150.

Hornstein, Norbert, Jairo Nunes, and Kleanthes K. Grohmann. 2005. *Understanding Minimalism*. Cambridge: Cambridge University Press.

Jiménez-Fernández, Ángel L. 2009. "On the Composite Nature of Subject Islands: A Phase-Based Approach." *SKY Journal of Linguistics* 22: 91–138.

Jiménez-Fernández, Ángel L., and Bożena Rozwadowska. Forthcoming. "The Information Structure of Dative Experiencer Psych Verbs." In *Various Dimensions of Contrastive Studies*, edited by Bożena Cetnarowska, Marcin Kuczok, and M. Zabawa, 91–111. Katowice: University of Silesia Press.

Jiménez-Fernández, Ángel L., and Vassilios Spyropoulos. 2013. "Feature Inheritance, vP Phases and The Information Structure of Small Clauses." *Studia Linguistica* 67 (2): 185–224.

Kayne, Richard S. 2005. "Prepositions as Probes." In *Movement and Silence*, edited by Richard S. Kayne, 83–104. Oxford: Oxford University Press.

Leonetti, Manuel. 2004. "Specificity and Differential Object Marking." *Catalan Journal of Linguistics* 3: 75–114.

Loebel, Elisabeth. 1994. "KP/DP-syntax: Interaction of Case Marking with Referential and Nominal Features." *Theoretical Linguistics* 20: 38–70.

López, Luis. 2012. *Indefinite Objects*. Cambridge, MA: MIT Press.

Lyons, Christopher. 1999. *Definiteness*. Cambridge: Cambridge University Press.

Marín, Rafael, and Louise McNally. 2011. "Inchoativity, Change of State, and Telicity: Evidence from Spanish Reflexive Psychological Verbs." *Natural Language and Linguistic Theory* 29: 467–502.

Ordóñez, Francisco. 1998. "Post-verbal Asymmetries in Spanish." *Natural Language and Linguistic Theory* 16: 313–46.

Ordóñez, Francisco, and Francesc Roca. Forthcoming. "Differential Object Marking (DOM) and Clitic Subspecification in Catalonian Spanish." In *The Syntactic Variation of Spanish Dialects*, edited by Ángel Gallego. Oxford: Oxford University Press.

Ormazabal, Javier, and Juan Romero. 2013a. "Differential Object Marking, Case and Agreement." *Borealis* 2 (2): 221–39.

Ormazabal, Javier, and Juan Romero. 2013b. "Object Clitics, Agreement and Dialectal Variation." *Probus* 25: 301–44.

Pineda, Anna. 2013. "Double Object Constructions in Spanish (and Catalan) Revisited." In *Romance Languages and Linguistic Theory 2011*, edited by Sergio Baauw, Frank Drijkoningen, Luisa Meroni, and Manuela Pinto, 193–216. Amsterdam: John Benjamins.

Pineda, Anna. 2014. "Les Fronteres de la (In)transitivitat." PhD diss., Universitat Autònoma de Barcelona.

Radford, Andrew. 1997. *Syntactic Theory and the Structure of English: A Minimalist Approach*. Cambridge: Cambridge University Press.

Riemsdijk, Henk C. van. 1978. *A Case Study in Syntactic Markedness: The Binding Nature of Prepositional Phrases*. Lisse: The Peter de Ridder Press.

Riemsdijk, Henk C. 2015. "Two Souls Alas Are Dwelling in the Breast of P: The Lexical and the Functional." Talk delivered at Workshop in honour of Bożena Rozwadowska, University of Wrocław, November 2.

Rodríguez-Mondoñedo, Miguel. 2007. "The Syntax of Objects: Agree and Differential Object Marking." PhD diss., University of Connecticut.

Rooryck, Johan. 1996. "Prepositions and Minimalist Case Marking." In *Studies in Comparative Germanic Syntax* II, edited by Höskuldur Thráinsson, Samuel D. Epstein, and Steven Peter, 226–56. Dordrecht: Kluwer Academic Publishers.

Suñer, Margarita. 2003. "The Lexical Preverbal Subject in a Romance Null Subject Language. Where Are Thou?" In *A Romance Perspective on Language Knowledge and Use*, edited by Rafael Núñez-Cedeño, Luis López, and Richard Cameron, 341–57. Amsterdam: John Benjamins.

Torrego, Esther. 1998. *The Dependencies of Objects*. Cambridge, MA: MIT Press.
Zdrojewski, P. 2013. “Spanish DOM as a Case of Lacking Case.” Talk at *Differential Object Marking Workshop*, University of Tromsø, May 23–24.

A New Syntactic Analysis of Dutch Nominal Infinitives

Kateřina Havranová

Palacký University, Olomouc, Czech Republic

katerina.havranova@upol.cz

Abstract: This paper deals with two types of Dutch nominalizations, the nominal infinitives (NIs) of two types. The first type are bare nominal infinitives (NI-Bs), the second type are determined nominal infinitives with the definite article *het* (NI-Ds). I will demonstrate that although their external syntax is basically the same, their internal structure differs. Unlike previous studies (e.g., Hoekstra 1985; Zubizarreta and Van Haaften 1988; Looyenga 1992; Hoekstra 1999; Schoorlemmer 2002; Reuland 2011; Broekhuis and Den Dikken 2012) that described Dutch NIs as "notoriously difficult to analyze" (Schoorlemmer 2002), I aim to show that these constructions are very systematic and logical if explained through a single operation that combines "Merge" and "Categorial Switch." Although other studies (e.g., Panagiotidis and Grohmann 2009) used the term "Categorial Switch," I reduce it to mechanisms used elsewhere in the grammar.

Keywords: Dutch nominalizations; nominalization process; Dutch nominal infinitives; external distribution; internal syntax

1. Introduction

This paper deals with two types of Dutch nominalizations, the nominal infinitives. Traditionally, four types of constructions are considered to be Dutch nominalizations: derived GE-nominals (*het GEtreiter van zijn klasgenoot* "the bullying of his classmate"), derived ING-nominals (*de behandelING van de patient* "the treatment of the patient"), and bare and determined nominal infinitives, discussed in more detail in this paper, since they all fulfill two basic criteria. Firstly, they inherit the denotation (namely the state of affairs) of the verb they are derived from and, secondly, they inherit the argument structure of that verb if interpreted as process (event) nominals. However, the main focus of this paper are the nominal infinitives (henceforth NIs) of two types.

2. Dutch Nominal Infinitives

Dutch nominal infinitives (NIs) are phrases that at first sight appear to be headed by an infinitival verb form (e.g. *lezen* "read," *schrijven* "write," *eten* "eat," etc.). Like derived nominals, they inherit the denotation as well as the argument structure of the verb they are derived from. However, one substantial difference between NIs and the derived nominals mentioned above is that nominal infinitives always denote the action of the verb as a process (event) and never a result.

As mentioned above, Dutch distinguishes two types of nominal infinitives. Both the first type, which I will from now on refer to as NI-Bs are bare (indefinite) nominal infinitives (1a), and the second type, which is henceforth referred to as NI-Ds, are nominal infinitives with the definite article *het* (1b), normally used for neuter nouns.

(1) (a) Boeken lezen is interessant.
books read is interresting
"Reading books is interesting."

(b) Het lezen van boeken is interessant.
the read of books is interesting
"The reading of books is interesting."

If we examine the two types of nominal infinitives from the point of view of their external syntax, it appears that they have exactly the same distribution as regular DPs with syntactic functions of subjects (2a–a'), direct objects (2b–b'), PP-objects (2c–c') or adverbials (2d–d'). Compare the pairs of sentences below, where the first sentence is always a bare nominal infinitive (NI-B) and the second sentence is a determined nominal infinitive (NI-D):

(2) NI-B as a subject
(a) Dat verslag zegt dat fruit eten gezond is.
that report says that fruit eat healthy is
"That report says that eating fruit is healthy."

(a') NI-D as a subject
Dat verslag zegt dat het eten van fruit gezond is.
that report says that the eating of fruit healthy is
"That report says that the eating of fruit is healthy."

(b) NI-B as a direct object
Ik haat boeken lezen.
I hate books read
"I hate reading books."

(b') NI-D as a direct object
Ik haat het lezen van boeken.
I hate the read of books
"I hate the reading of books."

(c) NI-B as a PP object
Ik ben dol op zeilen.
I am crazy on sail
"I am crazy about sailing."

(c') NI-D as a PP object
Ik ben dol op het zeilen.
I am crazy on the sail
"I am fond of the sailing."

(d) NI-B as an adverbial
Hoofdpijn gaat weg na water drinken
headache goes away after drink water
"Headache goes away after drinking water."

(d') NI-D as an adverbial
Zijn hoofdpijn ging weg na het drinken van water
his headache went away after the drink of water
"His headache went away after the drinking of water."

Another test for their external syntax is the coordination test. Since only constituents of the same type can be coordinated and nominal infinitives can co-occur with other DPs headed by nouns which are not derived from verbs (3a–b), they must be DPs themselves.

(3) (a) Voldoende water drinken en voldoende rust is gezond.
plenty of water drink and enough rest is healthy
"Drinking plenty of water and enough rest is healthy."

(b) Het voldoende drinken van water en voldoende rust
the plenty of drink of water and enough rest
is gezond.
is healthy
"The drinking of enough water and rest is healthy."

Moreover, nominal infinitives follow prepositions in PPs (4a–b) which is a typical position of noun phrases. The following examples illustrate that both bare and determined NIs behave in the same way in these tests:

(4) (a) Ik ben dol op films en boeken lezen.
I am crazy on films and books read
"I am fond of films and reading books."

(b) Ik ben dol op films en het lezen van boeken.
I am crazy on films and the read of books
"I am fond of films and the reading of books."

So far we have seen that externally both types of nominal infinitives have the same distribution as DPs.[1] In the next section I will treat each type of Dutch nominal infinitive separately, examine their internal syntax and compare their nominal and verbal properties.

2.1 Bare Nominal Infinitives

In this section, I will examine more closely the first type of nominal infinitives, that is bare (indefinite) nominal infinitives (NI-Bs). Just like English VP gerunds, NI-Bs seem to have the internal structure of VPs, with a verbal lexical head (Zubizarreta and van Haaften 1988). This for example means that in Dutch the object precedes the V frequently, rather than following it as a *van*-phrase (the Dutch equivalent of an English *of*-phrase).

In the infinitival construction with *te* (the Dutch counterpart of the English infinitive with *to*) (5a), the *van*-phrase is excluded completely (5b).

(5) (a) Het is leuk boeken te lezen.
it is nice books to read
"It is nice to read books."

(b) *Het is leuk te lezen van boeken.
it is nice to read of books

It should be pointed out that in Dutch objects of verbs normally precede their head in VPs, while they follow it in NPs in the form of a *van*-phrase, so that if the object can precede the infinitive as in (6a), then the infinitive must be verbal. If an object can follow it as in (6b), then it must be nominal as well.

1 Note incidentally that these two tests do not treat English infinitives in this way.

(6) (a) Sigaren roken is ongezond.
cigarettes smoke is unhealthy
"Smoking cigarettes is unhealthy"

(b) ?Roken van sigaren is ongezond.
smoke of cigarettes is unhealthy
"Smoking of cigarettes is unhealthy"

However, we must note here that the reported judgments with respect to the acceptability of bare NI-Bs with an object following the head in a *van*-phrase PP (6b) differ among authors. While some (e.g., Looyenga 1992) exclude this completely, others (e.g., Broekhuis and Den Dikken 2012) consider it a less preferred and more marked option. Referring to what was said before, if (6b) is acceptable and the object can follow the head, then the construction must have some nominal properties as well.

With respect to the object form in NI-Bs, it is further restricted in such a way that the object must be indefinite, which for example means that pronouns (7a), proper names (7b) and definite DPs (7c) are unacceptable in the pre-head position (Hoekstra 1999, 268).

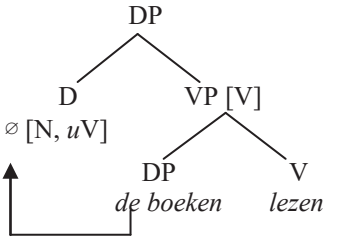
(7) (a) *Hen lezen vind ik saai.
them read find I boring

(b) *Jan opbellen heb ik geen tijd voor.
Jan call have.1.SG I no time for

(c) *De boeken lezen vind ik interessant.
those books read find I interesting

Since this phenomenon has not been otherwise accounted for in the literature, I propose to extend an idea of Jackendoff (1968) for percolation of "definiteness." I claim that the definiteness of the object should percolate to the VP as is illustrated under (8). This contradicts the indefiniteness of the bare nominal infinitive. This conflict then accounts for the acceptability judgments in (7).

(8)



Let's now look at the internal verbal and nominal properties of NI-Bs which can be tested by the modifiers that they take and by their ability to be pluralized, quantified and questioned.

With respect to modification, just like verbs, bare NIs can be modified by adverbs (9a). However, their adverbial status is sometimes questioned in the literature since the *-e* ending which marks adjectives (9c) appears only if an NP is determined by a definite determiner. Since there is no article with the NI-Bs, analysts waiver as to whether the lack of *-e* indicates adverbial status, or simply the lack of definiteness.

To illustrate the phenomenon, compare the following examples, which show that the word *goed* can be an adjective as well as an adverb depending on the preceding word. The *-e* ending that clearly marks *goed* as an adjective, and not an adverb, appears only if a definite article precedes it and the whole NP is thus definite. Compare: *goed luisteren* "listen well," *een goed boek* "a good book," but *het goede book* "the good book."

(9) (a) Frequent bomen kappen door de industrie is schadelijk.
frequently trees cut by the industry is harmful
"Cutting trees frequently by the industry is harmful."

(b) ?Frequent kappen van bomen door de industrie
frequently cut of trees by the industry
is schadelijk.
is harmful
"Cutting of trees frequently by the industry is harmful."

(c) *Frequente bomen kappen door de industrie is schadelijk.
frequent trees cut by the industry is harmful (Reuland 2011, 2)

Thus since NI-Bs in the preceding examples, unlike NI-Ds, need to be modified by the adverb *frequent* and not the adjective *frequente*, they must be verbal themselves.

Furthermore, with respect to the size of the verb, nominal infinitives of both types can contain auxiliary or modal verbs (10) while other types of nominalizations (e.g. derived nominals) exclude modals or auxiliaries as their input.

(10) (a) auxiliary verbs

Zo'n	boek	geschrijven	hebben	is	niet	genoeg.
Such a	book	write.$_{\text{P.PARTICIP.}}$	have	is	not	necessary
om	je	schrijver	te	noemen.		
to	you	writer	to	call		

"Having written such a book is not enough to call yourself a writer."

(b) modal verbs

Met	een	auto	kunnen	rijden	is	nodig.
with	a	car	can	drive	is	necessary

"Being able to drive a car is necessary."

Unlike countable nouns, bare NIs cannot co-occur with quantifiers (11a) and cannot be pluralized (11b) or questioned (11c) either. A sentence like *Veel sprookjes lezen elke dag is niet verstandig* "Reading a lot of fairytales every day is not sensible" would, however, be acceptable, since the quantifier clearly premodifies only the direct object itself and not the whole NI, which is also indicated by the agreement of the verb with a subject in singular.

(11) (a)

*Veel	sprookje	lezens	waren	saai.
many	fairy tale	reads	were	boring

(b)

*Peter	geniet	van	sprookje	lezens.
Peter	enjoys	of	fairy tale	reads

(c)

*Welk	sprookje	lezen	vind	je	het	leukst?
which	fairy tale	read	find	you	the	nicest

Different studies analyze the internal structure of NI-Bs differently. Looyenga (1992) for instance suggests that NI-Bs are internally IPs that appear in argument position. According to other studies (e.g., Hoekstra 1985) these constructions even have a PRO subject, a typical clausal property, which he claims is supported by the impossibility of examples such as (12).

(12)

iemand	geld	lenen	(* door Jan)
somebody	money	lend	(by John)

The analysis that I propose here is below in (13). Although, as explained above, NI-Bs are claimed to allow both the complement preceding the head (*bomen kappen* "cutting trees") as in (13a) as well as following the head (*kappen van bomen* "cutting of trees") as in (13b), the first "verbal" word order is preferred, unmarked and more frequent, probably because it is more economical for the bare nominal infinitive. Principles of economy are understood as in the Minimalist Program (Chomsky 1995) and favour simpler structures and prohibit superfluous steps in derivations.

Thus in order to utilize less structure, the verbal head of the NI-B can merge with the DP complement earlier, at the VP level, giving rise to the VP-type word order. The less frequent "nominal" word order (13b) is less economical, because it requires (i) a step where

the head changes its category from V to N and then (ii) a merge with a more complex *van*-phrase PP. Since DPs are for NI-Bs "cheaper" than PPs with DPs inside them, and Dutch makes it possible to express a DP argument with a V-headed construction (with no lexical N in the head position), it is more economical for the bare nominal infinitive V to merge with a DP rather than a PP complement later.[2]

Now to allow both possible word orders we have to use some operation that combines Merge, the central concept of the Minimalist Program, as well as some version of the "Categorial Switch" described by Panagiotidis and Grohmann (2009). However, in my view they interact and are not independent processes, which simplifies the operation and reduces it to mechanisms used elsewhere in the grammar. Merge tells us that only one complement/adjunct can enter a tree at a time, not two. And the patterns of NI-Bs explored above show us that in nominalizations, such constituents can merge either before a V becomes an N ("Categorial Switch") or after. In other words, with complements which are selected obligatorily in the lexicon, the satisfaction of selection (which is a property of LF) can "wait" until the final extended/highest projection of that lexical entry is reached. This scenario is thus a new type of independent evidence that all syntactic structure is binary branching, i.e., even lexically selected phrases enter trees one at a time.

The Switch Categorizer Hypothesis as formulated by Panagiotidis and Grohmann (2009) claims that between two types of domains in a derivation (e.g., verbal and nominal) there appears an additional "functional categorizer" that triggers a switch between the two categories. Moreover and crucially, the switch from one category to another can occur only once, so for example a change from verbal to nominal domain and then back to nominal again is not allowed.[3]

Applying this idea to nominal infinitives in Dutch, the switching category is a language particular lexical item, which must have an interpretable nominal feature [N] and an uninterpretable verbal feature [*u*V] that is checked against the interpretable feature [V] of the verbal chunk. In this way Categorial Switch brings about a change between the verbal and the nominal domain.

In my view, apparently counter to these authors, it is not necessary to postulate any new feature or category to effect the switch. In particular, although this is not uniform across languages, the lexical entry for the switching item in NI-B is just an

2 The same logic holds for Dutch APs without agreement (adverbs) which are "cheaper" than Dutch APs with agreement (adjectives).

3 The operation of Categorial Switch presupposes that, however complex the phrases might be themselves (e.g., the verbal phrase can in fact be the whole IP), they must remain coherent (Bresnan 1997). In other words the chunks making each phrase must be categorially uniform without any interspersed verbal elements within a nominal domain, or the other way around. This kind of stipulation is avoided in my model.

interpretable lexical D with an uninterpretable feature [*u*V] that ensures selection of an interpretable sister that is a verbal projection.[4] For NI-Bs, the lexical D is a null indefinite article. D, which like any functional category in the extended projection of N, has a nominal feature.

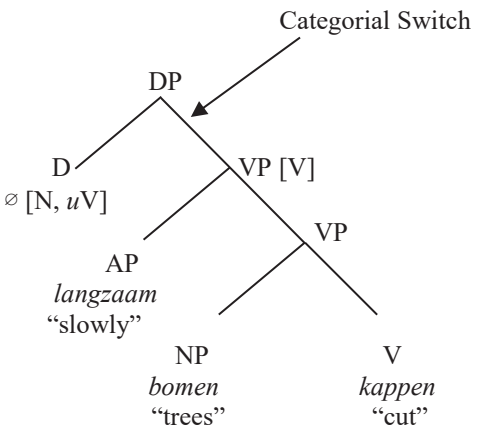
The main advantage of my approach is that I use the same formal mechanisms that other people use for selection—Merge. Another important principle of my theory is that complements and adjuncts of V are all optional unless a maximal extended projection is reached. The many examples presented here have shown repeatedly that this is true. And here we make use of it to explain why a V sister of an empty N can have unsatisfied selection features. These features can be satisfied in a subsequent derivational phase for DP, as will be exemplified below.

Finally, when we get to the maximal projection in case of NI-Bs, the D head will remain empty. Thus, because of the nature of Merge and the operation of Categorial Switch, the tree structure of the NI-B comes out automatically, a confirming result which has not previously been made explicit in other analyses.

The two examples below (13) are the two alternative options for the structure of NI-Bs:

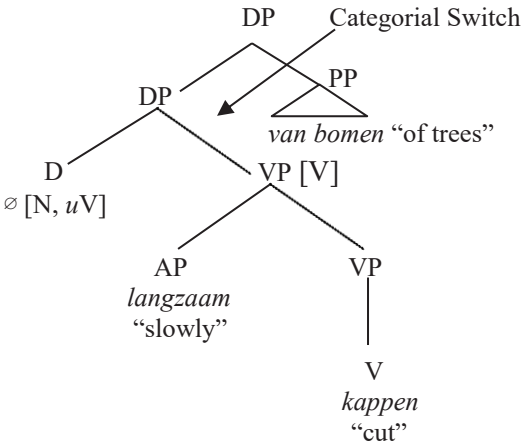
(13)

(a)



4 Note that this lexical entry D with an interpretable N feature and an uninterpretable V feature is missing in English.

(b)



Let's now consider the argument structure of a given NI-B. As mentioned before, bare nominal infinitives inherit their argument structure from the verb, and their thematic frame (14a) essentially remains unaffected by the derivational process. However, unlike with English verbs in a maximal verbal projection VP (in today's terms, a phasal domain vP), the arguments of an NI-B are not obligatorily expressed. Thus while the patient is most frequently realized as an NP in the pre-head position (14b), its realization can be delayed until the next phrase, where it possibly follows the nominal head in the form of a *van*-phrase (14c), although this is a more marked and less preferred option (as previously discussed).

(14) (a) Jan schrijft artikelen.
Jan write.3.SG artikels
"Jan writes articles."

(b) Artikelen schrijven kost veel tijd.
articles write cost.3.SG a lot of time
"Writing articles costs a lot of time."

(c) ?Schrijven van artikelen kost veel tijd.
write of articles cost.3.SG a lot of time
"Writing of articles costs a lot of time."

An agent phrase is neither a selected complement nor an adjunct in a VP, so it is not realized inside a maximal VP, whether the verb is intransitive (15a) or transitive (14a).

As an external argument of VP, if it is expressed, then it will be in a nominal projection, either following the head as a *van*-phrase (15b) or preceding it in the form of a possessive pronoun or a genitive noun phrase (15c).

(15) (a) Kinderen/Jan lachen/lacht.
children/Jan Jan laugh/laughs
"Children/Jan laugh/laughs."

(b) Lachen van kinderen was te horen.
laugh of children was to hear
"Laughing of children was to hear."

(c) Jans lachen was te horen.
John's laugh was to hear
"John's laughing was to hear."

A recipient, just like the agent is expressed optionally, and not necessarily in the verbal projection. In the nominal projection, it can appear either as an NP in the prenominal position (16b), or it can be realized as a PP in which case it will follow the patient and either appear in VP in the prenominal (16c) or in the NP in postnominal position (16d).

(16) (a) Jan schenkt geld aan de kerk.
Jan donates money to the church
"Jan donates money to the church."

(b) De kerk geld schenken is een goede zaak.
the church money donate is a good thing
"Donating money to the church is a good thing."

(c) Geld aan de kerk schenken is een goede zaak.
money to the church donate is a good thing
"Donating money to the church is a good thing."

(d) Geld schenken aan de kerk is een goede zaak.
money donate to the church is a good thing
"Donating money to the church is a good thing."

Verbs which select a PP complement can also be nominalized, and in this case with bare nominal infinitives, the PP complement will either precede a V head (17b) or follow an

N head (17c), but the more frequent and preferred word order is, as expected, the verbal one with the complement preceding the head.

(17) (a) Jan schiet op konijnen.
Jan shoots on rabbits
"Jan shoots at rabbits."

(b) Op konijnen schieten is een rare hobby.
On rabbits shoot is a strange hobby
"Shooting on rabbits is a strange hobby."

(c) Schieten op konijnen is een rare hobby.
shoot on rabbits is a strange hobby
"Shooting on rabbits is a strange hobby."

As mentioned before head nouns will never be preceded by a PP, while bare NIs may be (Hoekstra 1999, 267), which is another verbal property. Thus, if the PP complement precedes the head the merge must occur in the verbal domain prior to Categorial Switch, while if it follows the head, it must occur later in the nominal domain. The change from one to the other is affected by the indefinite empty head D which selects a verbal projection by means of an uninterpretable feature [*u*V]. The special property of Dutch is that this null lexical item in nominal infinitives seems indifferent to the level of the verbal projection. English has no such indefinite empty singular article, as is well known. Compare this to English where, unlike in Dutch, the PP complement will always follow a gerund.

2.2 Determined Nominal Infinitives

The second type of Dutch nominal infinitives, which I will discuss in this section, are determined nominal infinitives (NI-Ds). In comparison to NI-Bs, NI-Ds are internally a nominal construction with mixed nominal and verbal lexical heads (Zubizarreta and van Haaften 1988, 282). This can for example be shown by the fact that the object in determined NI-Ds can both precede the verb (18a), which is a property typical of VPs, as well as follow the infinitive as a *van*-phrase (18b), as is the case in NPs. However, unlike with NI-Bs, both of these forms are equally acceptable (Looyenga 1992; Hoekstra 1999; Reuland 2011, etc.).[5]

5 *Syntax of Dutch* (Broekhuis and Den Dikken 2012, 57) claims that the unmarked form is the exact opposite of bare nominal infinitives, that is with the object following the head in a DP. This opinion is not uniform in the literature.

(18) (a) Het boeken lezen vind ik vervelend.
the books read find I annoying
"I find the reading of books annoying."

(b) Het lezen van boeken vind ik vervelend.
the read of books find I annoying
"I find the reading of books annoying."

As these examples illustrate, this construction seems to be equivalent to both the "event" nominal and verbal English gerunds at the same time. Although it does not exactly exist in English, the closest counterpart would be the following example, where a limited list of determiners can take either gerund complements (19a) or derived nominals (19b):

(19) (a) John's/his/this/that/any/no reading books all night can be harmful.

(b) John's/his/this/that/any/no reading of books all night can be harmful.

On the other hand, determiners such as *some*, *each* or *every* are excluded in the gerund construction.

(20) * Some/each/every reading books can be harmful.

Chomsky (1970 does not treat such examples beyond mentioning them, and Emonds (2000) considers them peculiar and restricted.

When we test NI-Ds with modifiers, just like NI-Bs they preferably take adverbial modifiers, although some speakers accept both adjectives as well as adverbs as below in (21).

(21) (a) Het ?frequente/frequent bomen kappen door
the frequent/frequently trees cut by
de industrie is schadelijk.
the industry is harmful
"The frequent cutting of trees by the industry is harmful."

(b) Het ?frequente/frequent kappen van bomen
the frequent/frequently cut of trees
door de industrie is schadelijk.
by the industry is harmful
"The frequent cutting of trees by the industry is harmful."

In my analysis, the combination with an adjective can be explained by the fact that the [*u*V] selection feature is on both a lexical (free morpheme) D and the (bound) case inflection D on the adjective. However, this second word option is non-preferred and less economical.

However, if both an adjective and an adverb precede an NI-D, they must occur in the order Adj_Adv (22a), and the opposite results in an ungrammatical construction (22b).

(22) (a) Het irritante langzaam kappen van bomen was vervelend.
the irritating slowly cut of trees was annoying
"The irritating slow cutting of trees was annoying."

(b) *Het langzaam irritante kappen van bomen was vervelend.
the slowly irritating cut of trees was annoying

This fact can be easily explained by combining Merge and Categorial Switch, as the merge with an adverb has to occur lower down within the verbal domain (earlier in the derivation), while the merge with an adjective has to occur later, since the adjective carries the [*u*V] feature and triggers the Categorial Switch.

A similar principle can explain example (21a) where the whole verbal chunk consisting of the verb, its complement and the adverbial modifier can undergo the Categorical Switch together. The same holds for example (21b) with the difference that the obligatory complementation of the verb is satisfied later in the nominal domain by the "*van*-phrase."

The mixed properties of NI-Ds are also illustrated well by the fact that they can co-occur either with a PP modifier in the pre-head position, which requires a V category, or with a *van*-phrase, which requires an N category, in one construction (22). In this case again the PP modifier has to merge first in the verbal domain while the *van*-phrase merges after the Categorial Switch.

(22) Het met een mesje schillen van aardappels is gemakkelijk.
the with a knife peel of potatoes is easy
"The peeling of potatoes with a knife is easy."

Just like bare nominal infinitives, NI-Ds can contain complex verbal structures with a modal or auxiliary verb (23). This indicates that Categorial Switch can apply quite late in NIs.

(23) Het willen lezen van een boek is nodig.
the want read of a book is necessary
"The will to read a book is necessary."

The previous examples show that unlike in English, Dutch modals and auxiliaries are not in the I position but in the V position, and thus both the lexical and the modal/auxiliary verb undergo Categorial Switch together and then merge with the *van*-phrase higher up within the nominal domain.

Unlike ING- and GE-nominalizations, both types of NIs allow modals and auxiliaries in the nominalizations. However, bare nominal infinitives must realize the complement of the lexical verb as an NP in the pre-head position, while determined nominal infinitives as a *van*-phrase (Broekhuis and Den Dikken 2012, 50).

On the other hand, just like regular nouns NI-Ds can be determined by articles or demonstrative pronouns (24a), but unlike countable nouns, they can never be quantified (24b), pluralized (24c) or questioned (24d). These restrictions show that both Dutch bare and determined nominal infinitives are like neuter (or perhaps genderless) mass nouns; their indefinite and definite Ds are respectively ∅ and *het*.

(24) (a) Dat constant roken van cigaren was irritant.
that.NEUT constant smoke of cigarettes was irritating
"That constant smoking of cigarettes was irritating."

(b) *Veel lezens van boeken waren verschillend.
many reads of books were different

(c) *Peter geniet van de lezens van boeken.
Peter enjoys of the reads of books

(d) *Welk lezen van boeken vind je het leukst?
Which read of books find you the nicest

The internal structure of NI-Ds has been analyzed differently in previous studies. The most detailed analysis is provided by Looyenga (1992), whose determined NIs are NPs that consist of an IP to which an affix expressing nominal features has been attached. This affix which carries the nominal features has no morphological realization. The affix provides the NI-Ds with nominal characteristics and gives it, together with the determiner, the internal grammar of a nominal phrase (Looyenga 1992, 178). However, neither he nor the other authors explain why two word orders in both types of nominal infinitives are possible, although one word order is always preferred. Unlike Looyenga, I do not think that IP level is needed and consider the Categorial Switch analysis more systematic and elegant and thus an advantage over his approach.

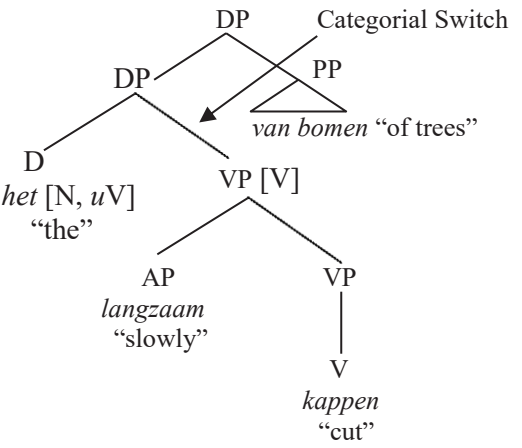
In my view, no redundant null affix expressing nominal features is needed, because the same operation that combines Merge and Categorial Switch that applies for NI-Bs is at work in NI-Ds too (25). As discussed in Section 2.1., just like bare nominal infinitives,

determined nominal infinitives also have two possible word orders which are alike in both types of NIs; one with the complement following the head (*het langzaam kappen van bomen* "the slow cutting of trees") (25a) and the other with the complement preceding the head (*het langzaam kappen van bomen* "the slow cutting of trees") (25b). Although both word orders are acceptable, the preferred order is the former one which copies the internal word order of DPs, since for determined NIs it is the more economical version. In this word order the merge with the complement has to occur higher up in the tree structure after *kappen* has switched its category from a V to an N so that it can merge with a *van*-phrase. In the latter word order, the merge between the head and the complement has to occur lower down at the V level. With respect to their preferred word orders NI-Ds are the opposite of NI-Bs since with determined nominal infinitives with a mixed nominal and verbal head, PPs (*van*-phrases) are required by the definiteness of the NI and apparently therefore "cheaper" and more economical than DP complements that are selected by Vs.

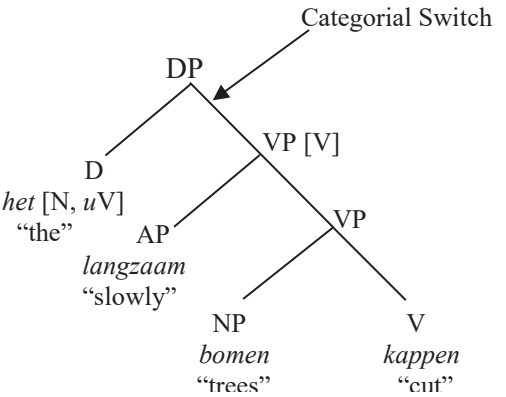
Again with Merge and Categorial Switch, as explained in the preceding section, the tree structure comes out right. Unlike in NI-Bs, the D head does not remain empty but is occupied by the definite article *het.* The definiteness of NI-Ds also clarifies why the DP word order is more economical. Compare the two optional word orders of NI-Ds below in (25).

(25)

(a)



(b)



Since the internal structures of NI-Bs and NI-Ds are very similar with only the difference that a different word order is preferred (not obligatory) in each type, with respect to their argument structure, NI-Ds seem to be the definite counterpart of the indefinite bare NI-Bs with their thematic frame inherited from the verbs they are derived from (26a) (Broekhuis and Den Dikken 2012). Thus the patient will most frequently follow the head N in the form of a *van*-phrase (26b), since this is the preferred word order, or possibly appear in front of the head verb as an NP (26c), as discussed above.

(26) (a) Jan leest boeken.
John read.3.SG books
"John reads books."

(b) Jan geniet van het lezen van boeken.
John enjoy.3.SG of the read of books
"John enjoys the reading of books."

(c) Jan geniet van het boeken lezen.
John enjoy.3.SG of the books read
"John enjoys the reading of books."

Just as with NI-Bs the agent does not have to be expressed if the verb is intransitive (27a). However, if it is expressed it may either precede the head NP as a genitive NP or a possessive pronoun (27b) or follow it in the form of a *van*-phrase (27c).

(27) (a) Jan/kinderen lacht/lachen.
John/children laugh.3.SG/laugh
"John/children laughs/laugh."

(b) Jans lachen was te horen.
John's laugh was to hear
"John's laughing was to hear."

(c) Het lachen van kinderen was te horen.
the laugh of children was to hear
"The laughing of children was to hear."

If the verb is transitive (28a), both arguments may be expressed. The patient will preferably appear as a *van*-phrase, and the agent will be expressed either in the form of a possessive pronoun or a genitive noun phrase (28b) or it can follow the head in the form of a *door*-phrase (the Dutch equivalent of the English *by*-phrase) (28c). Note that the *door*-phrase is not allowed with bare nominal infinitives. A less frequent (more marked) but also possible realization is with a patient preceding the head and an agent following it in the form of a *van*-phrase (28d).

(28) (a) Jan verzmelt postzegels.
Jan collect.3.SG. stamps
"John collects stamps."

(b) Jans/Zijn verzamelen van postzegels is tijdrovend.
John's/his collect of stamps is time-consuming
"John's collecting of stamps is time-consuming."

(c) Het verzamelen van postzegels door Jan is tijdrovend.
the collect of stamps by John is time-consuming
"John's collecting of stamps is time-consuming."

(d) ?Het postzegels verzamelen van Jan is tijdrovend.
the stamps collect of Jan is time-consuming
"John's collecting stamps is time-consuming."

The recipient may also be optionally expressed. In case of NI-Ds the recipient must follow both the head and the patient and cannot precede them (29), which means that the [def] feature must be added prior to satisfying the complement selection feature.

(29) Het schenken van geld (aan de kerk) is een goede zaak.
the donate of money to the church is a good thing
"The donating of money to the church is a good thing."

Furthermore, determined NI-Ds appear when the patient is expressed as a *van*-phrase in a generic example (30).

(30) Het vallen van bladeren gebeurt elk najaar.
the fall of leaves happens every autumn
"The falling of leaves happens every autumn."

With verbs which select PP-complements, although both word orders are acceptable (31a–b), there is a clear preference for placing the PP-patient in post-head position in NI-Ds (31b), unlike with NI-Bs.

(31) (a) Het op konijnen schieten is een rare hobby.
the on rabbits shoot is a strange hobby
"The shooting on rabbits is a strange hobby."

(b) Het schieten op konijnen is een rare hobby.
the shoot on rabbits is a strange hobby
"The shooting on rabbits is a strange hobby."

This is clearly a nominal property, since nouns in Dutch will be followed by a PP complement, while verbs will be preceded by it. The preferred word orders of PP-complements obviously copy the preferred word orders with respect to NP-patient complements of both NI-Bs and NI-Ds in (13) and (25), so the definiteness of the NI-D also affects the preference for the nominal word order. The merge with the PP complement in the preferred word order in (31) must occur higher up in the tree structure in the nominal domain, after the head changes its category from a V to an N.

3. Conclusions

When comparing Dutch bare nominal infinitives (NI-Bs) and determined nominal infinitives (NI-Ds) we have seen that both of them are externally DPs with all their typical syntactic functions, but internally they differ. While NI-Bs have mostly verbal properties, NI-Ds have mixed nominal and verbal properties. These predominant nominal or verbal properties of both types of nominal infinitive are also reflected with respect to definiteness, the modifiers that they take and the realization of arguments, especially that of patient.

The main focus of this paper has been the internal structures of Dutch nominal infinitives that result from a syntactic operation that combines Merge and Categorial Switch. I have shown that the fact that NIs can combine with their complements in two different ways can be easily explained by Merge, which is part of Universal Grammar. The V selects its obligatory complement(s) but the satisfaction of the selection can either occur lower down in the tree structure still within the verbal domain, or this can wait until the final highest/extended projection of that lexical item and occur higher up in the nominal domain. The operation of Categorial Switch, introduced before by Panagiotidis and Grohmann (2009), is in my view a language particular phenomenon which is triggered by items of the category D that carry the uninterpretable [*u*V] feature. Categorial Switch in combination with Merge provides a very systematic and logical treatment of the Dutch nominal infinitives, even though they have been previously described as "notoriously difficult to analyze" (Schoorlemmer 2001).

Works Cited

Bresnan, Joan. 1997. "Mixed Categories as Head Sharing Constructions." In *Proceedings of the LFG97 Conference*, edited by Miriam Butt and Tracy Holloway King. Stanford, CA: CSLI Publications.

Broekhuis, Hans, and Marcel den Dikken. 2012. *Syntax of Dutch*. Amsterdam: Amsterdam University Press.

Chomsky, Noam. 1970. "Remarks on Nominalization." In *Readings in English Transformational Grammar*, edited by Roderick Jacobs and Peter Rosenbaum, 184–221. Waltham, MA: Ginn and Company.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Emonds, Joseph. 2000. *Lexicon and Grammar: The English Syntacticon*. Berlin: Mouton de Gruyter.

Hoekstra, Teun. 1999. "Parallels between Nominal and Verbal Projections." In *Specifiers: Minimalist Approaches*, edited by David Adger, Susan Pintzuk, Bernadette Plunkett, and George Tsoulas, 163–87. Oxford: Oxford University Press.

Jackendoff, Ray. 1968. "Quantifiers in English." *Foundation of Language* 4 (4): 422–42.

Looyenga, Sietze. 1992. "A Syntactic Analysis of Dutch Nominal Infinitives." In *Linguistics in the Netherlands*, edited by Reineke Bok-Bennema and Roeland van Hout, 173–84. Amsterdam: John Benjamins Publishing Company.

Panagiotidis, Phoevos, and Kleanthes K. Grohmann. 2009. "Mixed Projections: Categorial Switches and Prolific Domains." *Linguistic Analysis* 35: 141–61.

Reuland, Eric. 2011. "What's Nominal in Nominalizations." *Lingua* 121 (7): 1283–96.

Schoorlemmer, Maaike. 2002. *Dutch Nominalised Infinitives As Non-Identical Twins.* Utrecht: Utrecht Institute of Lingustics OTS.

Zubizarreta, Maria Luisa, and Ton van Haaften. 1988. "English *-ing* and Dutch *-en* Nominal Constructions: A Case of Simultaneous Nominal and Verbal Projections." In *Morphology and Modularity*, edited by Martin Everaert, 361–94. Dordrecht: Foris.

Explaining Bobaljik's Root Suppletion Generalization as an Instance of the Adjacency Condition (and Beyond)

Pavel Caha

Masaryk University, Brno, Czech Republic

pcaha@mail.muni.cz

Abstract: Bobaljik (2012) observes on the basis of an impressive sample of languages that root suppletion is hardly ever conditioned by degree markers that do not form a word with the root. He calls this the Root Suppletion Generalization (RSG). If true, the generalization provides a possible argument for the lexicalist position: RSG can be seen as a consequence of the lexicalist architecture, where words are built pre-syntax, and therefore syntax cannot influence their shape (Williams 2007). Against this background, this paper discusses evidence (some of it presented already in Bobaljik's work) that the RSG (when stated over words) is empirically (sometimes) too weak and (sometimes also) too strong. In view of these observations, I suggest a way in which all of these examples can be captured in ways that do not lend any support to lexicalism, simply because the word is not the relevant notion for blocking suppletion.

Keywords: suppletion; comparatives; RSG; adjacency; words

1. Introduction: What Is RSG and How to Account for It?

This paper is an attempt at a reformulation of a generalization proposed in Jonathan Bobaljik's (2012) book. The book presents a theoretically oriented discussion of a large wealth of empirical data, focusing primarily on the attested and unattested patterns of root suppletion in adjectival degree expressions (of the sort *good*, *better*, *best*). Among the core generalizations of the book, we find the so-called Root Suppletion Generalization (RSG), which is given in (1) below.

(1) The Root Suppletion Generalization (Bobaljik 2012, ex. 3)
Root suppletion is limited to synthetic (i.e., morphological) comparatives.

The generalization (descriptively) rules out certain patterns of root suppletion as impossible, while allowing others. Specifically, when the comparative is expressed as an affix on the root, root suppletion is possible, see (2a). However, comparatives can also be periphrastic (2b). The RSG says that periphrastic marking never triggers root suppletion; so cases like (2c) should be unattested. Needless to say, Bobaljik's impressive sample of data supports that generalization almost without a single exception.

(2)		POS	CMPR	
	(a) English	good	bett-er	(morphological)
	(b) English	intelligent	more intelligent	(periphrastic)
	(c) the RSG rules out:	intelligent	more comptus	(periphrastic)

Theoretically, Bobaljik implements RSG in a Distributed Morphology architecture as a restriction on Vocabulary Insertion rules. VI rules specify how syntactic nodes are to be pronounced, and in doing so, they may inspect the surroundings of the particular node. For instance, the adjectival roots *good* and *bett-* seen in (2a) would have their specifications as shown in (3a, b).

(3) (a) *good* = A (b) *bett-* = A / _ CMPR

(3b) is a contextually dependent spell-out rule for the adjectival root, which applies when the root is embedded under a CMPR head. In this setting, where suppletive forms are introduced by contextually specified VI rules, Bobaljik proposes that if the search space of VI rules would be somehow restricted to synthetic forms, then the RSG follows.

Following the standard stance in DM and much other research, Bobaljik understands synthetic forms (i.e., words) as complex heads, formed either by head-movement or Merger. Hence, the concrete proposal is to restrict the search space to heads; see (4).

(4) Head locality: Spell-out rules cannot look outside of a complex head
(a) * alpha . . .]$_{XP}$ . . . beta
(b) ok alpha . . .]$_{X^0}$ . . . beta

This specific account of (1) raises some theoretical issues. For instance, Julien (2002), Koopman (2005) and a number of other researchers argued that words do not necessarily

correspond to heads. If that is so, the question is how to account for (1) without making reference to the head-phrase distinction.[1]

In a still wider theoretical perspective, the very statement (1) is actually surprising for any "neo-constructivist" theory that has adopted the move from lexicalism to something like "syntax all the way down." In order to see why that is so, consider the following passage from Edwin Williams' (2007) paper *Dumping Lexicalism.* He writes:

> The Lexical Hypothesis is about the organization of the grammar into modules. It suggests that the system of words in a language is independent of the system of phrases in a language in a particular way. . . . The essence of the hypothesis is the separation of the two systems and the asymmetric relation between them . . . [S]pecifically, the channel of communication is asymmetrical, by virtue of the fact that phrases are made out of words, but not vice-versa.
>
> The encapsulation prevents analyses. It narrows the scope of word/phrase interaction. For example, the parts of a word are not accessible in the phrasal system . . . From this flows many mundane but important facts. (Williams 2007)

As far as I can tell, the RSG—if correct—would be one of these "mundane but important facts." If looked upon from the lexicalist perspective, the generalization says that the shape of the morphemes in a word cannot be influenced by a category expressed outside of that word, as in (2c). However, if the very same category is expressed inside that word, it does have the power to influence the shape of other morphemes inside that word (2a). This perfectly instantiates the logic of "information encapsulation," which is at the heart of the lexicalist framework. So the question is how strong the empirical motivation for (1) actually is.

In this paper, I discuss a couple of data points which suggest that (1) can perhaps be stated in different terms. This is possible due to the extreme clarity with which Bobaljik presents and discusses his data, which much of this discussion heavily depends upon. Specifically, in Section 2, I suggest that some of the core examples discussed in Bobaljik's book in support of (1) are quite likely indecisive, because they are already ruled out by an independent condition, namely the Adjacency Condition. The crucial theoretical difference in explaining the patterns by adjacency is that adjacency is a concept that does not need to make a distinction between words and phrases, or heads and non-heads; in other words, there is no clear point in favor of the Lexicalist Hypothesis

1 In Bobaljik and Harley (forthcoming), the constraint on suppletion is shown to be actually compatible with word-phrase interactions; the idea is that suppletion can be triggered by elements that are inside the maximal projection of the root (and these can be phrasal). I ignore these later developments in this paper, and focus rather on the antecedent issue of whether (1) is the right way to look at the data to begin with.

to be made on the basis of such examples. In Section 3, I turn to Bulgarian, where Bobaljik (2012) found a surface counterexample to the RSG, one where suppletion seems to be triggered across a word boundary. I show how this particular example may be better explained under the adjacency-based reformulation of RSG.

In Section 4, I turn to additional data from Czech and argue that in this language, there seem to be cases where suppletion is blocked "inside" words in a way that is reminiscent of (2c), strengthening the point that the boundaries of words and boundaries for suppletion actually diverge. I follow this track and suggest that these cases can ultimately be attributed to the analytical/fusional expression of categories.

2. "Core RSG" as an Instance of Adjacency

Let us now turn to some of the core data that motivate Bobaljik's proposal. The strongest evidence in favor of RSG apparently comes from languages where periphrastic and morphological comparatives can be formed side by side. A couple of examples is given in (5). In these examples, comparative markers are set in bold.

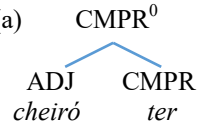
(5)	data from Bobaljik (2012)		POS	CMPR	
	(a) Greek	"good"	kak-ós	cheiró-**ter**-os	(morphological)
			kak-ós	**pjo** kak-ós	(periphrastic)
	(b) Georgian	"good"	k'arg-i	u-k'et-**es**-i	(morphological)
			k'arg-i	**upro** k'arg-i	(periphrastic)

What we see in these languages is that suppletion is found in cases where the comparative marker is an affix on the root. When the comparative marker is a separate word, no suppletion takes place. Obviously, the reason for the regular forms is not that the language would lack a suppletive root in the lexicon. We know that there is one (because we see it in the morphological comparative), it is simply not used. For instance, the Greek pair *kak*-ós—*chiró-ter-os* "good"—"better" tells us that there must be two VIs as shown in (6):

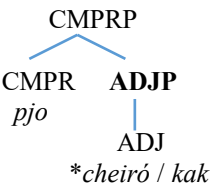
(6) (a) *kak* = ADJ
(b) *cheiró* = ADJ / _ CMPR

Now given the presence of suppletive lexical items in the relevant languages, the question is why they are not used in the periphrastic case. Bobaljik's idea, expressed in (1) and implemented in (4), is that this does not happen because the morphological form corresponds to a single head, see (7a), while the periphrastic form contains a phrasal node, boldfaced in (7b). This node intervenes in between ADJ and CMPR and blocks suppletion.

(7) (a)

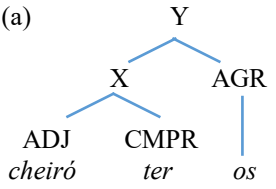


(b)

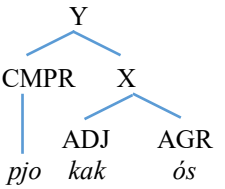


However, the two examples are not such a neat minimal pair as indicated in (7). Specifically, in both Greek and Georgian, there is also an agreement marker, which the parses in (7) simply ignore. It is possible that the agreement marker indeed plays no role, but the argument for (1) is exactly as strong as that assumption. If, on the other hand, the agreement marker is present in the structure, then the structures look like in (8a, b) respectively. The non-terminals in the trees are labelled in a way such that the labelling avoids making any reference to the head-phrase distinction.

(8) (a)



(b)



If (8) is the right way to depict the structures, then there is an additional difference between the two cases. Specifically, ADJ and CMPR are included in a single constituent (labelled X) in (8a), but there is no such constituent in (8b). In derivational terms, this means that the CMPR marker *-ter* is combined directly with the root, whereas the derivation of the periphrastic form has to first combine the root with *-ós* and add the comparative only later on. This in turn leads to the conclusion that in the periphrastic case, the root is never combined with the CMPR marker directly, and their interaction may therefore be blocked for this reason. If correct, this could be seen as an instance of the Adjacency Condition (Siegel 1978), according to which (in simple terms) the interaction between morphemes is only allowed if they attach one after the other.

Let me add that the adjacency condition is independently used in Bobaljik's work to rule out suppletion in cases which are analogous to (8b). For instance, the adjective *good-ly* has the comparative *good-li-er* and not **bett-li-er*, because here the comparative morpheme *-er* is separated from the root by *-ly*. What I suggest, then, is that the account of *good-li-er* is simply extended to cases such as (8b), where the role of the intervening *-ly* is taken on by the agreement marker. If this analysis is correct, the generalization in (1) is a red herring; what matters in cases such as (5) is not the

fact that the comparative marker is outside of the root's word, but the (structural and derivational) separation of the root from the comparative by an agreement marker.

3. Bulgarian *po-veče* "more"

The natural thing to do now is to look at cases where agreement is missing, as in English and other languages like that. If (1) is a side-effect of agreement intervention, we should find cases where—in the absence of agreement—suppletion can be triggered across a word boundary. One such case is in fact found in Bulgarian—and discussed in Bobaljik's book as a potential counterexample to (1)—and I turn to this example presently.

The first thing to note is that Bulgarian is a language where adjectives generally agree with the head noun, as illustrated in (9).

(9) (a) dobăr-ø măž
good-m man
"a good man"

(b) dobr-a žen-a
good-f woman-f
"a good woman"

(c) dobr-o det-e
good-n child-n
"a good child"

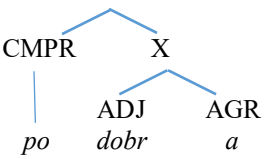
Comparatives are formed by putting the marker *po* to the left of the agreeing adjective. Bobaljik independently shows that the marker *po* is a phrasal marker that can attach to a variety of categories, some of them obviously phrasal; see (10a, b) for examples.

(10) (a) na jug
to/on south
"to the south"

(b) po na jug
po to/on south
"more southerly (more to the south)"

Given these facts, both adjacency and word-locality predict that there should be no suppletion in Bulgarian. From the perspective of RSG, this is because the comparative marker is periphrastic. Form the perspective of the adjacency-based explanation, this is because the agreement marker is closer to the root than the compartive marker, and blocks their interaction, see (11).

(11)



This is borne out, and there are no agreeing suppletive adjectives in BG. So, for instance, one of the most frequently suppletive root (judging from Bobaljik's sample of languages) is the root for the meaning "good," whose positive forms are in (9). The comparatives are shown in (12), and we see no root suppletion.

(12) (a) po-dobăr-ø măž
po-good-m man
“a better man”

(b) po-dobr-a žen-a
po-good-f woman
“a better woman”

(c) po-dobr-o dete
po-good-n child
“a better child”

This contrasts with the majority of the other Slavic languages, where comparative markers attach to the root, and may trigger suppletion. To illustrate this, let me turn to Czech which we will look at in the next section in more detail. What we see in this language is that the comparative marker *-š* comes in between the root and the agreement marker, see (13).

(13) (a) star-ého muže
old-m.gen man.gen
“of an old man”

(b) star-š-ího muže
old-er-m.gen man.gen
“of an older man”

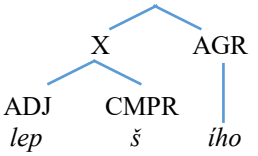
Given this, we expect that root suppletion is possible in Czech. And this expectation is borne out, see (14). Note that the positive in (14a) is obviously cognate with the BG root.

(14) (a) dobr-ého muže
good-m.gen man.gen
“of a good man”

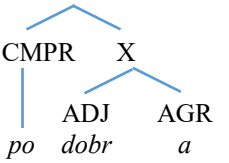
(b) lep-š-ího muže
bett-er-m.gen man.gen
“of a better man”

The difference between the Czech and Bulgarian comparatives is thus exactly the same as the difference between the two different ways of forming comparatives in Greek, compare (15) with (8), and both theories make the same predictions.

(15) (a) CZ



(b) BG



A difference appears when we look at non-agreeing modifiers. A case in point are quantificational adjectives like “much”—“more”—“most.” These show no agreement in BG:

(16) (a) mnogo snjag
much snow.m
“a lot of snow”

(b) mnogo voda
much water.f
“a lot of water”

(c) mnogo mljako
much milk.n
“a lot of milk”

Given the absence of agreement, the adjacency based theory would seem to allow for an exception to the general pattern of non-suppletion in exactly such cases. And that is in fact what we find; the form meaning "more" is suppletive in Bulgarian (and non agreeing):

(17) (a) po-veče snjag
po-more snow.m
"more snow"

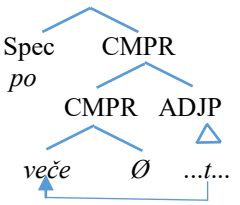
(b) po-veče voda
po-more water.f
"more water"

(c) po-veče mljako
po-more milk.n
"more milk"

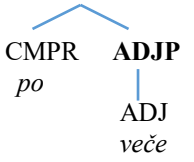
Such data are surface problematic for the RSG. Here it seems that a comparative marker that is phrasal, as in (10b), apparently triggers suppletion across a word boundary. If one wants to show that phrasal syntax has the power to influence the shape of morphemes inside words, this is the kind of example one would want to find.

Theoretically, it seems tempting to attribute to the comparatives in (17) the structure in (18b), where the comparative *po* takes the ADJP as a complement (with no AGR present). This structure, is, however, identical to (7b), so the phrasal ADJP is expected to block suppletion per (4). Under an adjacency-based account, such an expectation does not arise, and we correctly allow the interaction between the two markers.

(18) (a)



(b)



Bobaljik accommodates this example by proposing a slightly more complex structure for BG. He suggests that *po* is not the true comparative marker, but rather an obligatory reinforcer, which occupies a higher position in the tree than CMPR, perhaps the Spec position, as shown in (18a). In such an analysis, the true comparative marker is silent, and forms a complex head with the adjective. The silent comparative marker (rather than the overt one) is then the real trigger for suppletion.

The account clearly works for the suppletive case, but it no longer explains why BG is special in the context of Slavic, recall the contrast between (12) and (14). The initial insight was that BG differs from related languages like Czech because it has a comparative marker which is periphrastic; that is why BG has so little suppletion compared to the related languages. But this explanation is now lost; according to the new analysis in (18), BG also has a word internal comparative marker. Hence, the proposal now fails to explain the observed contrast between BG and the majority of other Slavic languages. The adjacency based alternative fares well: the phrasal nature

of the comparative marker leads to it appearing outside of agreement, which (when present) blocks suppletion. When agreement is absent, suppletion may still arise.

The place where this brings us is that the evidence for proposing something like the RSG as an independent generalization (over and above the Adjacency Condition) is weakened. The empirical record in favor of RSG which remains after agreement intervention is admitted to be a potential confound, needs to be re-established and re-evaluated, a task which is beyond the scope of this paper. It is clear though that some cases will remain; for instance, the English pattern in (2) does not seem to be due to agreement intervention. What can be said about such cases? The following section presents a short case study of Czech comparatives that may have some bearing on the answer.

4. A Restriction on Suppletion in Czech Comparatives

BG has provided us with a case where the RSG seems to be too restrictive: comparative markers may—in special (and admittedly rare) cases—apparently trigger suppletion across a word boundary. In this section, I discuss data from Czech suggesting that the RSG may also be too permissive. Specifically, I argue that in Czech, there is a systematic restriction on suppletion that is in a way analogous to (2c), but which in fact restricts suppletion inside a single word. This will lead me to formulate a generalization that will apply to both English and Czech, and make the RSG superfluous in (2).

Let me then turn to the Czech data which are going to be crucial for what follows. Below in (19) I give a couple of adjectives in their positive and comparative degree. What we see is that the comparative is formed by attaching *-ějš* to the root. The sign *ě* corresponds to an *e* which triggers the palatalization of the preceding consonant, a process which only happens "word internally." The bracketed segments are concord markers.

(19)	gloss	POS	CMPR
	fast	rychl-(ý)	rychl-ejš-(í)
	red	červen-(ý)	červen-ějš-(í)
	stupid	hloup-(ý)	hloup-ějš-(í)
	wild	bujar-(ý)	bujař-ejš-(í)

However, there are reasons to think that *-ějš-* should be split into two morphemes, *-ěj* and *-š*, because each of these markers leads an independent life. The first piece of evidence for this comes from comparative adverbs, seen in the second column of (20). Here the *-š-* part of the comparative adjective is systematically missing.

(20)	gloss	CMPR ADJ	CMPR ADV (no š)
	fast	rychl-ej-š-(í)	rychl-ej-(i)
	red	červen-ěj-š-(í)	červen-ěj-(i)
	stupid	hloup-ěj-š-(í)	hloup-ěj-(i)
	merry	bujař-ej-š-(í)	bujař-ej-(i)

The absence of *-š* is hard to attribute to phonology, because the adverbial marker *-i* has the same quality as the agreement marker *-í*, and the two differ only in length. Therefore, it seems necessary to separate the comparative marker *-ějš* into two parts, *-ěj* and *-š*. The description then says that the first part of the comparative is preserved in the adverb, while the second part is lost. The separation of the comparative into two markers is similar to Bobaljik's proposal for BG. The difference is that neither *-ěj* nor *-š* can be considered a Spec, because they are both inside one and the same word. Therefore, I propose that in Czech (and probably more generally) there are two comparative heads, CMPR1 and CMPR2; see also the Georgian morphological comparative in (5b).[2, 3]

The second thing to note concerning the separation of *-ěj-* and *-š-* is the fact that some adjectives lack the first part of the comparative marking and only have the second part. (Velars are subject to palatalization before *-š*.) Again, this points to the conclusion that *-ěj* and *-š* are separate, because some forms lack one but have the other.

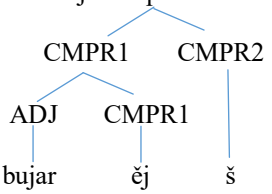
(21)

gloss	POS	CMPR (no -ěj-)
old	star-(ý)	star-ø-š-(í)
hard	tvrd-(ý)	tvrd-ø-š-(í)
expensive	drah-(ý)	draž-ø-š-(í)
silent	tich-(ý)	tiš-ø-š-(í)

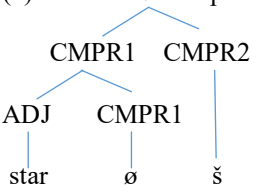
Let me suppose, for the start, that there is simply a zero allomorph of the CMPR1 *-ěj*, as indicated in the second column. The structures of the two types of comparatives would then look as follows:

(22)

(a) The -ěj- comparative



(b) The ø comparative



2 For languages, where we only see one of them, we can consider the other as null for the moment; later, I will develop a phrasal-spell-out account of the phenomenon, proposing that CMPR1 and CMPR2 may be pronounced by a single morpheme.

3 A virtually identical approach is proposed in DeClerq and Vanden Wyngaerd (2016) and embedded within a more general account of adjectival meaning.

The question I turn to now is whether the facts and generalizations that we have seen up to now lead us to expect anything about the distribution of suppletion in the two sets of cases. The answer is, I think, "no." There is no reason why CMPR1 *-ěj* should refuse to trigger suppletion, or why its silent counterpart should do so. Similarly, if there is in fact no silent CMPR, and the CMPR1 node is radically missing, then *-š*, being both adjacent to the root as well as being word internal, should be able to trigger suppletion.

However, contrary to the expectations, there is in fact an asymmetry in suppletion patterns between (22a) and (22b). Specifically, suppletion takes place only in cases like (22b), but never in those which are like (22a). The data in (23a) illustrate this for the roots which are "radically" suppletive, the data in (23b) illustrate this for "mildly" suppletive roots. (I consider them both suppletive, endorsing a theory without morphologically triggered readjustment rules.) Just for completeness, I give the forms of the comparative adverbs. These essentially retain suppletion (and are subject to palatalization and vowel lengthening), but lack the *-š* just like their regular counterparts.

(23)		gloss	POS	CMPR	CMPR.ADV
	(a)	good	dobr-ý	lep-š-í	lép-e
		bad	špatn-ý	hor-š-í	hůř-e
	(b)	small	mal-ý	men-š-í	mén-ě
		big	velk-ý	vět-š-í	víc-e

So the generalization is that there is an asymmetry such that *-š* comparatives allow suppletion (adverbs even in the absence of *-š*), while *-ěj-š* comparatives do not. The generalization can be stated in the following shape:

(24) The Czech suppletion generalization (CSG)
When the comparative degree is expressed by two overt morphemes in addition to the root, there is no suppletion.

I will try to implement this generalization theoretically in the next section. What is relevant for me now is that if the CSG were generalized beyond Czech, it would also be relevant to the English examples in (2), repeated below in (25).

(25)		POS	CMPR	
(a)	English	good	bett-**er**	(mono-morphemic)
		intelligent	**mo**-**re** intelligent	(bi-morphemic)
(b)	the CSG rules out:	intelligent	mo-re comptus	

In order to show the relevance of (24), I have to first make explicit an analysis of *more* which I am assuming, namely that *mo-re* is actually bi-componential, corresponding

to the comparative form of *much* (Corver 1997; Bobaljik 2012, a.o.). If that is so, the original data set from (25) shows not only an asymmetry in terms of word internal/external expression of the comparative, but also an asymmetry in terms of complexity. Specifically, in *mo-re intelligent*, the comparative is expressed by the combination of two markers, and hence the phrase is an instance of a "bi-morphemic" comparative. This means that the lack of suppletion in these cases may be the consequence of (24).

Summing up this section: in Czech, there is a restriction on root suppletion that is unrelated to the word/phrase distinction, but seems to care instead about how many pieces of morphology there are in the comparative. This generalization—applied to the case of English—yields the same cut between *A-er* comparatives and *mo-re A* comparatives as the RSG. Since the latter are bi-morphemic, they are expected to trigger no suppletion. Importantly, the blocking of suppletion has nothing to do with whether the two comparative morphemes are in the same word or not. They happen to be so in Czech, but not in English; yet this is irrelevant for how the condition is applied. In the next section, I turn to some ideas as to what theory may lie behind the CSG.

5. The Underpinnings of the Czech Suppletion Generalization

In Bobaljik's (2012) book, there are two ways to be suppletive. For some pairs of roots, Bobaljik proposes that the suppletive form corresponds to a lexical item which spells out a complex non-terminal containing the ADJ node and the CMPR feature. Such a pair is for instance *bad* and *worse*, seen in (26a, b). This seems to me an intuitive way of encoding that *worse* conveys the meaning of both *bad* and the meaning of CMPR.

(26) Two ways to suppletion in Bobaljik (2012)

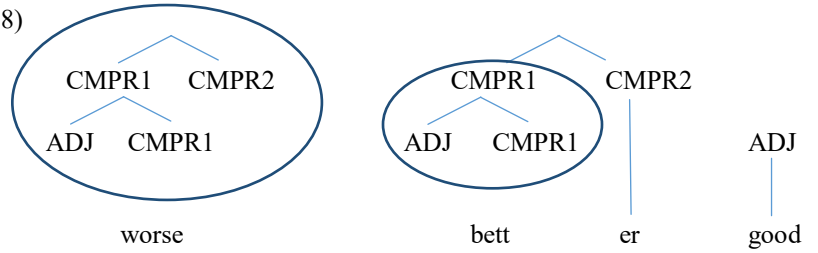
(a) ADJ <=> bad
(b) [ADJ CMPR] <=> worse
(c) ADJ <=> good
(d) ADJ \ _CMPR <=> bett-

However, for pairs such as *good–bett-er*, Bobaljik finds this account unsatisfactory. That is because in the suppletive form *better*, it is only the *bett-* part which is suppletive, while the *-er* part is fully regular. Therefore, Bobaljik proposes that in *bett-er*, *-er* spells out the CMPR node as usual, which only leaves the ADJ node for spell out. Hence, there must be a second way to suppletion, which is provided by rules such as (26d). These rules say that the form of the root meaning "good" is *bett-* in the context of CMPR. These lexical entries produce structures such as (27), where the arrow indicates that insertion under ADJ is sensitive to the presence of CMPR.

(27)

However, once CMPR is split into two parts, it is no longer necessary to use two distinct mechanisms for suppletion. The difference between *bett-* and *worse* can be modelled by the proposal that they differ in how many CMPR heads they spell out. Specifically, *worse* spells out both CMPR1 and CMPR2 with the adjectival root, while *bett-* spells out only the lower CMPR1. This is shown in (28). For clarity, I also illustrate the tree for the simple positive form *good*.

(28)



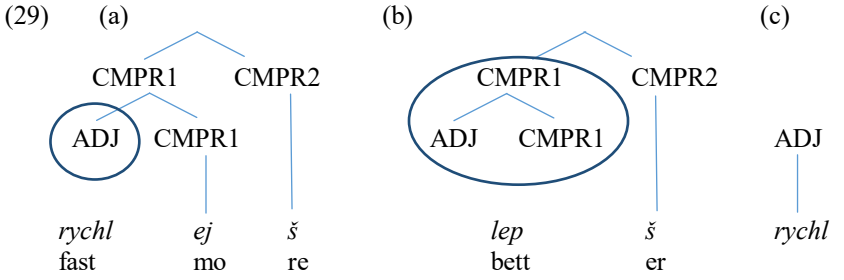
Under this approach, any adjective in English that combines with *-er* must have an entry like *bett-*. If that adjective is suppletive, it means that there is an entry like (26a) which only targets the ADJ node. If such an adjective turns out not to be suppletive, it just means that for the ADJ node, there is no dedicated competitor, and that such a root is able to be inserted also under the simple ADJ node. The precise mechanisms for such "shrinking" has been elaborated within the Nanosyntax framework (Starke 2009; Caha 2009) and I refer the interested reader to this literature.[4]

The idea that there is only a single route to suppletion leads to an explanation for the CSG. The starting point is the fact that we get a suppletive root pair only when we have two entries for a root. The positive-degree entry spells out ADJ—see *good* in (28)—and the comparative degree root spells out either one (*bett*) or both (*worse*) comparative heads. In such a scenario, it is impossible to have a suppletive root for the comparative and simultaneously leave *both* comparative heads intact and empty for insertion. And since bi-partite comparative markers may arise only if both heads are in fact available for insertion, we derive the fact that a bi-partite comparative is incompatible with suppletion, the content of CSG.

4 A reviewer asks about adjectives such as *worse*. In principle (as the reviewer correctly notes), such adjectives could also lack a competitor for the root position, in which case *worse* would be ambiguous between the comparative and the positive degree. For English, one may be tempted to rule this out, but in a larger perspective, it is not clear that this is a malign consequence. That is because the most common strategy for marking comparatives in various languages is to leave the adjective unchanged (Bobaljik 2012, ch. 1.4); so In Japanese, one says literally "John is smart from Bill." Bobaljik suggests that minimally in some languages, CMPR is present, but phonologically null. The question of how to treat this zero marker awaits future research.

Let me now turn to Czech and show how exactly the theory derives the forms and the CSG. In (29a), we see the comparative of an adjective that has a bi-morphemic comparative. The root is inserted under ADJ, leaving the two comparative nodes available for insertion. By necessity, the positive degree will have the same root in this case, because the positive corresponds exactly to the node that the root occupies in the comparative, see (29c).

(29)



On the other hand, in suppletive forms, the root must be different from the one found in the positive degree, and must therefore spell out minimally CMPR1 (as the root *lep* "bett"). If that is so, then it is no longer the case that both CMPR1 and CMPR2 are available for insertion in comparatives, see (29b); the one closer to the root disappears.

The idea that suppletion arises simply due to the spell out of CMPR1 explains also what happens in the comparative adverbs. Recall that the comparative adverb lacks the CMPR2 -*š*, and it is based on the shape of CMPR1. For the suppletive cases, this entails that we will only see the suppletive root followed by the adverbial marker, but with no -*š*. This prediction is borne out, see (23): the form is *lép-e* "bett-ly." This seems to confirm the idea that suppletion does not arise as a consequence of a contact between the root and the -*š*; rather, suppletion is connected to the non-terminal spell out of CMPR1 by the root.

6. Conclusions

The goal of this paper was to discuss one of the generalizations proposed in Jonathan Bobaljik's recent book, namely the RSG. The RSG says that suppletion is restricted by wordhood: comparatives expressed word externally may not condition suppletion. My goal was to suggest that the empirical evidence in favor of such a condition is weaker than initially thought, because agreement markers represent a confounding factor that needs to be controlled for. Further, I suggested a way in which some of the residual cases may be reinterpreted, arguing that the relevant dividing line runs between mono-morphemic and bi-morphemic forms. Whether this reinterpretation can be maintained in the face of the complete record of the phenomenon remains, however, an open question.

Funding Acknowledgement

This contribution is funded by the grant no. GA17-10144S (Exploring Contiguity) issued by the Czech Science Foundation.

Works Cited

Bobaljik, Jonathan. 2012. *The Universals of Comparative Morphology.* Cambridge, MA: MIT Press.

Bobaljik, Jonathan, and Heidi Harley. Forthcoming. "Suppletion is Local: Evidence from Hiaki." In *The Structure of Words at the Interfaces*, edited by Heather Newell, Maíre Noonan, Glynne Piggot, and Lisa Travis. Oxford: Oxford University Press.

Caha, Pavel. 2009. "The Nanosyntax of Case." PhD diss., University of Tromsø.

Corver, Norbert. 1997. "Much-Support as a Last Resort." *Linguitic Inquiry* 28: 119–64.

De Clercq, Karen, and Guido Vanden Wyngaerd. 2016. *Negative Adjectives: Evidence from Czech.* A paper presented at the workshop CRISSP 10, Brussels, Belgium, April 1.

Julien, Marit. 2002. *Syntactic Heads and Word Formation*. Oxford: Oxford University Press.

Koopman, Hilda. 2005. Korean (and Japanese) Morphology from a Syntactic Perspective. *Linguistic Inquiry* 36: 601–33.

Siegel, Dorothy. 1978. "The Adjacency Constraint and the Theory of Morphology." In *Proceedings of NELS 8*, edited by M. Stein, 189–97. Amherst: GLSA.

Starke, Michal. 2009. "Nanosyntax. A Short Primer to a New Approach to Language." In *Nordlyd 36: Special Issue on Nanosyntax*, edited by Peter Svenonius, Gillian Ramchand, Michal Starke, and Tarald Taraldsen, 1–6. Tromsø: University of Tromsø.

Williams, Edwin. 2007. "Dumping Lexicalism." In *Oxford Handbook of Linguistic Interfaces*, edited by Gillian Ramchand and Charles Reiss, 353–82. Oxford: Oxford University Press.

Right Branching in Hungarian: Moving Remnants

Gábor Alberti[a] and Judit Farkas[b]

[a]Department of Linguistics, University of Pécs, Hungary
[b]Research Institute for Linguistics, Hungarian Academy of Sciences

[a]alberti.gabor@pte.hu; [b]juttasusi@gmail.com

Abstract: The wide-spread opinion that Hungarian is basically a head-final language is claimed to rest upon misconception. Different head types can have arbitrarily complex right-branching zones. What actually holds for Hungarian is that such right-branching zones should almost always be extracted. The paper overviews several instances of the scenario of an extracted right-branching domain in Hungarian, starting with the case of complex aspectualizing arguments to be raised into the (right-branching refusing) specifier of Aspectual Projections. Then we show how to raise highly complex noun phrases—deverbal nominal constructions with arguments, for instance—into the specifier of a (right-branching refusing) focus layer. The same type of remnant movement can also happen as an option if a contrastive topic layer is targeted. A section is also devoted to a special *indeed*-construction.

Keywords: right- and left-branching phrases; operator layers; remnant movement

1. Introduction

According to a wide-spread opinion, "Hungarian is a more or less regular head-final language below the level of the (tensed) sentence, that is, in its NPs, APs, PPs, etc." (Kenesei 2014, 225). The source of this opinion is Szabolcsi and Laczkó's (1992, 189–90) argumentation against the mere existence of the postnominal complement zone in Hungarian noun phrases on the basis of a constituency test resting upon *focus* constructions. Alberti et al. (2015), however, points out the inadequacy of the focus test as a constituency test on the basis of the property of the Hungarian focus that it cannot host right-branching phrases by any means, and it proposes a contrastive-topic-based

constituency test, exploiting the fact that the specifier of this layer readily tolerates right branching. In the light of this, Hungarian already proves to be not (or only "statistically") head-final.[1]

This paper is about the other side of the coin. There are (indeed) several syntactic positions in Hungarian which do not tolerate right branching (making it seem as if Hungarian were a head-final language). Nevertheless, even such a position, marked as (Spec,αP) in Figure 1 below, can be applied to host a right branching constituent, βP, at the cost of extracting the right-branching part, γP, in order to provide βP with the pragmaticosemantic contribution peculiar to the operator hosted in α. As shown in Figure 1, we should make the relevant syntactic scenario more precise with at least two respects. The raised phrase βP can be regarded as right branching not only relative to its lexical head β_1 but also relative to (some of) its functional heads; β_2 refers to the head whose complement is extracted. The coincidence between β_2 with β_1 is obviously not excluded. The constituent εP is the one which hosts the extracted γP.

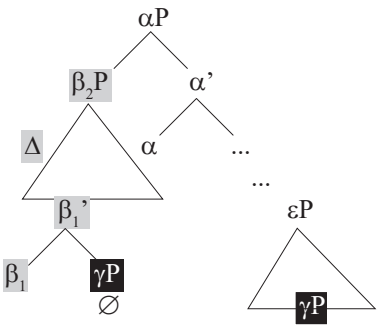


Figure 1. Extracting the right-branching domain (γP) of a complex projection (βP) raised into the specifier of an operator projection (αP)

It is at this point that some general remarks on the (unfavorably highly model-specific) decision of the syntactic category of εP are worth making, in order for us to be able to concentrate on the more relevant nodes αP and βP in Sections 2–5.

In earlier syntactic models of the Hungarian sentence such as that of É. Kiss (2002, 61, 120), in which no morphosyntactic positions (i.e., abstract agreement layers) are considered and the VP is assumed to be flat (but a rich system of topic, quantifier, focus

1 The following abbreviations are used in the glosses:
(i) case suffixes: ACC(USATIVE), ABL(ATIVE), DAT(IVE), DEL(ATIVE), SUB(LATIVE), SUP(ERESSIVE);
(ii) other suffixes on nouns: PL(URAL), POSS (possessedness suffix);
(iii) affixes on verbs: 1SG/.../3PL (agreement suffixes);
(iv) derivational suffixes: INF(INITIVE).

and aspectual layers is applied to account for the different word-order permutations and intonational variants occurring in Hungarian), it is straightforward to assume εP on the right periphery to be identical to this VP itself, with γP occupying an extra position as a sister of V (i.e., as a daughter of V'), at least *as a default*.

If a fully hierarchized (bifurcating) Grohmann-style (2003) model with three Prolific Domains is assumed, the default solution to the problem of εP is as follows. (i) As there is *ab ovo* no reason to assume that the extracted component γP as such gets a (new) thematic role or operator function, εP is not assumed to belong to the thematic domain (Θ) or the operator domain (Ω).[2] (ii) As the order of postverbal arguments depends on their phonetic weights rather than their thematic roles (see É. Kiss [2009] on the role of the Behaghel Law in Hungarian), the Φ domain must be made responsible for the order

2 The thematic domain (Θ) of a hierarchized Hungarian clause structure is analysed in Surányi's (2009, 234, 237, 238) sophisticated model as follows. Besides the customary VP layer ("containing oblique, goal and theme arguments, as well as internal stative locatives") and *v*P layer ("hosting the external argument subjects, and probably also dominating source and orientation of trajectory adverbials"), we need a position for preverbs and other verbal modifiers "below the base position of those elements that cannot 'incorporate' [into the verb] and above the base position of those that can." The given layer can be termed PredP, because the (phrasal) verbal modifier and the verb form a *complex predicate*. Sentences (i)–(ii) can serve as a sketchy illustration of this syntactic model. The base positions from bottom up are as follows: the accusative case-marked theme is base-generated in the VP layer, then the adverbial phrase (*rá*) *a megcímzett képeslapra* "(onto) the addressed postcard.SUB" is in (Spec,PredP), and the nominative case-marked subject is hosted in (Spec,*v*P). Another relevant point of the structure is the analysis of the phonological unit containing the verb (with one **stressed** syllable, which is the first syllable, as always in Hungarian). Surányi (2009, 226–229) accounts for this unit by assuming the following specifier–head configuration: the surface position of the verb is the T(ense) head, and the verbal modifier, which is the accusative case-marked bare noun phrase in (i) and an adverbial phrase in (ii), occupies (Spec,T). At this point, however, we prefer the somewhat different earlier solutions (Piñón 1995; Alberti 2004), based on the assumption that such verbal modifiers are in (Spec,AspP); thus we assume a separate aspectual layer over TP (in Section 2).

(i) **I**li **vé**gül **bé**lyeget ragasztott a **meg**címzett **ké**peslapra.
Ili finally stamp-ACC glued the addressed postcard-SUB
"Finally, Ili put a stamp (or more stamps) on the addressed postcard."

(ii) **I**li **vé**gül **rá**-ragasztott egy **bé**lyeget a **meg**címzett **ké**peslapra.
Ili finally onto-glued a stamp-ACC the addressed postcard-SUB
"Finally, Ili put a stamp on the addressed postcard."

of postverbal dependents; and γP should be placed somewhere, depending on its position in the given word order, in the specifier of a special extracted-component-hosting εP layer within this domain, containing layers such as Asp(ectual)P, T(ense)P, Agr_SP, and Agr_OP.

All in all, in what follows εP will be discussed only in cases in which the extracted component γP should be assumed to have any operator function.

The operator domain is assumed to contain layers headed by the five operators listed in Table 1. They can be identified on the basis of the system of the five types of pragmaticosemantic content given in the table as follows. If the reference r of a noun phrase is associated with a particular operator character in an utterance, by referring to r a whole set of its pragmatic alternatives is evoked as background knowledge shared by the interlocutors. Such alternatives are thus not referred to explicitly, but only implicitly. Due to the given operator, some logical claim is predicated of the implicit referents.

	✓	¬
∃	Q: *also*-quantifier	CTop: contrastive topic
∀	Q: *each*-quantifier	Foc: (contrastive) focus
	TopP: (non-contrastive) topic	

Table 1. The system of operators in Hungarian

In all five examples shown in Table 2, the set of implicit referents consists of persons who can be regarded in a given context as alternatives to a person who is called Lilla. They all together form the *relevant set*. Suppose the implicit participants are Anna, Bea and Cili; so the relevant set now consists of four persons. The corresponding sentence with an *also*-quantifier then provides the additional semantic information—in addition to the "explicit content" that Lilla came here, which is true in all the five variants—that what holds for Lilla also holds for (✓) at least one (∃) implicit participant. The additional information due to the contrastive topic is that what holds for Lilla does not hold for (¬) at least one (∃) implicit participant. The contribution of focus is captured in the table as follows: what holds for Lilla is a piece of information that uniformly (∀) does not hold for (¬) the implicit participants. The *each*-quantifier realizes the fourth logical possibility in the following sense: everyone is referred to implicitly (since the general expression *mindenki* "everyone" can have no other function in the given context than evoking what is termed above the relevant set), and hence the corresponding sentence can be interpreted as claiming that the information "someone came here" holds truly (✓) for each implicit participant (∀). As for the fifth operator, the non-contrastive topic, it can be placed in the system just sketched as an operator realizing the logical alternative of providing no information on the implicit participants. The translations illustrate these (context-based) semantic contributions.

	✓	¬
∃	**Lil**la is **el-** jött. Lilla also away came "Lilla also came here."	^**Lil**la # **el-** jött. Lilla away came "As for Lilla, she came here; but there is another person who did not come here."
∀	**Min**denki **el-** jött. everyone away came "Everyone came here."	**LIL**la jött el. Lilla came away "It was Lilla who came here."
	Lilla **el-** jött. Lilla away came "Lilla came here."	

Table 2. Illustration of the system of operators in Hungarian

As for the formal cues of these operators, relative to the basic variant with a topic, the two types of quantifier can be recognized by means of characteristic elements such as *is* "also" and the morpheme *mind-* "each." The two contrastive operators can be recognized on the basis of peculiar intonational and word-order phenomena. The contrastively topicalized element bears a special rising and then falling intonation (^) and is followed by a short pause (#). The focused element bears a strong **FOCUS STRESS** and seems to substitute for the preverb compared to the neutral word order. Note that we follow Brody (1990) in analyzing the placement of the verb in a focus construction in terms of head movement of the verb to the head of a Foc(us) functional projection that hosts the operator in its specifier (similar to Puskás [2000], but in contrast to É. Kiss [2002], for instance[3]).

We are going to overview several instances of the scenario of an extracted right-branching domain in Hungarian, sketched in Figure 1, starting with a case discussed in Alberti (2004): the case of complex aspectualizing arguments (βP) to be raised into the (right-branching refusing) specifier of Aspectual Projections (αP=AspP). Section 2 evokes this topic. Section 3 shows how to raise highly complex noun phrases (βP)—deverbal nominal constructions with arguments, for instance—into the specifier of a (right-branching refusing) focus layer (αP=FocP). It will also be shown that the same type of remnant movement can also happen as an option if the targeted layer αP is CTopP. Section 4 works up a special *indeed*-construction, which is based on the raising of (remnants of) different kinds of entire clauses into the specifier of an *also*-quantifier. Section 5 is a short summary.

3 It is however more economical to assume a system in which the V-to-F head movement is dispensed with, certain word-order variants (e.g., ex. [22] in É. Kiss [2002, 86]) cannot be convincingly accounted for in the simpler syntactic model.

2. Aspectualizing Arguments: Climbing Preverbs and Roll-Up Structures

On the basis of Alberti (2004), we consider the topic of expressing aspect in Hungarian an ideal point of departure for this paper, devoted to the overview of constructions undergoing constraints on (right) branching.

Aspect is claimed to be often expressed in Hungarian by raising a typically (right) branching argument selected by the verb to serve as its *aspectualizer* into a position, (Spec,AspP), left-adjacent to the surface position of the verb stem, in which (right) branching is not tolerated, unless the given aspect (in the context of the given, "self-aspectualizing," verb) is expressed exactly by raising nothing into (Spec,AspP). It is to this tension between the opposite requirements that Alberti (2004) attributes the Hungarian-specific *climbing-preverb* ([1b], [3b], [3c]) and *roll-up* (4a–b) structures (see É. Kiss and van Riemsdijk [2004] and the rich underlying literature therein).

The difference between the progressive aspect in (1a) and the perfect aspect in (1b) is expressed by the difference that in (1a) the verb stem functions as a self-aspectualizer in the above sense while in (1b) the argument with the structure [$_{\text{AdvP}}$ *up* + sublative case-marked DP] is raised into (Spec,AspP) at the cost of divorcing from its right branching component, the sublative case-marked DP (see footnote 2). The argument for considering the sequence *fel a fára* to be a constituent is that it can serve as a possible short answer to such questions as "Where are you climbing?"

(1) (a) (Éppen) mászom fel a fára.
just climb-1SG up the tree-SUB
"I am climbing up the tree."

(b) Fel-mászom a fára.
up-climb-1SG the tree-SUB
"I (will) climb up the tree."

Table 3 below provides the relevant syntactic details on the basis of Figure 1 in Section 1.[4]

	α	β_1	β_2	д	γP	remark
(1b)	Asp	Adv	β_1	Adv	KP_{Sub}	up climb the tree-SUB

Table 3. The remnant of a right-branching aspectualizing argument in (Spec,$Asp_{(Inf)}$P)

4 Δ in Tables 3–7 and in Figure 1 demonstrates the linearized content of the phonetic material of the remnant of βP in (Spec,αP) "after" extracting what counts as right branching in βP. Δ does not necessarily form a phrase. As will be discussed in connection with (5b), if Δ happens to form a phrase XP, ambiguity may arise, since on an "immediate" reading XP itself is interpreted according to the operator character due to αP while on another reading it is the whole βP expression that should be interpreted in this way.

In (2a–b), let us consider the infinitival phrase(s) with the lexical head "to climb" (from now on, see Table 4 at the end of the section). It is illustrated that the event which is hated can be viewed both progressively (2a) and with perfect aspect (2b). The syntactic difference also concerns the lative expression *fel a fára* "up to the tree," which remains *in situ* in (2a), while in (2b), it is raised into (Spec,Asp_{Inf}P) at the cost of divorcing from its right branching component, the sublative case-marked DP (NB: the difference has also shown that infinitival expressions have their own internal aspect).

(2) (a) Utálok éppen mászni fel a fára, . . .
hate-1SG just climb-INF up the tree-SUB
"I hate to be in the middle of climbing up the tree . . ."

(b) Utálok fel-mászni a fára.
hate-1SG up-climb-INF the tree-SUB
"I hate to climb up the tree."

The minimal pair of examples in (3a–b) should be compared to the pair in (2). The source of the difference in word order is the difference between the finite verbs "hate" and "want." While "hate" behaves as a self-aspectualizing verb, which "blocks" the filling of its (Spec,AspP) left-adjacent to it (2a–b), "want" "uses" the infinitival phrase (which is its argument referring to the object of demand) as its aspectualizer (3a–b). If the InfP is progressive (3a), its syntactic structure can be regarded as right branching relative to its Inf head, so this Inf head will constitute the remnant appearing in (Spec,AspP). However, if the InfP is perfective (3b), its syntactic form starts with the adverbial head "up" in (Spec,Asp_{Inf}P), and now the InfP will qualify as the right branching part. Table 3 above provides the relevant syntactic details on the basis of Figure 1 in Section 1.

(3) (a) Mászni akarok éppen fel a fára, . . .
climb-INF want-1SG just up the tree-SUB
"I want to be in the middle of climbing up the tree . . ."

(b) Fel akarok mászni a fára.
up want-1SG climb-INF the tree-SUB
"I want to climb up the tree."

(c) Fel fogok akarni mászni a fára.
up will-1SG want-INF want-1SG the tree-SUB
"I will want to climb up the tree."

The word order in the *climbing preverb* construction in (3c) can be derived by the triple successive cyclic application of the raising rule aiming at the corresponding (Spec,Asp$_{(Inf)}$P) positions, all refusing right branching. Rows (3c.1–3) in Table 4 provide the relevant details.

The pair of examples in (4a–b), potential short answers to questions like "what was the most remarkable mistake?", shows another clustering of verbs, dubbed *role-up* structures. As indicated in the two β rows in Table 4, the syntactic difference between the analogous examples (3b–c) and (4a–b) obviously depends on the selection of the head of the aspectualizing expression relative to which right branching is considered. In (4a–b), but not in (3b–c), left branching is tolerated in the relevant (Spec,Asp$_{(Inf)}$P) positions (while right branching is not tolerated in either cases). Thus, in (4a–b) right branching is calculated relative to the infinitival lexical head while in (3b–c) relative to a higher functional head. Alberti (2004) attributes this difference to the difference that in (3b–c) the ultimate aim is to fill in the specifier of the aspectual layer belonging to the finite verb while in (4a–b) the relevant "ultimate" aspectual layer (see rows [4a.2] and [4b.3] in Table 4) belongs to an infinitival head. Entering into the presumable phonetic background of the phenomenon would go beyond the scope of this paper, but see Alberti (2004).[5]

(4) (a) fel-mászni próbálni egy ilyen fára
up-climb-INF try-INF a such tree-SUB
"to attempt to climb up such a tree"

(b) fel-mászni próbálni akarni egy ilyen fára
up-climb-INF try-INF want-INF a such tree-SUB
"to want to attempt to climb up such a tree"

It is worth testing, however, whether an infinitival expression with another type of functional head on its left periphery can serve as an aspectualizer in the (Spec,AspP) position belonging to the finite verb "want." A focused infinitival construction is tested in (5a), as the comparison between the intended meaning given in (5a) and the non-intended one in (5c), in which the focus semantically belongs to the finite verb "want" instead of the infinitive "to climb," clearly shows.

(5) (a) *HÉTfőn fel-mászni a fára akartam.
Monday-SUP up-climb-INF the tree-SUB wanted-1SG
Intended meaning: "To climb up the tree exactly ON MONDAY, that is what I wanted."

5 Also see the comments on example (30) in Alberti et al. (2015, 34), which offers a global analysis on differently "heavy" phrases depending on left/right branching and the positions available for them.

(b) HÉTfőn akartam fel-mászni a fára.
Monday-SUP wanted-1SG up-climb-INF the tree-SUB
"To climb up the tree exactly ON MONDAY, that is what I wanted."

(c) HÉTfőn akartam fel-mászni a fára.
Monday-SUP wanted-1SG up-climb-INF the tree-SUB
"To climb up the tree, that is what I used to want ON MONDAY."

The word order in (5a) is ill-formed. But what is to do is nothing else but to extract the component that counts as right branching relative to the highest functional head in the expression with the infinitive as its lexical head, which is the Foc_{Inf} head, responsible for the focus interpretation within the infinitival expression. The resulting, well-formed, word order is shown in (5b). It coincides with the word order presented in (5c). The source of the ambiguity is obviously the two possible affiliation of the focused temporal expression, of which the structurally more complicated variant (5b) offers the more natural reading.

	α	β_1	β_2	д	γP	remark
(1b)	Asp	Adv	β_1	Adv	KP_{Sub}	up climb the tree-SUB
(2b)	Asp_{Inf}	Adv	β_1	Adv	KP_{Sub}	up climb-INF the tree-SUB
(3a)	Asp	Inf	Asp_{Inf}	Inf	AdvP	climb-INF want up the tree-SUB
(3b.1)	Asp_{Inf}	Adv	β_1	Adv	KP_{Sub}	up climb-INF the tree-SUB
(3b.2)	Asp	Inf	Asp_{Inf}	Adv	InfP	up want climb-INF the tree-SUB
(3c.1)	Asp_{Inf}	Adv	β_1	Adv	KP_{Sub}	up climb-INF the tree-SUB
(3c.2)	Asp_{Inf}	Inf	Asp_{Inf}	Adv	InfP	up want-INF climb-INF the...
(3c.3)	Asp	Inf	Asp_{Inf}	Adv	InfP	up will want-INF climb-INF the...
(4a.1)	Asp_{Inf}	Adv	β_1	Adv	KP_{Sub}	up climb-INF a s. tree-SUB
(4a.2)	Asp_{Inf}	Inf	β_1	Adv+Inf	KP_{Sub}	up climb-INF try-INF a s. tree-SUB
(4b.3)	Asp_{Inf}	Inf	β_1	Adv+ +Inf+Inf	KP_{Sub}	up climb-INF try-INF want-INF a s. tree-SUB
(5b)	Asp	Inf	Foc_{Inf}	N_{Sup}	Asp_{Inf} P	on-M want climb-INF a tree-SUB

Table 4. Summary: remnants of right-branching aspectualizing arguments in (Spec,$Asp_{(Inf)}$P)

3. Complex Noun Phrases in Specifiers of Operator Projections

It is investigated in (6) how we can focus a noun phrase which is so complex that, relative to the lexical noun head, it has both right- and left-branching parts—by raising its appropriate remnant into (Spec,FocP) (introduced in Table 1 in Section 1).

(6) (a) *[Móricznak a versikéjét a tehenekről] mondja el.
Móricz-DAT the rhyme-POSS.ACC the cow-PL.DEL recites away

(b) **MÓ**ricznak a **VER**sikéjét mondja el a **TE**henekről.
Móricz-DAT the rhyme-POSS.ACC recites away the cow-PL.DEL

(c) **MÓ**ricznak mondja el a **VER**sikéjét a **TE**henekről.
Móricz-DAT receites away the rhyme-POSS.ACC the cow-PL.DEL
"It is Móricz's rhyme about the cows that he is going to recite (of several literary works)." (6a–c)

(Spec,FocP) does not tolerate right branching (from the lexical head) (6a), but it tolerates left branching (6b), at least as an option in addition to another option according to which the remnant in (Spec,FocP) only consists of the dative case-marked possessor on the left periphery of the nominal expression (6c). Table 5 provides the relevant technical details. What is crucial in this section is that the dative case-marked possessor is assumed (Alberti et al. 2015) to be hosted in a separate PosP layer built upon the DP layer on the left periphery of the noun phrase, on which even ω_{Pos}P operator layers can be based in the spirit of the clausal-DP hypothesis (Grohmann [2003, 200]; see [9b–c]), which "argues that essentially all types of properties found in the clause can also be found in the nominal layer."

	α	β_1	β_2	д	γP	remark
(6b)	Foc	N	β_1	N D N	KP_{Del}	εP: FocP
(6c)	Foc	N	Pos	N_{Dat}	$[_{DP}$ D N_{Acc} $KP_{Del}]$	εP: FocP
(7a)	Foc	N	Pos	N_{Dat}	$[_{DP}$ D N_{Acc} $KP_{Del}]$	εP: default
(7b)	Foc	N	β_1	N D N	KP_{Del}	εP: default
(8a)	Foc	N	β_1	Det A N	DP_{Dat}	εP: FocP
(8b)	Foc	N	β_1	Det A N	DP_{Abl}	εP: FocP
(9b)	Foc	N	Q_{Pos}	Det N_{Dat}	$[_{PosP}$ D $N_{Acc}]$	
(9c)	CTop	N	Q_{Pos}	Det N_{Dat}	$[_{PosP}$ D $N_{Acc}]$	

Table 5. Remnants of focused and contrastively topicalized nominal expressions functioning as operators

As illustrated in (7a), the word order presented in (6c) can be associated with two meanings, which is a newer instance of the systematic-ambiguity phenomenon

discussed in footnote 4. In variant (7a), only the possessor is focused, that is, it is presupposed that rhymes about cows of different poets are recited, which is a presupposition much more specific than in the case of variant (6c). This difference in meaning comes with the difference in stress pattern that in (6c), in contrast to (7a), even the nominal head *versikéjét* "rhyme-POSS.ACC" and the delative case-marked noun *tehenekről* "cow-PL.DEL" are focus-stressed, besides the possessor. This can be accounted for by assuming that in the syntactic structure of (6c) the extracted part (γP) is also hosted in the FocP layer (εP), in some way or another. It is a theory-dependent question whether we follow É. Kiss (1992, 99–104) in assuming a *mirror focus* construction with a right branching (Spec,FP) position (besides the customary left branching [Spec,FP]) or Alberti–Medve (2000, 95–105) in assuming an extra position dominated by F'. Note that assuming, as is suggested by É. Kiss (2002, 99) in a similar context, that γP is hosted among arguments *in situ* on the right periphery in the analysis of (6c) is an approach in which the intonational and semantic differences between (7a) and (6c) are not accounted for.

(7) (a) **MÓ**ricznak mondja el a versikéjét a tehenekről.
Móricz-DAT recites away the rhyme-POSS.ACC the cow-PL.DEL

(b) **MÓ**ricznak a versikéjét mondja el a tehenekről.
Móricz-DAT the rhyme-POSS.ACC recites away the cow-PL.DEL
"It is Móricz whose nursery rhyme about the cows he is going to recite (of several rhymes about cows)." (7a–b)

It is presented in (7b) that the word order in (6b) can also be associated with the meaning associated with (7a). The interesting experience is that the meaning (practically the ratio of the presupposition within the meaning) depends on which words are focus-stressed, independently of what is extracted and what remains in the remnant in (Spec,FocP). The latter factor is ruled by branching questions, rather than semantic ones.

For the sake of completeness, let us consider examples with alternative argument structures. In (8a), the writer of the rhyme is expressed, again, as a dative case-marked possessor, but situated in the complement of the complex noun phrase (see Alberti et al. [2015, 18–33]). In (8b), the writer is referred to by an ablative case-marked nominal expression, also situated there. As clearly shown by the well-formed word orders, what only counts with respect to the raising of the complex noun phrase (βP) is also branching (see also the corresponding rows in Table 5).

(8) (a) Egy **TR**éfás **VER**sikéjét mondja el **MÓ**ricznak.
a funny rhyme-POSS.ACC recites away Móricz-DAT
"He is going to recite a funny nursery rhyme of Móricz."

(b) Egy **TR**éfás **VER**sikét mond el **MÓ**ricztól.
a funny rhyme-ACC recites away Móricz-ABL
"He is going to recite a funny nursery rhyme by Móricz."

As mentioned above, the Hungarian noun phrase structure is "clausal" in that it can contain operator layers (Farkas and Alberti 2017). In (9a), the dative case-marked possessor referring to "both colleagues" is assumed to occupy the specifier of a $Q_{Pos}P$ layer over the PosP layer on the left periphery of the complex accusative case-marked noun phrase. As this noun phrase is not right branching relative to its N head "sending," it can remain as a non-split unit (9a). However, we can also have recourse to the option of extracting the part which can be regarded as right branching relative to the Q_{Pos} functional head, either the matrix (Spec,αP) belongs to a focus (9b) or a contrastive topic (9c) construction (see Table 1).

(9) (a) Mindkét kollégának az elküldését ellenzi.
both colleague-DAT the sending- POSS.ACC opposes
1. "He is against the option according to which both colleagues would be sent away [as for him, one of them can be sent away]."
2. "It holds for both colleagues that he is against the option according to which the given colleague would be sent away [he thinks that neither of them should be sent away]."

(b) Csak mindkét kollégának ellenzi az elküldését.
only both colleague-DAT opposes the sending- POSS.ACC
"It is only the option according to which both colleagues would be sent away that he is definitely against [as for him, one of them can be sent away]."

(c) ^Mindkét kollégának # **el**lenzi az elküldését.
both colleague-DAT opposes the sending-POSS.ACC
"As for the option according to which both colleagues would be sent away, he is definitely against that [but there are options that he is not against]."

This option is another instance of a surprisingly large "distance" between the semantic content expressed and the word order, ruled by branching factors, since the extracted possessor as a quantifier belongs to the deverbal nominal head (as in meaning [9a.1]) in both cases (9b–c), instead of belonging to the finite verb (as in meaning [9a.2]). We will discuss the semantic details in a series of other papers. The only thing relevant here is that whether the possessor "ran away from home" (Szabolcsi 1983), as in (9b–c), or not, as in (9a), has no impact on the option that an *each*-quantifier possessor in a deverbal

nominal construction can be interpreted in the information structure of the verb which is the derivational basis of the given construction, as in the case of the meanings shown in (9a.1), (9b) and (9c).

4. A Special *Indeed*-Construction

This section discusses a special construction in which the matrix αP in Figure 1 is chosen to be the *also*-quantifier layer. It can be used only as a continuation of a text in which what is claimed in the given sentence to take or have taken place has been "promised" as a plan or a prediction; see the translations associated with (10a–b), for instance. Thus, the particle *is* "also" refers to a plan as the presupposition (see Table 1 in Section 1 concerning the logical interpretation of this operator) underlying the fact that the shaded string of words refers to in (10a–d).

(10) (a) És **HÉT**főn is másztam fel a fára!
and Monday-SUP also climbed-1SG up the tree-SUB
"And it was on Monday, indeed, that I climbed up the tree";
as a continuation of (5b) in Section 2.

(b) És fel is mászom a fára!
and up also climb-1SG the tree-SUB
"And I WILL climb up the tree," as a continuation of (3b) in Section 2.

(c) Havazott is!
snowed also
"It was snowing, indeed."

(d) És nem is **ÉN** mentem el!
and not also I went-1SG away
"And it was not me, indeed, who went away."

The shaded strings of words constitute a FocP, an AspP, a VP, and a NegP, respectively. The *is* particle is inserted in the strings immediately after these highest operator layers (β_2P), triggering the extraction of the phonetic material in their complement (γP), at least according to our approach sketched in Section 1 and exemplified in Sections 2–3. More precisely, (10c) is an exception, the "degenerate case" of the special *indeed*-construction, in which there is no operator layer and no right branching, and hence the verb itself is raised into (Spec,QP) and nothing is extracted.

	α	β_1	β_2	д	γP
(10a)	Q	V	Foc	N_{Sup}	AspP
(10b)			AspP	Adv	*V*p
(10c)			β_1	V	–
(10d)			Neg	*nem*	FocP

Table 6. Remnants of different types of finite construction in (Spec,Q_{also}P)

As presented in (11a), a Q_{each}P construction cannot be raised into (Spec,Q_{also}P), probably due to some kind of incompatibility between the two types of quantifiers (see also Table 7).

(11) (a) *És mindenki is elment.
and everyone also away went

(b) És mindenki el is ment.
and everyone away also went
"And, indeed, everyone went away." (11a–b)

Nevertheless, the intended meaning in (11a) can be expressed as follows: a smaller part of the Q_{each}P construction should be extracted, namely, the right branching complement of the aspectual head occupied by the preverb *el* "away" (11b).

It is even more dispreferred for a Q_{also}P construction to be raised into (Spec,Q_{also}P), obviously to the total incompatibility (12a).

(12) (a) *És Lilla is (is) el (is) ment.
and Lilla also also away also went

(b) És el is ment Lilla is.
and away also went Lilla also
"And, indeed, Lilla also went away." (12a–b).

	α	β_1	β_2	д	γP	remark
(11a)	Q	V	Q	N	AspP	unacceptable
(11b)			AspP	N Adv	*v*P	remnant of βP=QP is raised
(12b)			AspP	Adv	*v*P	*is*-quantifier: postverbal

Table 7. Remnants of further types of finite construction in (Spec,Q_{also}P)

Now it is the fact that quantifiers are allowed to remain *in situ* in the postverbal periphery (É. Kiss 2002, 119–22) that offers a solution (12b).

5. Concluding Remarks

We claim that the structures proposed in Sections 2–4 on the basis of the general scheme presented in Section 1 precisely account for the complex meanings and special stress patterns that sentences (1a)–(12b) are associated with.[6]

The rich domain of data and the well-functioning analyses have led us to the conclusion that Hungarian is not a head-final language. Different head types can have arbitrarily complex right-branching zones. What actually holds for Hungarian is that such right-branching zones should often be extracted.

We conclude this paper by formulating the conjecture that several further constructions also function according to the scheme in Figure 1, whose verification we have intended to devote two further papers.

Acknowledgement

We are grateful to OTKA NK 100804 (*Comprehensive Grammar Resources: Hungarian*) and OTKA NF-84217 for their financial support.

Works Cited

Alberti, Gábor. 2004. "Climbing for Aspect—with No Rucksack." In *Verb Clusters; A study of Hungarian, German and Dutch*. *Linguistics Today* 69, edited by Katalin É. Kiss and Henk van Riemsdijk, 253–89. Amsterdam: John Benjamins.

Alberti, Gábor, Judit Farkas, and Veronika Szabó. 2015. "Arguments for Arguments in the Complement Zone of the Hungarian Nominal Head." In *Approaches to Hungarian* 14, edited by Katalin É. Kiss, Balázs Surányi, and Éva Dékány, 3–36. Amsterdam: John Benjamins.

Brody, Michael. 1990. "Some Remarks on the Focus Field in Hungarian." *UCL Working Papers in Linguistics* 2: 201–26.

É. Kiss, Katalin. 2002. *The Syntax of Hungarian*. Cambridge: Cambridge University Press.

É. Kiss, Katalin. 2009. "Is Free Postverbal Order in Hungarian a Syntactic or a PF Phenomenon?" In *The Sound Pattern of Syntax*, edited by Nomi Erteschik-Shir and Lisa Rochman, 53–71. Oxford: Oxford University Press.

Farkas, Judit, and Gábor Alberti. 2017. "The Hungarian HATNÉK-noun Expression: A Hybrid Construction." In *Constraints on Structure and Derivation in Syntax, Phonology and Morphology*, edited by Anna Bondaruk and Anna Bloch-Rozmej, 71–100. Frankfurt am Main: Peter Lang.

6 In order to be convinced of the *really* complex meanings, on which typically complex propositions obtain additional operator-based semantic factors instead of the simple entities appearing in the brief introduction associated with Table 1, it is worth reconsider the instances of word-order-level ambiguity discussed in footnote 4.

Grohmann, Kleanthes K. 2003. *Prolific Domains: On the Anti-Locality of Movement Dependencies. Linguistik Aktuell* 66. Amsterdam: John Benjamins.

Kenesei, István. 2014. "On a Multifunctional Derivational Affix: Its Use in Relational Adjectives or Nominal Modification, and Phrasal Affixation in Hungarian." *Word Structure* 7 (2): 214–39.

Piñón, Christopher. 1995. "Around the Progressive in Hungarian." In *Approaches to Hungarian* 5, edited by István Kenesei, 153–90. Szeged: JATEPress.

Puskás, Genoveva. 2000. *Word Order in Hungarian. The Syntax of A-bar Positions.* Amsterdam: John Benjamins.

Surányi Balázs. 2009. "Verbal Particles inside and outside *v*P." *Acta Linguistica Hungarica* 56: 201–49.

Szabolcsi, Anna. 1983. "The Possessor That Ran Away from Home." *The Linguistic Review* 3: 89–102.

Szabolcsi, Anna, and Laczkó, Tibor. 1992. "A főnévi csoport szerkezete." *Strukturális magyar nyelvtan 1. Mondattan,* edited by Ferenc Kiefer, 179–298. Budapest: Akadémiai Kiadó.

Syntactic Features and Their Interpretations

Preverbal Focus and Syntactically Unmarked Focus: A Comparison

Enikő Tóth[a] and Péter Csatár[b]

University of Debrecen, Debrecen, Hungary

[a]toth.eniko@arts.unideb.hu; [b]csatar.peter@arts.unideb.hu

Abstract: This paper presents the results of an experiment exploring the factors influencing the interpretation of Hungarian preverbal (PVF) and syntactically unmarked focus (SUF). We used a sentence-picture verification task where participants rated utterances on a 6 point Likert-scale. There was no empirical difference found between PVF and SUF with respect to the two factors, exhaustivity and expectedness. Exhaustivity had a main effect, while expectedness did not. Our findings are in line with Gerőcs et al.'s (2014) results and provide empirical evidence in favor of Surányi's (2011) claim that SUF can also receive an exhaustive interpretation, at least in a context which strongly supports exhaustivity. In addition, the results contradict those views that treat PVF as necessarily exhaustive and SUF as necessarily non-exhaustive when they form an answer to a *wh*-question. Our study implies that the exhaustivity of PVF is not inherent in nature, and it should be treated as a pragmatic phenomenon.[1]

Keywords: focus; exhaustivity; expectedness; experimental pragmatics

1. Introduction

In Hungarian, two types of focus have traditionally been differentiated: identificational or preverbal focus (PVF), and syntactically unmarked or information focus (SUF) (É. Kiss 1998). PVF is marked by stress, the focused constituent moves into a preverbal position and if the verb contains a verbal particle, then the particle is stranded by the movement of the verb:[2]

1 We would like to thank Miklós Fegyveres for his help in conducting the experiment. We are also grateful to Kálmán Abari, Gábor Alberti, György Rákosi and the anonymous reviewer of this paper for their valuable comments.

2 Boldface signals prosodic prominence throughout the paper.

(1) Mari **egy** **kalap-ot** nézett ki magá-nak.
Mary a hat-ACC picked PRT herself-DAT
"It was **a hat** that Mary picked for herself." (É. Kiss 1998, 249)

SUF, on the other hand, is marked by prosodic prominence only, and the focused constituent remains in situ.

(2) Mari kinézett magá-nak **egy** **kalap-ot**.
Mary picked out herself-DAT a hat-ACC
"Mary picked for herself **a hat**." (É. Kiss 1998, 249)

From a semantic perspective, PVF was first considered to be exhaustive in nature, and its exhaustivity was described as an inherent semantic feature, i.e., its exhaustivity was handled as part of the truth-conditions of the sentence under interpretation (É. Kiss 1998, Szabolcsi 1981). However, recently this view has been challenged from an empirical perspective. For example, Wedgwood (2005) argued—relying on corpus-linguistic data—that the exhaustive interpretation arises as a result of a pragmatic implicature. In addition, several experimental studies also questioned the inherent exhaustivity of PVF (see Onea and Beaver 2011, Kas and Lukács 2013) and claimed that the exhaustivity of PVF should be treated as a pragmatic phenomenon. Regarding SUF, É. Kiss (1998) claimed that its function is to mark new, non-presupposed information. She also pointed out that "if the answer [to a *wh*-question] is exhaustive, [. . .] it must be put as a preverbal identification focus" (É. Kiss 1998, 250), and this implies that SUF cannot express exhaustive identification.

It was Surányi (2011) who first raised the possibility that SUF might also receive an exhaustive interpretation. He also called for experimental investigations, since he relied only on his own intuitions and on a limited spectrum of native speaker judgments that he had collected in a non-systematic way. Taking his view as a starting point, the aim of the present paper is threefold: (i) to collect experimental evidence either in favor of, or against, Surányi's (2011) claim regarding the exhaustivity of SUF, (ii) to collect further data regarding the exhaustivity of PVF within the same experimental setting, and (iii) to compare the results obtained in order to gain a more accurate picture of how these focus structures are interpreted.

2. Background

2.1 Previous Experimental Work on PVF

One of the most fundamental theoretical problems with respect to PVF is how the exhaustive interpretation associated with it can be accounted for. As mentioned above, several experimental studies argued that exhaustivity is not an inherent semantic feature of PVF and it was also suggested that the exhaustive interpretation arises as a pragmatic

implicature (Onea and Beaver 2011, Kas and Lukács 2013).[3] More recent experimental studies also support the pragmatic view regarding the exhaustivity of PVF. For example, Babarczy and Balázs (2016) argue in a relevance-theoretic framework that the exhaustivity of PVF is due to pragmatic inferences, and they consider it to be a scalar implicature. This means that in a non-exhaustive context a PVF construction is under-informative and therefore its interpretation requires more cognitive effort. They assumed that in a sentence-picture verification task, where a PVF construction is accompanied by a matching, but non-exhaustive picture, kindergarten children will produce a non-uniform rating pattern on a ternary-scale, while adults, due to their sensitivity to pragmatic meaning, should be able to opt for the middle point on the scale (neither matching, nor non-matching). They tested 4-, 6- and 8-year-old children and an adult control group and found a correlation between the cognitive maturity of children and the ability to interpret under-informative structures (PVF), i.e., the cognitively more developed children showed more adult-like behavior in the rating task.[4] In other words, Babarczy and Balázs (2016) proved their hypothesis, and concluded that the results support the pragmatic approach to the exhaustivity of PVF.

Pintér (2016) also used a sentence-picture verification task to test the exhaustivity of PVF within four age groups (6-, 7- and 9-year-old children, and an adult control group). Moreover, she also wanted to show that binary judgement tasks are not suitable for testing intuitions about PVF. To prove this claim she used two different experimental designs. In the first experiment participants had to decide whether a sentence with PVF was true or false (binary scale) with respect to a given picture. In the second experiment, however, she also employed a ternary scale (cf. Babarczy and Balázs 2016). Her results confirmed her expectations: she found that while a binary scale is not an appropriate method to discern the results of the different age groups, with the help of a ternary scale it is possible to detect subtle differences across the age groups. Accordingly, 7-year-olds and older children showed adult-like behavior when producing ternary judgements, but preschoolers did not show sensitivity to PVF. Pintér (2016) also found a significant difference across her experimental conditions (true, false, false in an exhaustive reading). Analyzing reaction time in the case of the adult control group there was no significant difference across conditions, i.e., participants did not need more time to reject a sentence in a non-exhaustive setting than to accept it in the true control condition. Pintér (2016) argues that these results suggest that the exhaustivity of PVF should be treated as a presupposition, and not as an inherent semantic feature or an implicature.

3 For a detailed discussion of these experiments see Gerőcs et al. (2014).

4 They used independent standard tests, such as the N-back test and the Dimensional Change Card Sort Task to examine the cognitive abilities of the participants (Babarczy and Balázs 2016, 156–57).

From a cross-linguistic point of view, Zimmermann (2008) introduced an entirely different perspective and suggested that discourse factors, such as discourse expectability, might trigger focus-fronting (a marked construction) in various languages. He emphasizes the importance of the background assumptions of both the speaker and the hearer, and he argues that the less expected a given piece of information is for the hearer as judged by the speaker, the more likely the speaker is to put it into a marked, focus position.[5] However, Zimmermann (2008) points out that Hungarian might be an exception to this generalization.

Skopeteas and Fanselow (2011) also conducted two experiments to investigate whether the use of focus constructions is motivated by discourse-related factors in four languages (German, Spanish, Greek and Hungarian). First, they examined whether the exhaustive interpretation is obligatory or not, since if it is not obligatory then it must be dependent on discourse-related features. Participants filled in a questionnaire, where relevant test items consisted of a *wh*-question and an answer with an object constituent in PVF. After reading the question-answer pair participants had to judge the extent to which the answer was exhaustive while answering a *yes-no* question: *Is it possible that Matthias also fished other fishes?* (Skopeteas and Fanselow 2011, 1695), then they indicated their judgment on a 7 point Likert-scale (7: further alternatives are possible, 1: further alternatives are excluded). A test-item taken from Skopeteas and Fanselow (2011, 1694–95) is shown below:

(3) Q: Mi-t fogott Matyi?
what-ACC caught Matthias
"What did Matthias catch?"

A: **Pisztráng-ot** fogott Matyi.
trout-ACC caught Matthias
"It was **trout** (and not other types of fish) that Matthias caught."[6]

Their findings set Hungarian apart from the other languages they examined, and they observe that in Hungarian the exhaustive interpretation of PVF is obligatory, and it arises independently of the context. However, we would like to point out that their Hungarian test sentences are infelicitous, or at least marked,[7] according to our native speaker intu-

5 Destruel and Velleman (2014) pursue the same line of argumentation when they describe the use of *it*-clefts in English (which is usually associated with PVF in Hungarian) as marking a conflict with the various expectations of the interlocutors.

6 Translation adjusted by the authors.

7 The answer has only one available interpretation in Hungarian, which means that we are comparing different types of fish and not discussing a particular entity that Matthias has caught, as the test items taken from the other languages suggest.

itions (see also Pintér's [2016] comments on their Hungarian data), which means that one should treat their conclusion with caution. There is also a methodological problem regarding their experiment, i.e., participants had to provide a scalar judgement as an answer to a *yes-no* question.

In the second experiment, which used the same design, Skopeteas and Fanselow (2011) examined whether speakers will select a marked construction in order to signal to the hearer that the information in the focus position is not predicted when compared to the background assumptions of the hearer. Again they tested object constituents in PVF, where the entity referred to by the object was either predictable (a trout) or not predictable in the given context (a bottle). As we have seen, in their first experiment they observed that focus-fronting in Hungarian is not dependent on contextual factors. Therefore, they did not anticipate finding an effect of predictability of the focused constituent in Hungarian. The results satisfy their prediction. However, they do not provide a list of the Hungarian test sentences, which might be problematic, since they used infelicitous sentences in the first experiment. Therefore, their overall conclusion—that Hungarian is the only language where exhaustivity is a structural property and the interpretative properties of PVF are not sensitive to contextual factors—should be re-examined.

2.2 Previous Experimental Work on SUF

As mentioned above, Surányi (2011) raised the issue of whether SUF can also be interpreted exhaustively. He analyzes SUF in a question semantic framework and characterizes it as an answer given to a "Mention some!" question. Such questions require at least one relevant answer which is contextually appropriate, out of the semantically possible alternative answers. However, "Mention some!" questions might also receive answers which contain all the alternatives, i.e., answers that are exhaustive in the given context. Based on these assumptions, the exhaustive interpretation of SUF is context-dependent. This means that there is an important difference between the exhaustivity of PVF and SUF: whereas, according to the standard theory, the exhaustivity of PVF is obligatory, SUF is only optionally exhaustive. Hence, the exhaustive interpretation of SUF is context-dependent and, as Surányi (2011) argues, it might arise as a pragmatic implicature.

Adopting Destruel et al.'s (2015) view on answers to questions, we suggest that the same difference between PVF and SUF can be captured from another perspective, too. Destruel et al. (2015) argues that answers can be labelled as maximal if "no true answer to the question under discussion . . . is strictly stronger" (Destruel et al. 2015, 136). Maximal answers are in fact exhaustive, therefore they cannot be followed by another question which seeks information about other entities satisfying the previous question. For instance, in the example below the second question that comes after the exhaustive answer with PVF, is infelicitous, provided that the speaker and the hearer share the background assumption that we usually read aloud one tale each evening to our children.

(4) Q: Mi-t olvastál fel nekik elalvás előtt?
what-ACC read PRT them sleeping before
"What did you read aloud to them before bedtime?"

A: A **Hamupipőké-t** olvastam fel.
the Cinderella-ACC read PRT
"I read aloud **Cinderella**."
És még mit olvastál fel?
"And what else did you read aloud?"

However, since the exhaustivity of SUF is optional, such a continuation is more acceptable in (5):

(5) Q: Mi-t olvastál fel nekik elalvás előtt?
what-ACC read PRT them sleeping before
"What did you read aloud to them before bedtime?"

A: Felolvastam a **Hamupipőké-t**.
read the Cinderella-ACC
"I read aloud **Cinderella**."
És még mit olvastál fel?
"And what else did you read aloud?"
(The question-answer pair is taken from Surányi 2011, 283)

From an empirical perspective, Gerőcs et al.'s (2014) two experiments are pertinent. First, the exhaustivity of PVF and SUF was tested; second, two other types of focus structures (*only*-focus, cleft-constructions) were examined, as well.

The starting point of the first experiment was the relevance-theoretical hypothesis that any decrease in cognitive resources results in limited information processing. Regarding the interpretation of PVF and SUF this means that if the exhaustivity of PVF is semantic in nature, then it will be processed even when the cognitive resources are artificially limited within an experimental setting. If, on the contrary, there is no such limitation present in the experimental setting, then a pragmatic implicature can also be formulated, i.e., not only PVF, but also SUF can be interpreted exhaustively. The limitation on cognitive resources was controlled by manipulating the time window in which participants were required to finish the task.

At the beginning of the task participants listened to a background story (a young girl finds a corpse with a piece of paper in its pocket), which was followed by a *wh*-question. After that, the target sentence was presented auditorily, and at the same time a picture appeared on the screen depicting a non-exhaustive scenario. If the picture matched the

target sentence, participants had to give a *yes* answer. If, on the contrary, the picture did not match the target sentence, they had to provide a *no* answer—which represents the exhaustive interpretation. Participants were divided into two groups based on the length of the time window: in the Long condition participants had 3000 ms to give a *yes/no* answer, while in the Short condition the limit was only 1000 ms. A test item taken from Gerőcs et al. (2014) is presented below (Gerőcs et al. 2014, 186):[8]

(6) Q: Mi-t karikázott be az áldozat?
what-ACC circled PRT the victim
"What had the victim circled?"

A (PVF): Az áldozat a **piramis-t** karikázta be.
the victim the pyramid-ACC circled PRT
"It was the **pyramid** that the victim had circled."

A (SUF): Az áldozat bekarikázta a **piramis-t**.
the victim circled the pyramid-ACC
"The victim circled the **pyramid**."

The accompanying picture showed a crown, a fish and a pyramid, where the crown and the pyramid were each circled.

Gerőcs et al. (2014) found that in the Long condition SUF and PVF sentences were interpreted exhaustively almost in the same proportion (63% vs. 72%). The authors claim that this result may be attributed to the effect of the introductory *wh*-question, which could have served as a trigger for implicature generation.

In the Short condition, however, the results cannot be explained in the same way. The shorter time window resulted in a smaller proportion of exhaustive answers and participants performed around chance level in the case of both PVF and SUF. On the one hand, Gerőcs et al. (2014) argue that the results might be explained by the fact that the limited time was not enough to process the target sentence and participants were only guessing. On the other hand, the results might also be accounted for assuming that the exhaustive interpretation arises as an implicature, in the case of both PVF and SUF. Leaving it out of consideration which explanation is more plausible, it is important to note here that both explications point toward the conclusion that the exhaustivity of PVF is not semantic in nature.

In order to explore this claim further, the authors conducted a second experiment where they compared PVF to other types of focus constructions, such as SUF, *only*-focus and cleft-constructions. The authors expected that PVF, *only*-focus and cleft-constructions

8 Boldface marking prosodic prominence added by the authors.

would reveal a strong preference for an exhaustive interpretation, while SUF would be less likely to be interpreted exhaustively, since they tested isolated sentences only.

Participants were introduced to a story about a thief hunted by the police. While reading an eye-witness description of the thief on the computer screen, participants saw the pictures of four individuals. The participants had to select that/those picture(s) that depicted an individual matching the eye-witness description. For each test sentence there was an exhaustive and a non-exhaustive picture, and two distractors.

Each test item consisted of a single sentence, and four focus constructions were tested:[9]

(7) (a) A **kalap-ot** próbálta fel.
the hat-ACC tried PRT
"He tried on the **hat**."

(b) Felpróbálta a **kalap-ot**.
tried PRT the hat-ACC
"He tried on the **hat**."

(c) Csak a **kalap-ot** próbálta fel.
only the hat-ACC tried PRT
"He only tried on the **hat**."

(d) A **kalap** volt az, ami-t felpróbált.
the hat was it that-ACC tried
"It was the **hat** that he tried on."

The results reveal a significant difference between any two pairs of focus constructions: *only*-focus almost always received an exhaustive interpretation, while SUF had a really low proportion of exhaustive answers. Surprisingly, PVF and cleft-sentences also showed a significant difference, clefts more frequently were interpreted exhaustively (54% vs. 35%). Since the exhaustivity of clefts is a semantic entailment, the authors assume that the exhaustivity of PVF is not semantically encoded, and should rather be treated as an implicature. Regarding SUF, the outcome of the experiment confirmed the view that the exhaustivity of SUF is due to contextual factors. This raises the question whether there is a difference between the exhaustivity of PVF and that of SUF. More specifically, if the exhaustivity of PVF is an implicature, as the authors suggest, then the exhaustivity of PVF should also be treated as a context-dependent phenomenon.

9 It is not clear from the description of the design whether stress was somehow marked in the last three test sentences.

It must be added, however, that if we compare the two designs, not only the introductory *wh*-question, but the auditory clues were also removed from the target items. We believe that the lack of prosodic marking in the case of SUF may have influenced the results and might have given rise to a much smaller proportion of exhaustive answers.

3. The Experiment

When designing our experiment we wanted to compare the exhaustivity of PVF and SUF within the same experimental setting. The aim of our experiment was twofold: (i) to examine whether native speakers give higher ratings for PVF/SUF constructions in exhaustive contexts than in non-exhaustive ones, (ii) to investigate whether a marked structure (PVF/SUF) receives higher ratings when its use is motivated by the unexpectedness of the focused constituent. In order to be able to test these assumptions in an experimental framework, we assessed two factors: EXHAUSTIVITY and EXPECTEDNESS. In what follows we describe the working definitions we constructed for the purposes of our experiment.

Adopting Kamp's (MS.) classification of contexts as cited in Riester (2008), an articulated context consists of a dynamic discourse context, an environment context, a generic context and an encyclopedic context. Since in our experiment we only used question-answer pairs, the dynamic discourse context and the generic context are not relevant for our purposes. The environment context contains all entities in the immediate physical environment, while the encyclopedic context "consists of the entities that the speaker may assume his addressee to have knowledge about" (Riester 2008, 517). The encyclopedic context also contains all kinds of information about the entities in question. In our experiment exhaustive vs. non-exhaustive settings are differentiated, depending on which entity or entities of the environment context are being acted upon as depicted in the accompanying picture. This means that when only the entity being referred to by the focused constituent is shown, then the setting is exhaustive.

Regarding the other factor, EXPECTEDNESS, we follow Destruel and Velleman's (2014) distinction between expectations about the world and expectations about the discourse. For our purposes it is enough to consider expectations about the world which involve "beliefs about the world, expressed as assertions or presuppositions" (199). Hence, we make a distinction between expected and unexpected patients within an event being described. A patient is (un)expected in an event when its particular appearance is (in)compatible with our general assumptions about the event in question (based on our encyclopedic knowledge). Our target items were utterances where (un)expected patients occurred in focus.

We used a mixed factorial design in our experiment, testing two factors with two levels each: EXHAUSTIVITY: exhaustive vs. non-exhaustive settings and EXPECTEDNESS OF THE FOCUSED ELEMENT: expected vs. unexpected patients in focus. We presented 5 items

in each condition, i.e., we had 20 test sentence-picture pairs and 12 fillers. We also had a between-subjects variable: FOCUS TYPE, PVF vs. SUF. One group of participants was tested only for PVF structures, and the other only for SUF.

The four conditions are illustrated below, preceded by a sample test item illustrating both types of answers: PVF/SUF (also see Figure 1 and Figure 2):

1. an exhaustive setting with an expected patient in focus
2. a non-exhaustive setting with an expected patient in focus
3. an exhaustive setting with an unexpected patient in focus
4. a non-exhaustive setting with an unexpected patient in focus

(8) Q: Mi-t fogott ki Bence?
what-ACC caught PRT Bence
"What did Bence catch?"

A (PVF): Bence egy **hal-at** fogott ki.
Bence a fish-ACC caught PRT
"It was a **fish** that Bence caught."

A (SUF): Bence kifogott egy **nyaklánc-ot.**
Bence caught a necklace-ACC
"Bence caught a **necklace**."

Figure 1. Conditions 1 and 2

Figure 2. Conditions 3 and 4

3.1 Method

We used a sentence-picture verification task, where pictures were accompanied by a short dialogue embedded in a context. Background information was provided by an introductory sentence presented as part of the picture (*Bence a vidámparkban a horgászmedencénél játszott* "Bence was playing at the fish pool at the funfair"). First, participants read these very brief descriptions while looking at the picture. Following this, they heard an auditory stimulus, a question-answer pair. We used auditory and not written stimuli in order to exploit the prosodic clues, since these play an important role in the processing of the focus structures under investigation. The utterances were recorded by two different native speakers of Hungarian as a dialogue in order to create as natural stimuli as possible. Participants listened to a given dialogue only once while looking at the picture depicting the situation. Their task was to rate the acceptability of the answer on a 6 point Likert scale (1: totally unacceptable, 6: totally acceptable).[10] They had to press a button on a laptop to indicate their choice without time limit restrictions. We used the Pypres toolkit developed by Daniele Panizza to conduct the experiment; test items were presented in an individually randomized order for each participant.

10 We decided to use a 6 point Likert scale, since Pintér (2016) showed that binary judgement tasks are not appropriate to study the exhaustivity of PVF. Experiments relying on rating tasks, however, were able to detect a difference between neutral and focus structures even in the case of children.

3.2 Participants

66 university students participated in the experiment; they were all native speakers of Hungarian and were not receiving linguistic training at the time of the experiment. The subjects were randomly selected and they were randomly assigned to the two groups. The PVF and SUF groups involved 32 and 34 students, respectively. Further details about the participants are shown in the table below. Each participant received a small gift at the end of his or her session.

	Men	**Women**	**Total**	**Average age**
Group 1: PVF	14	18	32	22
Group 2: SUF	13	21	34	21

Table 1. Participants

3.3 Predictions

It was mentioned above that Gerőcs et al. (2014) argued that the exhaustivity of PVF is an implicature, while Surányi (2011) claimed that SUF might also be interpreted exhaustively. In other words, it might be assumed that the exhaustive interpretation of both focus structures is triggered by contextual factors. Taking this assumption as a starting point, we expected to get similar results regarding the acceptability of these focus structures in exhaustive vs. non-exhaustive settings (EXHAUSTIVITY). The role of the other factor (EXPECTEDNESS) in motivating the use of a marked structure has been examined by Skopeteas and Fanselow (2011), but only for PVF. In our experiment we wanted to test and compare the acceptability of both PVF and SUF constructions. If the exhaustivity of PVF and SUF is context-dependent, and EXPECTEDNESS is a contextual factor, then we can expect a similar type of behavior from both focus structures with respect to EXPECTEDNESS.

3.4 Results

Descriptive statistics across the four conditions and a diagram representing the overall results are shown below in Table 2 and Figure 3, respectively. As is shown in Figure 3, PVF and SUF sentences received similar ratings across the conditions.

	PVF/ SUF	**Mean**	**Standard deviation**
Exhaustive, expected	PVF	5.69	.74
	SUF	5.8	.41
Exhaustive, unexpected	PVF	5.4	.83
	SUF	5.61	.58
Non-exhaustive, expected	PVF	3.03	.88
	SUF	3.36	.92
Non-exhaustive, unexpected	PVF	3.14	.85
	SUF	3.36	.96

Table 2. Descriptive statistics

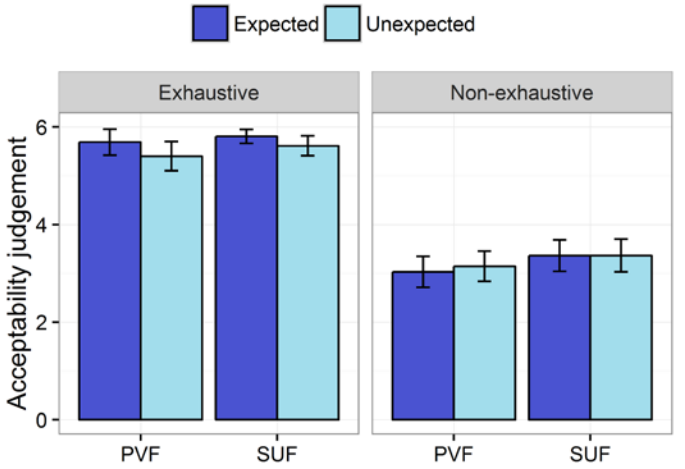


Figure 3. Overall results

We carried out a mixed design ANOVA in order to analyze the results. There was a main effect of EXHAUSTIVITY ($F(1, 64) = 406.9$, $p < .001$, $\eta_p^2 = .864$), which means that exhaustive settings received significantly higher ratings than non-exhaustive settings. The main effect of EXHAUSTIVITY is shown in Figure 4.

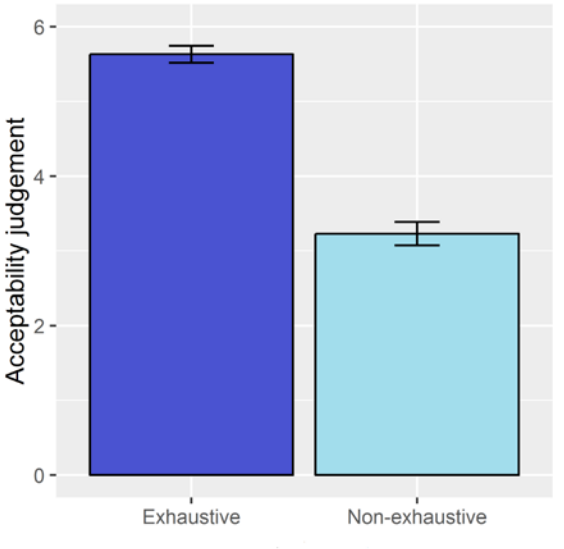


Figure 4. The main effect of EXHAUSTIVITY

There was no effect of EXPECTEDNESS ($F(1, 64) = 3.614$, *ns.* $\eta_p^2 = .053$), i.e., mean ratings were almost the same for expected/unexpected patients in focus. There was no effect of FOCUS TYPE either ($F(1, 64) = 2.573$, *ns.* $\eta_p^2 = .039$), i.e., if we ignore all other factors, PVF ratings were basically the same as SUF ratings. The diagrams below clearly illustrate that these measures did not have an effect.

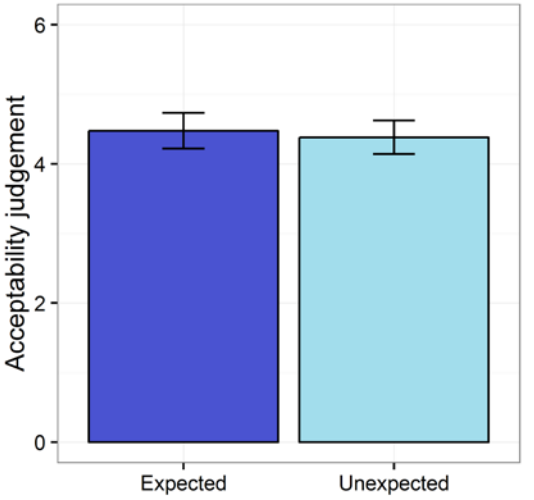


Figure 5. No effect of EXPECTEDNESS

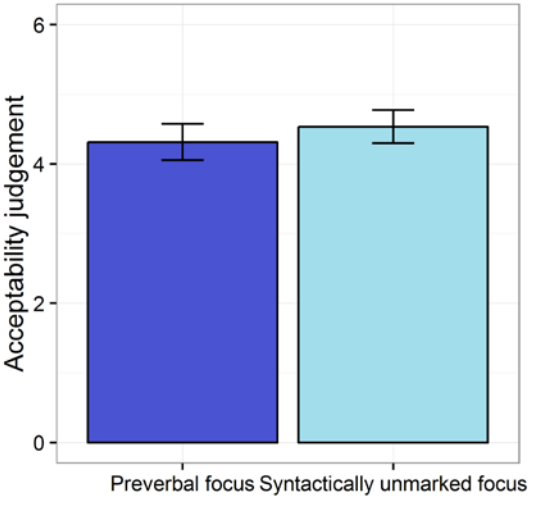


Figure 6. No effect of FOCUS TYPE

There was a significant interaction between EXHAUSTIVITY and EXPECTEDNESS ($F(1, 64) = 9.32$, $p < .05$, $\eta_p^2 = .127$), which means that if we ignore FOCUS TYPE, the profile of ratings across different levels of exhaustivity was different for expected and unexpected patients (see Figure 7). More specifically, the difference between the exhaustive and non-exhaustive conditions was smaller when the focused constituent is unexpected than when it is expected. It seems to be the case that the interaction is slightly stronger for PVF (the differences between the exhaustive vs. non-exhaustive mean ratings in the expected condition are: 2.66 for PVF and 2.44 for SUF, while in the unexpected condition they are 2.26 and 2.25, respectively). However, the interaction is quite weak, accounting for only 12.7 % of the total variability in the acceptability judgements.

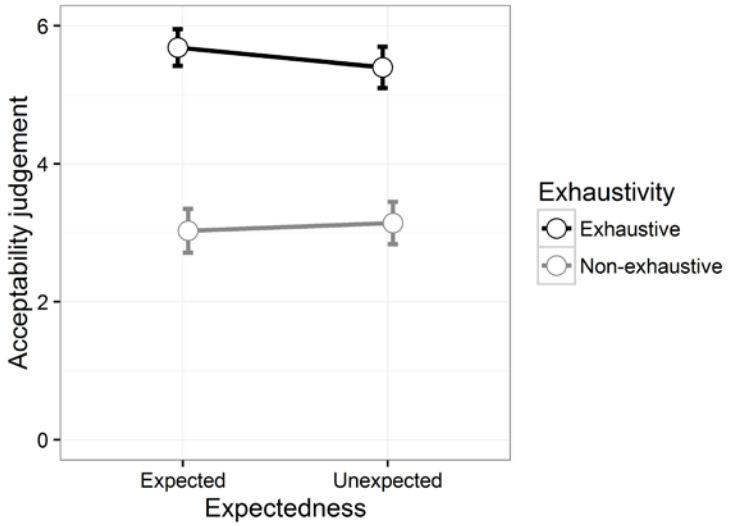


Figure 7. The interaction between EXHAUSTIVITY and EXPECTEDNESS

3.5 Discussion

As outlined above, we found a main effect of EXHAUSTIVITY, i.e., there was a significant difference between the ratings of exhaustive and non-exhaustive settings. More specifically, mean ratings of both focus structures were significantly smaller in the case of non-exhaustive settings. These results are in line with Gerőcs et al.'s (2014) findings, who found that the presence of an introductory *wh*-question resulted in a higher proportion of exhaustive interpretations for both PVF and SUF.

We leave it to future research to investigate the extent to which exhaustivity is primed by the presence of the introductory *wh*-question, by the presence of the prosodic clues or by the presence of these two features together, and plan to conduct a follow-up experiment

using the same stimuli without the presence of the introductory *wh*-question. The results of the follow-up experiment might also contribute to the ongoing debate about the exhaustivity of PVF. For instance, Kas and Lukács's (2013) results based on an experiment using a sentence-picture verification task with binary judgements questioned the exhaustivity of PVF. However, the authors admit that there was huge individual variation even among the adult participants. As mentioned above, Pintér (2016) pointed out that binary judgement tasks might not be subtle enough to explore the exhaustivity of PVF, hence the inconsistencies found by Kas and Lukács (2013) might stem from a methodological flaw.

In essence, we believe that the most important finding here is the fact that EXHAUSTIVITY had the same effect on both PVF and SUF. Therefore, experimental evidence has been provided in favor of Surányi's (2011) claim about the exhaustivity of SUF. In addition, the results contradict those views that treat PVF as necessarily exhaustive and SUF as necessarily non-exhaustive when they form an answer to a *wh*-question (cf. for example É. Kiss 1998, Horváth 2006).

There was no effect of EXPECTEDNESS; it did not matter whether the patient in focus was considered expected or unexpected for the hearer as judged by the speaker, which shows that this contextual factor does not influence the acceptability of the focus structures in question. Skopeteas and Fanselow (2011) also showed that Hungarian PVF is interpreted independently of the predictability of the referent. However, we believe that the conclusion that contextual factors do not play any role in the interpretation of PVF and SUF in Hungarian is too strong, since other extra-grammatical factors not tested here may still influence the use of these focus structures.[11]

It is important to note here that there was a weak interaction between EXHAUSTIVITY and EXPECTEDNESS, and the interaction is stronger for PVF. This weak interaction suggests that the exhaustivity of PVF is not entirely independent of contextual factors, but it is too marginal to draw stronger conclusions.

Our results also confirm the widely accepted view that PVF constructions in sentential answers given to a *wh*-question are interpreted exhaustively. This means that the *wh*-question expresses a request to the addressee and the speaker's expectation is that the answer will specify the exact subset of entities of which the question predicate holds, i.e., in other terms the answer is a maximal answer (cf. Balogh 2009, Surányi 2011, Destruel et al. 2015).

11 Gábor Alberti (pers. comm.) drew our attention to the possibility that SUF constructions seem to be more acceptable when evidentiality plays an important role in the given situation. For example, if someone has accidentally thrown away the remote control, one might react as follows:

(9) Nézd már mit csinált ez a lüke! Kidobta **a távirányítót**!
"Look what this nutter has done! He has thrown away **the remote control**!"

Further experiments are called for to explore the role of evidentiality in interpreting SUF.

4. Conclusions

In this paper we reported the results of an experiment we conducted to compare Hungarian preverbal or structural focus (PVF) and information or syntactically unmarked focus (SUF) within the same experimental setting. Previous literature on SUF argued that, as opposed to PVF, SUF cannot receive an exhaustive interpretation. However, Surányi (2011) questioned this view and argued for the possibility that SUF might also be interpreted exhaustively. Our aim was to collect empirical evidence for or against Surányi's (2011) view, and to gather further data on the interpretation of both focus structures.

We used a sentence-picture verification task where participants rated utterances on a 6 point Likert scale. Target sentences were always presented in a context after an introductory *wh*-question. We tested two factors, EXHAUSTIVITY and EXPECTEDNESS, and we also had a between-subjects variable, FOCUS TYPE.

First of all, experimental evidence was collected in favor of the claim that SUF might receive an exhaustive interpretation. We found that participants gave higher ratings to both structures in exhaustive settings, and the results were almost the same for the two types of focus structures. This also means that there is no clear-cut difference between the exhaustivity of PVF and SUF, i.e., it is not the case that PVF is necessarily exhaustive, while SUF is necessarily non-exhaustive.

To sum up, our results reinforce the outcomes of the first experiment in Gerőcs et al.'s (2014) study, concerning the lack of empirical difference between PVF and SUF. The results might be interpreted as a challenge to the standard view, i.e., if SUF is exhaustive (provided that there is an introductory *wh*-question), then it might be the case that the exhaustivity of PVF is not necessarily inherent in nature. Therefore, we conclude that the exhaustive interpretation is pragmatic in nature that arises from the presence of the *wh*-question in both cases (PVF and SUF). At this point it is not clear whether it is due to a presupposition, as Pintér (2016) claims, or an implicature (cf. Gerőcs et al. 2014, Babarczy and Balázs 2016). The pragmatic explanation also accounts for the similar results obtained for both structures.

Funding Acknowledgement

This research was supported by the National Research, Development and Innovation Office (NKFIH), grant No. K 111918.

Works Cited

Babarczy, Anna, and Andrea Balázs. 2016. "A kognitív kontroll és a preverbális fókusz értelmezése." In *Szavad ne feledd! Tanulmányok Bánréti Zoltán tiszteletére*, edited by Bence Kas, 151–63. Budapest: Magyar Tudományos Akadémia Nyelvtudományi Intézet.

Balogh, Kata. 2009. *Theme with Variations: A Context-Based Analysis of Focus.* Amsterdam: ILLC Dissertation Series.

Destruel, Emilie, and Leah Velleman. 2014. "Refining Contrast. Empirical Evidence from the English *it*-Cleft." In *Empirical Issues in Syntax and Semantics 10. Selected papers from CSSP 2013,* edited by Christopher Pinon, 197–214. http://www.cssp.cnrs.fr/eiss10/

Destruel, Emilie, Daniel Velleman, Edgar Onea, Dylan Bumford, Jingyang Xue, and David Beaver. 2015. "A Cross-linguistic Study of the Non-at-issueness of Exhaustive Inferences." In *Experimental Perspectives on Presuppositions*, edited by Florian Schwarz, 137–56. Berlin: Springer.

É. Kiss, Katalin. 1998. "Informational Focus vs. Identification Focus." *Language* 74 (2): 245–73.

Gerőcs, Mátyás, Anna Babarczy, and Balázs Surányi. 2014. "Exhaustivity in Focus: Experimental Evidence from Hungarian." In *Language Use and Linguistic Structure*, edited by Joseph Emonds and Markéta Janebová, 181–94. Olomouc: Palacký University.

Horváth, Júlia. 2006. "Separating "Focus Movement" from Focus." In *Phrasal and Clausal Architecture,* edited by Simin Karimi, Vida Samiian, and Wendy K. Wilkins, 108–45. Amsterdam: John Benjamins.

Kamp, Hans. 2008. "Discourse Structure and the Structure of Context." MS., IMS, Universitaet Stuttgart.

Kas, Bence and Ágnes Lukács. 2013. "Focus Sensitivity in Hungarian Adults and Children." *Acta Linguistica Hungarica* 60 (2): 217–45.

Onea, Edgar and David Beaver. 2011. "Hungarian Focus Is Not Exhausted." In *Proceedings of the 19th Semantics and Linguistic Theory Conference*, edited by Satoshi Ito Cormany and David Lutz, 342–59. Ithaca: Cornell University.

Pintér, Lilla. 2016. "A magyar szerkezeti fókusz kimerítő értelmezésének kísérletes vizsgálata." In *STUDIA VARIA: Tanulmánykötet*, edited by József Balázs, Anita Bojtos, Tamás Paár, Zsófia Tompa, Gergő Turi, and Noémi Vadász, 191–212. Budapest: Pázmány Péter Katolikus Egyetem.

Riester, Arndt. 2008. "A Semantic Explication of Information Status and the Underspecification of the Recipients' Knowledge." In *Proceedings of SuB12*, edited by Atle Grønn, 508–22. Oslo: ILOS.

Skopeteas, Stavros and Gisbert Fanselow. 2011. "Focus and the Exclusion of Alternatives: On the Interaction of Syntactic Structure with Pragmatic Inference." *Lingua* 121 (11): 1693–1706.

Surányi, Balázs. 2011. "A szintaktikailag jelöletlen fókusz pragmatikája." *Általános Nyelvészeti Tanulmányok* 23: 281–313.

Szabolcsi, Anna. 1981. "Compositionality in Focus." *Folia Linguistica* 15: 141–61.

Wedgwood, Daniel. 2005. *Shifting the Focus. From Static Structures to Dynamics of Interpretation.* Amsterdam: Elsevier.

Zimmermann, Malte. 2008. "Contrastive Focus." In *The Notions of Information Structure.* (*Working Papers of the SFB 632*, *Interdisciplinary Studies on Information Structure* 6), edited by Caroline Féry, Gisbert Fanselow, and Manfred Krifka, 147–59. Potsdam: Universitätsverlag.

Hungarian Focus: Presuppositional Content and Exhaustivity Revisited

Tamás Káldi,[a] Anna Babarczy,[b] and Ágnes Bende-Farkas[c]

[a,b]Research Institute for Linguistics (HAS), Budapest, Hungary; [c]Department of Cognitive Science (BME), Budapest University of Technology and Economics, Hungary

[a]kaldi.tamas@nytud.mta.hu; [b]babarczy@cogsci.bme.hu; [c]agnesbf@gmail.com

Abstract: Hungarian has a syntactically marked focus construction which has been associated with exhaustive interpretation. The factors behind exhaustivity have generated an extensive debate: some theorists argue that this interpretation is determined at the syntax-semantics interface, while others argue that it is the result of a pragmatic inference. Previous experimental work supports the latter view. In the present study we hypothesized that within pragmatic inferences the exhaustivity associated with preVf is the result of scalar implicature generation. To test our hypothesis we conducted three eye-tracking experiments using a lexically marked focus construction as a baseline. Our results support the hypothesis: a strong context dependence and a delay in processing relative to the baseline are in line with earlier experimental data on scalars. We thus suggest that future research on the exhaustivity of Hungarian focus should concentrate on potential contextual effects.

Keywords: focus interpretation; exhaustivity; eye-tracking; scalar implicatures; contextual effects

1. Introduction

In Hungarian, a discourse configurational language, information structural functions like Topic, Focus and Comment are assigned different syntactic positions within the sentence. Among these functions Focus is one that has been subject to extensive research, and whose semantic and pragmatic properties have been debated vigorously. In the center of this debate lies the issue of exhaustive interpretation: although there is a consensus that the

Hungarian Focus construction has (or tends to have) an exhaustive interpretation, traditional generativist accounts claim that exhaustivity is computed at the syntax-semantics interface, whereas alternative, pragmatic approaches favor the idea that exhaustivity is the result of a pragmatic inference. Experimental work supports the latter view.

The aim of the present study is twofold: i) we make an attempt to adapt the visual world paradigm in order that the online investigation of the Focus structure becomes possible, ii) we use this method to test hypotheses derived from the results of earlier experimental work.

The results of our work are in line with those of earlier experimental research and mark out a new possible line of research on Hungarian Focus.

1.1 Theories of Focus Interpretation

The Focus structure investigated in the current paper is presented in (1a) together with its neutral, canonical pair in (1b).

(1) (a) János ‘Marit hívta meg.
John-NOM Mary-ACC call-3Sg-PAST verbal prefix
“It was Mary who John invited.”

(b) János meg hívta Marit.
John-NOM verbal prefix call-3Sg-PAST Mary-ACC
“John invited Mary.”

Formally, the differences between the two sentence-types are the following: in (1b) the verbal prefix is located before the verb constituting one phonological word with it, the NP referring to Mary is in post-verbal position. In the Focus construction in (1a), however, the corresponding NP is in pre-verbal position while the verbal prefix is situated post-verbally. Now it is the immediately pre-verbal NP that constitutes a phonological word with the verb.[1] The construction in (1a) has been assigned various labels depending on what was taken to be its most prominent semantic property. For example É. Kiss (1998), in her influential paper calls (1a) and (1b) identification- and information-focus respectively making a case for the terminological distinction on the following grounds. While identification-focus identifies a subset of the contextually available set of entities for which the predicate holds, and this identification is exhaustive, information-focus merely conveys new, non-presupposed information and exerts no exhaustivity effect. Since in the present work we will argue for an alternative

1 Moreover, as first Kálmán and Kornai (1989) noted, elements in the pre-verbal position receive an eradicating stress, meaning that the most prominent stress is placed on this element, while the domain following it has a flat prosody (i.e., is void of other elements bearing main stress).

approach to focus interpretation we will use the theory-neutral terms pre-verbal focus for (1a) (preVf henceforth) and neutral sentence for (1b).

As mentioned earlier, there is a heated debate about the interpretational characteristics of preVf in the literature: while no one questions the observation that preVf has an exhaustive interpretation, different theoretical frameworks explain this characteristic by postulating different interpretational processes.

Generative frameworks predominantly associate preVf with a [+ exh] operator (e.g., Szabolcsi 1981; Szabolcsi 1994; Kenesei 1994; É. Kiss 1998; É. Kiss 2004; Horváth 2010). For example É. Kiss (2004) provides a widely accepted analysis in the Minimalist framework considering movement to the pre-verbal position an operator movement where the focused element moves to spec-FP, where the head of FP contains the feature [+exh]. The trace of the moved element is bound by FP, and "the scope of focus . . . is the domain c-commanded by the constituent in spec-FP" (É. Kiss 2004, 86–87). This accounts for the observation that the c-commanded predicate part of the sentence exhaustively holds for the referent of the focused element. The operator analysis thus posits a deterministic relationship between form and meaning (i.e., exhaustive meaning) in the case of preVf, in which exhaustive interpretation is considered an entailment.

Alternative accounts on the other hand claim that the exhaustive interpretation in the case of preVf is not semantically determined but results from pragmatic processes. Wedgwood (2003), for example, upholds that it is unnecessary to posit an operator; preVf is underspecified for exhaustivity. As Wedgwood's reasoning goes, any component of the meaning of a structure can only be considered semantically determined if that aspect is context independent. If, however, this component depends on context, it is best seen as a pragmatic phenomenon. Through a corpus analysis of the contexts preVf appears in, Wedgwood (2003) comes to the conclusion that the exhaustive interpretation of preVf is variable, which he considers to be a strong empirical argument for the pragmatic view of preVf interpretation.

A third line of inquiry proposes that the exhaustivity of Focus is a (semantic) presupposition as has been suggested for the exhaustivity of English it-clefts (Velleman et al. 2012; Buring and Kruz 2013). There is some disagreement, however, in the literature on presupposition whether it should be considered a semantic phenomenon or a pragmatic implicature with some recent work convincingly arguing for the latter view (Schenkler 2008; Chemla and Bott 2013).

The theoretical debate on the nature of exhaustivity of Focus inspired an extensive array of experimental work (see, e.g., Onea and Beaver 2011; Kas and Lukács 2013; Gerőcs et al. 2014; Káldi 2015), which uniformly supports the pragmatic view of preVf interpretation. For the purposes of the present work Gerőcs at al. (2014) provides the most relevant results. In their first experiment the authors compared the interpretation of preVf sentences in a short- and a long time-condition in a picture-sentence verification experiment: participants heard preVf or neutral sentences, after

which they saw images corresponding to exhaustive or non-exhaustive scenarios. The experimental task was to judge whether the image matched the sentence or not. In the short-condition participants had to respond within 1000 ms after visual stimulus onset (the end of time limit was signaled with a beeping sound), while in the long-condition they had 3000 ms to give their answer. Based on the hypothesis that during sentence processing, pragmatic interpretation is preceded by semantic processing, Gerőcs at al. (2014) predicted that the limitation of time, and therefore of cognitive resources available for the process of interpretation will result in responses that reflect the semantic meaning of the presented sentences, whereas if more time is available (and thus pragmatic enrichment can take place), responses will reflect pragmatic meaning. The results were in line with the authors' predictions: in the short-condition the rate of exhaustive responses in the case of preVf sentences was around chance level, while in the long-condition the rate of exhaustive responses was significantly higher (72%) but still well below 100%. Gerőcs at al. (2014) concluded that the exhaustive interpretation of preVf emerges as a pragmatic inference. In their second experiment the authors compared the interpretational characteristics of lexically marked (*only*) focus (*only*-f henceforth), preVf and cleft sentences in a sentence—picture matching paradigm. Participants read a sentence of one of the above types and had to decide which one or more of four images matched the sentence best. The set of four images included one depicting an exhaustive interpretation, a non-exhaustive image and two distractors. According to the results participants gave an exhaustive response in 98% of the trials in the *only*-condition, while the rate of exhaustive responses was well below that in the cleft and preVf-conditions (54% and 35% respectively). Gerőcs at al. (2014) concluded that these results support the view that exhaustivity is not entailed but emerges as a result of a pragmatic inference.

1.2 The Role of Pragmatics in the Exhaustive Interpretation of preVf

As mentioned above the results of the bulk of experimental research on the exhaustive interpretation of preVf support the view that exhaustivity is tied to some sort of pragmatic inference. The purpose of the present work is to investigate the hypothesis that the pragmatic inference in question is scalar implicature. Let us briefly present the theory behind scalars and its relevance for accounting for the exhaustiveness effects related to preVf.

According to neo-Gricean accounts (see, e.g., Horn 1972; Gazdar 1979), certain sets of terms can be ordered on a scale of strength of meaning.[2] A classic example is the scale of the connectives *or* and *and*. The expression containing *or* in (2a) in a given context may be interpreted exclusively, i.e., excluding the possibility of Peter buying both an apple and an orange. This interpretation is called upper bounded in the theory,

2 Apart from the conjunctions discussed here scalar expressions include certain quantifiers (e.g., *some* < *most* < *all*) and adjectives (*warm* < *hot*) etc.

as not all of its (possible) referents are included in the set to which the predicate can apply. For this reason, *or* is also referred to as a weak term. However, as shown in (2b), the upper bounded interpretation is cancellable, demonstrating that its meaning can be compatible with the meaning of the stronger term on the scale and can have an inclusive, lower bounded interpretation, which corresponds to its logical interpretation.

(2) What do you think Peter bought at the market?
   (a) I think he bought an apple or an orange.

   (b) I think he bought an apple or an orange… Actually, I think he bought an apple *and* an orange.

In neo-Gricean terms, thus the upper bounded interpretation of *or* emerges as a scalar implicature in line with Grice's (1975, 45) Maxim of Quantity (3): the speaker in (2a) could have used the more informative (or stronger) expression *and* but opted for the weaker *or*. Assuming that the speaker observed the Maxim of Quantity, she must have used the weaker term because the stronger term did not apply.

(3) (a) Make your contribution as informative as required (for the current purposes of the exchange).

   (b) Do not make your contribution more informative than is required.

In sum, the interpretation of *and* is unambiguously inclusive at the level of semantics, while the interpretation of *or* is compatible with the interpretation of *and* at the semantic level, and its upper bounded interpretation is associated with a scalar implicature.

At this point it is important to note that although these examples of Gricean reasoning may create the impression that Grice and those in the neo-Gricean tradition attempted to describe actual psychological interpretational processes, these theories, as for example Geurts (2016) argues in detail, have no intended psychological reality. For this reason psycholinguistic studies turn to theories that allow researchers to make valid predictions on the mental interpretational processes themselves related to such expressions. One of these is Relevance Theory (Sperber and Wilson 1995). Relevance Theory claims that scalar expressions are semantically underspecified for certain aspects of meaning: for example *or* is underspecified for exclusive interpretation. As far as the process of interpretation is concerned Relevance Theory relies on the notion of Relevance, which is defined in the following way: everything else held constant, the greater the positive cognitive effect of an input, the higher its Relevance, while the greater the processing cost related to that input, the lower its Relevance. Scalar implicatures arise only if they are Relevant.

Fekete et al. (2014) examined the above Relevance Theoretic hypothesis concerning the underspecification of scalar terms. The authors employed a shallow processing paradigm in which they compared the interpretational characteristics of *or* and *and*. During the experiment subjects heard sentences in which object NPs were either coordinated with *and* or *or*, e.g., *John peeled the orange and/or the banana*. Following the sentences, images were shown that depicted the referents of the NPs in two possible states: either only one of the objects was manipulated or both of them (e.g., an orange intact and a banana peeled, or both peeled). The former was congruent with an exclusive meaning (the pragmatic interpretation of *or*), while the latter was congruent with an inclusive meaning (the meaning of and the logical meaning of *or*). The experimental task was to decide if the objects in the pictures had been mentioned in the sentences or not regardless of their state in the image. The dependent measure was reaction time (RT). Since in a shallow processing task reaction times are known to be slower when the state of the objects in the picture fails to match the meaning of the sentence, participants should take longer to respond to exclusive pictures than to inclusive pictures in the *and*-condition. Fekete et al. further predicted that if *or* is underspecified for exclusiveness as hypothesized by Relevance Theory, reaction times should not be affected by picture type in the *and*-condition. This is exactly what they found. RT was significantly slower for exclusive pictures than for inclusive pictures in the *and*-condition indicating that the inclusive interpretation of *and* was automatically processed. In the *or*-condition, however, there was no difference in RT between the inclusive and exclusive picture conditions suggesting that the exclusive and inclusive interpretations of *or* were equally active.

Relevance Theory also provides a framework for the interpretation of the results of Gerőcs et al (2014) on preVf. In the short-condition, due to the limited amount of available time, the cognitive resources necessary for implicature generation were not available, hence the exhaustive interpretation was less likely to be calculated. PreVf is underspecified for exhaustivity; the exhaustive interpretation corresponds to the upper-bounded, whereas non-exhaustive interpretation corresponds to lower-bounded reading.

In line with the reasoning outlined above our hypothesis is that the exhaustive interpretation of preVf emerges as a scalar implicature. To test our hypothesis we conducted three eye-tracking experiments. Experiment 1 examines two operators that show a similar difference in interpretational characteristics to what we hypothesize in the case of *only*-f and preVf sentences. These are conjunction (*and*) and disjunction (*or*). Results will enable us to infer what the correlates of semantic and pragmatic interpretational processes are in the eye-tracking data. Experiment 2 examines whether people are sensitive to the theoretically motivated presuppositional differences between *only*-f and preVf sentences. This is important if we would like to use *only*-f sentences as a baseline condition in the study on preVf sentences. Finally, experiments 3 and 4 investigate the interpretation of preVf.

2. Experiment 1

2.1 Materials and Method

The eye-tracking experiments presented here were recorded with an EyeLink 1000 Desktop Mount eye-tracker. In the first experiment 35 adult native Hungarians participated with normal or corrected to normal vision. 7 participants were excluded from the analysis as data collected from them was not interpretable (due to a high amount of track losses, calibration issues, etc.).

During the experiment 8 critical linguistic stimuli were presented. These were sentences that contained object NPs coordinated by either *and* or *or* (4).

(4) (a) *and*-cond.

Félbevágta	az	epret
cut-in-half-3Sg-PAST	the	strawberry-ACC
és	a	kivit.
and	the	kiwi-ACC

"She split the strawberry and the kiwi."

(b) *or*-cond.

Félbevágta	az	epret
cut-in-half-3Sg-PAST	the	strawberry-ACC
vagy	a	kivit.
or	the	kiwi-ACC

"She split the strawberry or the kiwi."

Linguistic stimuli were recorded in an adult male voice. Visual stimuli were presented in four image quadrants on a screen. The quadrants included an inclusive and an exclusive scenario and two distractors (Fig. 1).

Figure 1. Example display (monochrome version) for the eye-tracking experiments (i.e., experiment 1, 3, 4). Inclusive scenario: bottom-left; exclusive (or exhaustive) scenario: upper-right, distractors: upper-left and bottom-right.

The procedure went as follows. Participants saw a fixation cross which they had to fixate for 1700 ms. The cross disappeared and the visual and linguistic stimuli were presented simultaneously. The task of the participant was to choose the image that best matched the linguistic stimulus. Responses were given with a hand-held 5-button response box (type: RESPONSEPixx Handheld).

The experiment consisted of three blocks: two practice blocks of 17 and 8 trials and a test block of 26 trials. The test block consisted of 4 sentences in the *and*- and 4 in the *or*-condition and 18 additional fillers. Linguistic stimuli were presented in a random order in such a way that each sentence was presented in only one condition. Also, the position of image types and the order of objects within the images were balanced across the whole experiment.

Since the use of an NP coordinated by *or* in a situation whose circumstances are known to all interlocutors is infelicitous, it was necessary to "distance" the linguistic stimuli from the situations depicted by the images. For this reason we created a context story: at the beginning of the second practice block participants were asked to imagine that they were listening to the sentences of eye-witnesses of crimes, and as paralegals, they had to decide which scenario best matched the witnesses' description. Participants were given feedback on their response after each trial. At the beginning of the third block participants were informed that they are not paralegals any more, but real decision-makers and will not be given any further feedback.

The dependent measures were choice of image and the proportion of fixations (PoF) on the quadrants as a function of time. Regarding the choice of images we expected the interpretation of *and*-sentences to be invariantly inclusive, while the interpretation of *or*-sentences—in the absence of a disambiguating context—to be divided between inclusive and exclusive. Regarding the eye-tracking data we expected the fixations to converge faster on the inclusive image in the *and*-condition than in the *or*-condition. This would mean that eye-gaze data show a greater degree of hesitation in the case of processing *or*-sentences than in the case of *and*-sentences even if the explicit choice of images does not reveal a difference between the two conditions.

2.2 Results

The results of the experiment were in line with our expectations. The choice of exclusive image in the *and*-condition was 100%, while in the *or*-condition responses diverged: 18% of participants always chose the inclusive image, 57% always chose the exclusive image and 25% varied in their choices.

As a direct comparison of eye-tracking data was possible only in those cases where the participants gave an inclusive response, we report data from these trials (data gathered from 12 participants).

Figure 2 presents the PoF on the inclusive image as a function of time calculated as the number of fixations on the inclusive image as a proportion of the number of fixations on the inclusive + the exclusive image. We split the trials into three interest periods (IP).

(5) IP1: From stimulus onset to connective onset.
IP2: From connective onset to sentence offset.
IP3: From sentence offset to average RT.

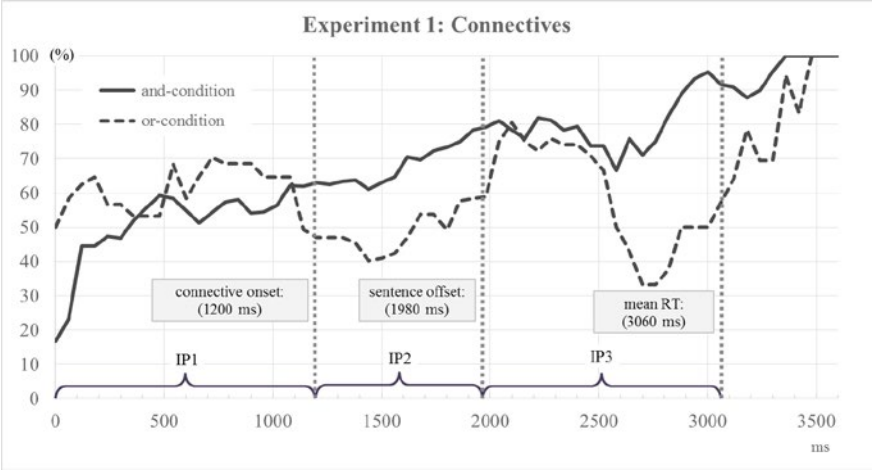


Figure 2. PoF on the inclusive quadrant (%)

As Figure 2 shows, looks on the two competing image types are around chance level in both conditions before the connective. After connective onset, however, PoF starts to converge on the inclusive image in the *and*-condition, while in the *or*-condition it heavily fluctuates and stays around 50%. This pattern in the *or*-condition is carried over to IP3 despite the fact that only data from the inclusive response type are presented here. Data were analyzed with a two-way repeated measures ANOVA. Factors were IP (2nd and 3rd) and Connective type (*and* and *or*). According to the results, the main effect of Connective Type was highly significant ($F(1, 11) = 35.91$; $p < .001$), however, there was no main effect of IP and interaction. Data thus reflect that already immediately after the connective and before the offset of the sentence PoF diverged in the two conditions: while in the case of *and* participants quickly look away from the exclusive image, in the *or*-condition the hesitation remains regarding the meaning of the connective. This hesitation in our interpretation of the data reflects the underspecification of *or* for exclusive reading.

3. Experiment 2

3.1 Materials and Method

As mentioned earlier the purpose of our investigation was to examine the interpretational processes associated with the (tendentiously) exhaustive reading of preVf using *only*-f sentences as a baseline. However, theoretical concerns have been raised

regarding this comparison on the grounds that the two structures differ with respect to their division into presupposed and asserted content (cf., e.g., Horn 1969; Bende-Farkas 2009). We provide a simplified comparison of the relevant aspect of these differences in Table 1.

Sentence type	Example	Presupposed content	Asserted content
***only*-f**	Ő csak a kivit vágta félbe. "(S)he only split a kiwi."	(S)HE SPLIT THE KIWI	NOTHING ELSE
preVf	Ő a kivit vágta félbe. "It was the kiwi (s)he split."	(S)HE SPLIT SOMETHING	IT WAS THE KIWI

Table 1. An outline of the differences in the presuppositional and asserted content of *only*-f and preVf sentences

As Table 1 shows, although both structures are associated with an exhaustive reading, this reading is encoded in them in different ways. This underlying difference, as critics of the method claim, may result in confound effects in the comparison.

In order to investigate the validity of this criticism we conducted an online sentence completion survey. The survey consisted of 8 different test sentences (4 *only*-, 4 preVf) with two possible emphatic continuation phrases (6).

(6) (a) … mást nem.
else-ACC not
"(and) nothing else"

(b) … nem mást.
not else-ACC
"(and) not something else" (i.e., "[and] exactly that")

The phrases in (6a) and (6b) reflect on the presupposed content of the *only*-f and preVf sentences respectively, in a way that they are synonymous with the asserted content of the two sentence types: in the case of only-f, (6a) is literally synonymous with it, whereas in the case of preVf (6b) expresses the identification that the asserted part of the focus expresses. Therefore, we could rather say that the endings in (6) are emphatic and spelled out counterparts of the asserted contents of the respective sentence types. Along this line our prediction was that if respondents are sensitive to the presupposed content of *only*-f and preVf sentences, they would predominantly choose (6a) as a continuation for the former and (6b) for the latter. The survey contained an additional 24 filler items. All items were randomized in a way that minimally one filler item intervened between

two test items. The survey was administered through Google Sheets. We analyzed the data gathered from 50 adult native Hungarians.

3.2 Results and Conclusion

Participants completed *only*-f sentences with (6a) 97.5% (SD = 7.57) of the time, whereas that ending was chosen only 45.5% (SD = 36.17) of the times in the preVf-condition. A one sample t-test revealed that this rate is not significantly different from chance level. Based on the result that respondents found both (6a) and (6b) compatible with preVf sentences, we concluded that there is a strong possibility that people are not sensitive to the difference in the presupposed and asserted content of out-of-context preVf sentences as predicted by theory. Consequently, the above theoretical criticism does not hold, and *only*-f sentences can be used as a baseline condition in experimental investigations on the exhaustive interpretation of preVf.

4. Experiment 3

Experiment 3 compared the processes associated with the exhaustive interpretation of *only*-f and preVf sentences. In the light of the results of Experiment 1 and the theoretical considerations outlined in 1.2 we conjectured that data obtained in the *and*-condition in Experiment 1 would pattern with data obtained from the *only*-condition, whereas data from the *or*-condition would pattern with those from the preVf-condition.

4.1 Materials and Method

18 native Hungarian adults participated in Experiment 3, 2 of whom had to be excluded for technical reasons. None of the participants of Experiments 1 or 2 participated in Experiment 3. The procedure and data recording were identical to those of Experiment 1. The third experimental block contained 12 test trials (6 *only*-f and 6 preVf) and 24 fillers. Examples of linguistic stimuli associated with the visual stimuli (see example in Figure 1) are given in (7). Linguistic stimuli were recorded in an adult male voice.

(7) (a) *only*-cond.

Csak	a	kivit	vágta	félbe.
only	the	kiwi-ACC	cut-Sg-PAST	in-half

"(S)he split only the kiwi."

(b) preVf-cond.

A	'kivit	vágta	félbe.
the	kiwi-ACC	cut-Sg-PAST	in-half

"It was the kiwi (s)he split."

We used the context story presented in 2.1 with no modification. The dependent measures, just as in the first experiment, were the choice of image and PoF on the exhaustive image as a function of time.

4.2 Results

Regarding choice of image, the data were uniform: participants chose the image associated with the exhaustive scenario in 100% of the trials in both conditions.

For the analysis of eye-tracking data we created three IPs (8).

(8) IP1: From stimulus onset to verb onset.
IP2: From verb onset to sentence offset.
IP3: From sentence offset to average RT.

Figure 2 displays the PoF on the exhaustive image (number of fixations on the exhaustive image as a proportion of the duration of fixations on the exhaustive [target] + the non-exhaustive [alternative] image) as a function of time. In IP2, PoFs on the target and the alternative images are approximately equally distributed. In IP3, looking preference gradually moves towards the target image. There was no significant difference between the two conditions. Data were analyzed in a two-way repeated measures ANOVA (IP and Sentence Type). Only IP had a significant main effect ($F(1, 15) = 14.03$; $p = .002$).

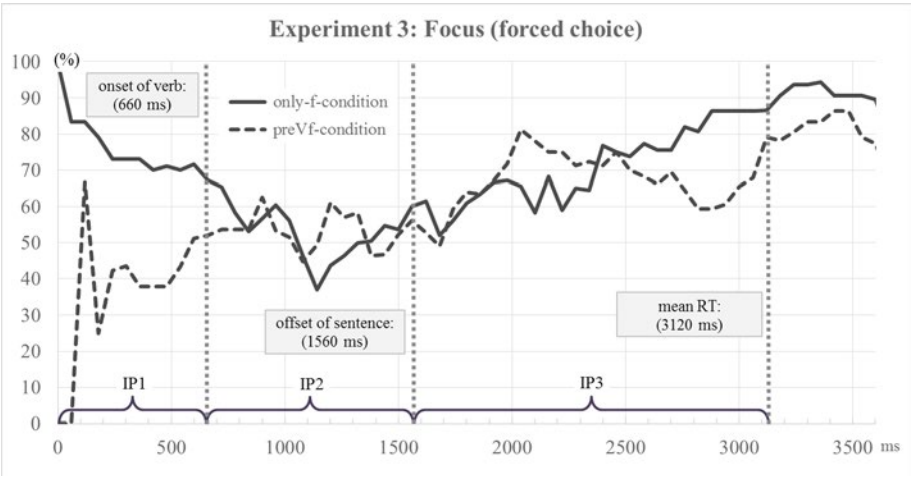


Figure 3. PoF on the exhaustive quadrant

4.3 Discussion

In Experiment 3, the explicit choices of image suggest that—contrary to earlier experimental results—participants uniformly exhaustively interpreted not only the lexically marked *only*-f sentences but also preVf sentences. Eye-tracking data also suggest that in the given experimental setup the interpretational process associated with the two structures is identical: PoF on the alternative non-exhaustive image decreases over time at the same rate in the three IPs.

The observed similarities of the interpretational processes associated with the two structures may support the traditional generative theories. However, they do no not contradict the pragmatic theories either as it can easily be appreciated that the experimental task introduced a contextual factor possibly having an effect on the results. As participants were instructed to choose only one image, the context of the task may have implied that there is one unambiguously "appropriate" choice in each trial. We thus assumed that the uniformity in both choices of image and eye-tracking data were the result of a forced-choice effect.

We based our assumption regarding the force-choice effect on earlier experimental work (see, e.g., Grodner et al. 2010; Bergen and Grodner 2012) which showed that the complexity of experimental task and other contextual factors have an effect on the interpretation of scalar expressions identifiable even in online, processing related data. In line with these, we hypothesized that the results of Experiment 3, though important, are not conclusive, and we have to control for the potential confound by allowing participants to choose any number of images.

5. Experiment 4

5.1 Materials and Method

30 native Hungarian adults participated in the third experiment. The context story for the experiment remained unchanged, and no modification was made to the experimental instructions except for one: participants were informed that they could choose any number image quadrants during the trials. Responses with the button-box were given as follows: participants pressed the button(s) corresponding to the quadrant(s) they felt matched the linguistic stimulus and then pressed a fifth (middle) button to signal the end of responding. Since this way of responding is slightly more complicated, the number of practice trials was increased to 31 in the first and 17 in the second block. The test (third) block consisted of 12 critical trials (6 *only*-f, 6 preVf) and 24 fillers.

5.2 Results

The frequencies of image choices are presented in Table 2. Response types were defined as follows. Exhaustive responses were those where participant chose only the exhaustive image, and non-exhaustive responses where participant chose both the exhaustive and the non-exhaustive image or only the non-exhaustive image. We excluded trials in which distractor images were chosen (3 trials). Exhaustive responses were given 93.33% of the time in the *only*-condition but only 65.0% of the time in the preVf-condition.

Our analysis, as in the first experiment, is restricted to trials in which participants gave the response that is congruent with both of the test structures (i.e., the exhaustive reading in this case). As 1 out of 30 participants gave a non-exhaustive response in all trials, we analyzed data gathered from 29 participants.

	Only Exh. image Mean (SD)	Only non-exh. image	Both exh & non-exh. images Mean (SD)	One distractor Mean (SD)
***Only*-f**	93.33 (17.29)	0	5.00 (13.77)	1.67 (6.34)
PreVf	65.00 (35.72)	0	34.17 (34.42)	0.83 (4.56)

Table 2. Distribution of choice of images as percentages (with SD)

IPs defined for Experiment 4 were identical to those defined for the Experiment 3 (see [8]). Figure 4 displays the PoF on the exhaustive image as a percentage of PoF on the exhaustive and non-exhaustive images as a function of time.

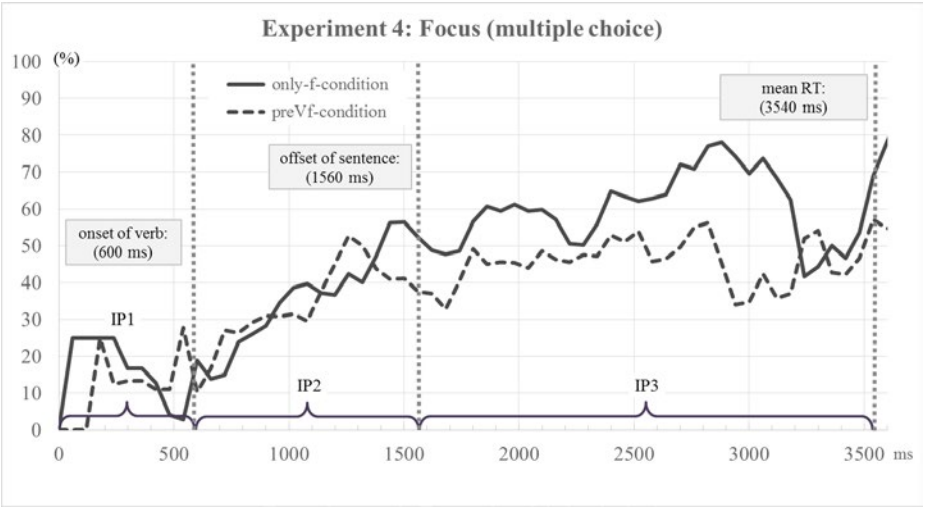


Figure 4. PoF on the exhaustive quadrant (%)

As Figure 4 shows PoF on the exhaustive image after the verb onset is proportionately distributed between the exhaustive and non-exhaustive images. Later, PoFs in the two conditions diverge, as PoF on the exhaustive image gradually increases in the *only*-condition, while it remains at chance level in the preVf-condition despite the fact that at the end of the trial the exhaustive image was chosen. A repeated measures factorial ANOVA (IP x Sentence Type) revealed a significant main effect of IP ($F(1.28) = 17.49$; $p < 0.001$) and of Sentence Type ($F(1.28) = 17.49$; $p = .001$). The interaction of the two variables was not significant.

6. Conclusion

The purpose of the current investigation was to examine the interpretational processes associated with the exhaustive interpretation of preVf. Our hypothesis was that the

exhaustive interpretation of preVf is tied to scalar implicature generation. The hypothesis was tested with three eye-tracking experiments and an online survey.

First we compared the interpretation of a minimal pair of sentences which had been shown to be interpreted via different processes. These were sentences containing NPs coordinated by *and* and *or*. Our results supported the view that the inclusive reading of *and* is tied to semantic, while the exclusive reading of *or* is tied to pragmatic (scalar implicature) interpretational processes. Explicit behavioral data showed that there were uniformly inclusive responses in the *and*-, and both inclusive and exclusive responses in the *or*-condition. Eye-tracking data also revealed a difference: PoF in the *or*-condition showed a greater hesitation than in the *and*-condition. Since these results are in line with the results of earlier investigations, they could provide a reliable basis for the comparison of *only*-f and preVf sentences.

In order to support the view that this comparison may lead to valid results we examined whether an important and theoretically motivated objection holds. According to this objection, the two structure types have different presuppositional and asserted content. If this is indeed the case, there may be a third, confounding factor in experiments that compare interpretational processes associated with the two structures and the results may be inconclusive. For this purpose we conducted an online survey in which participants were asked to choose one of two possible emphatic continuation phrases for the two structures. The two phrases were word order variations of each other. One of the phrases reflected on the presupposed content of *only*-f sentences used in the comparisons, while the other reflected on the presupposed content of preVf sentences. Both were paraphrases of the asserted content of their respective sentence type. We hypothesized that a consistent choice of completion phrases would reflect that people's intuition about how the presupposed and asserted content is divided in the two structures is consistent with current semantic analyses. Results revealed that people are consistent in the case of *only*-f sentences, but not in the case of preVf: the presupposed content of out of context preVf sentences is sometimes identified as those of *only*-f sentences. This rate is at chance level. Thus, the criticism may not be upheld: *only*-f sentences provide a reliable basis for comparison in the experimental investigation of the exhaustive interpretation of preVf.

In Experiment 3 we used the method of Experiment 1 to compare the interpretational processes associated with *only*-f and preVf sentences hypothesizing a similar difference in data patterns between *only*-f and preVf to those between NPs coordinated by *and* and *or*. Our hypothesis was not supported by the data: the interpretation of both sentence types was uniformly exhaustive, and the eye-tracking data did not correspond to the tendencies observed in the first experiment. These results may support the semantic accounts of preVf interpretation, but it is also possible that the uniformity of the data gathered in the two conditions was due to a forced choice task.

The purpose of Experiment 4 was to control for this effect. We repeated Experiment 3 with one modification: the number of images that could be chosen was not

restricted. The results of Experiment 4 were in line with results of earlier experimental works (see, e.g., Huang and Snedeker 2009) and matched the pattern obtained in Experiment 1. While a uniformly exhaustive interpretation was observed in the case of *only*-f sentences, the interpretations of preVf sentences were distributed between exhaustive and non-exhaustive readings. Eye-tracking data also revealed a significant difference similar to that associated with the processing of NPs coordinated with *and* and *or*: the data of preVf sentences (just as that of NPs coordinated with *or*) reflected a high degree of hesitation. Thus the results of Experiment 3 support our hypothesis that the exhaustive interpretation of preVf is associated with the generation of scalar implicatures.

The difference between the results of Experiments 3 and 4 suggests that although the exhaustivity of preVf is a pragmatic phenomenon (which as such is cancellable in certain contexts), the structure has a strong tendency to be associated with this reading. Consequently, although the traditional view suggesting a deterministic relationship between structure and interpretation may not be tenable, the intuition that preVf has a strong exhaustivity effect can be demonstrated.

Finally, the observation that the exhaustive interpretation of preVf is variable and that this variability is heavily context dependent can be taken as another strong case for its status as an implicature. To the best of our knowledge, the context dependence of preVf has not been demonstrated or even investigated in any previous experimental work so far. Thus, the direction of further research is given: we propose that in order to gain a deeper understanding of the interpretational processes related to preVf, research in this area could be extended to a principled investigation of the possible interpretations and functions of preVf and the interaction of these interpretations and potential contextual factors.

Works Cited

Bende-Farkas, Ágnes. 2009. “Adverbs of Quantification, *it*-Clefts and Hungarian Focus.” In *Adverbs and Adverbial Adjuncts at the Interfaces*, edited by Katalin É. Kiss, 317–48. Berlin, New York: Mouton de Gruyter.

Bergen, Leon, and Daniel Grodner. 2012. “Speaker Knowledge Influences the Comprehension of Pragmatic Inferences.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38 (5): 1450–60.

Büring, Daniel, and Manuel Kriz. 2013. “It’s That, and That’s It! Exhaustivity and Homogeneity Presuppositions in Clefts (and Definites).” *Semantics and Pragmatics* 6 (6): 1–29.

Chemla, Emmanuel, and Lewis Bott. 2013. “Processing Presuppositions: Dynamic Semantics vs Pragmatic Enrichment.” *Language and Cognitive Processes* 28 (3): 241–60.

É. Kiss, Katalin. 1998. “Identificational Focus versus Information Focus.” *Language* 74: 245–73.

É. Kiss, Katalin. 2004. *The Syntax of Hungarian*. Cambridge: Cambridge University Press.

Fekete, István, Mátyás Gerőcs, Anna Babarczy, and Balázs Surányi. 2014. "Logical and Pragmatic Meaning in the Interpretation of Connectives: Scalar Implicatures and Shallow Processing." In *Language Use and Linguistic Structure. Proceedings of the Olomouc Linguistics Colloquium,* edited by Joseph Emonds and Markéta Janebová, 171–80. Olomouc: Palacký University.

Gazdar, Gerald. 1979. *Pragmatics: Implicature, Presupposition and Logical Form*. New York: Academic Press.

Gerőcs, Mátyás, Anna Babarczy, and Balázs Surányi. 2014. "Exhaustivity in Focus: Experimental Evidence from Hungarian." In *Language Use and Linguistic Structure. Proceedings of the Olomouc Linguistics Colloquium,* edited by Joseph Emonds and Markéta Janebová, 181–94. Olomouc: Palacký University.

Geurts, Bart. 2016. "A Wish List for Experimental Pragmatics." In *Pre-proceedings of* "*Trends in Experimental Pragmatics*," edited by Fabianne Salfner and Uli Saurland. Berlin: Zentrum für Allgemeine Sprachwissenschaft.

Grice, H. Paul. 1975. "Logic and Conversation." In *Speech Acts*, vol. 3 of *Syntax and Semantics*, edited by Peter Cole and Jerry L. Morgan, 41–58. New York: Academic Press.

Grodner, Daniel J., Natalie M. Klein, Kathleen M. Carbary, and Michael K. Tanenhaus. 2010. "'Some,' and Possibly All, Scalar Inferences Are Not Delayed: Evidence for Immediate Pragmatic Enrichment." *Cognition* 116 (1), 42–55.

Horn, Laurence, R. 1969. "Presuppositional Analysis of 'Only' and 'Even.'" In *Papers from the Fifth Regional Meeting. Chicago Linguistic Society*, edited by Robert I. Binnick, Alice Davison, 98–107. Chicago, IL: Department of Linguistics, University of Chicago.

Horn, Laurence, R. 1972. "On the Semantic Properties of the Logical Operators in English." PhD diss, UCLA, Los Angeles, CA. IULC, Indiana University, Bloomington, IN.

Horváth, Julia. 2010. "'Discourse Features,' Syntactic Displacement, and the Status of Contrast." *Lingua* 120 (6): 1346–69.

Huang, Yi Ting, and Jesse Snedeker. 2009. "On-line Interpretation of Scalar Quantifiers: Insight into the Semantic-Pragmatics Interface." *Cognitive Psychology* 58 (3): 376–415.

Káldi, Tamás. 2015. "A magyar preverbális fókusz interpretációjának tulajdonságai egészséges és Broca-afáziás személyeknél." In *LingDok 14. Nyelvészdoktoranduszok dolgozatai*, edited by Zsuzsanna Gécseg, 105–24. Szeged: University of Szeged.

Kálmán, László, and András Kornai. 1989. "Hungarian Sentence Intonation." In *Autosegmental Studies on Pitch Accent*, edited by Harry van der Hulst and Norval Smith, 183–95. Dordrecht: De Gruyter Mouton.

Kas, Bence, and Ágnes Lukács. 2013. "Focus Sensitivity in Hungarian Adults and Children." *Acta Linguistica Hungarica* 60 (2): 217–45.

Onea, Edgar, and David Beaver. 2011. "Hungarian Focus Is Not Exhausted." In *Proceedings of the 19th Semantics and Linguistic Theory Conference: SALT 19*, edited by Satoshi Ito Cormany and David Lutz, 342–59. Ithaca: Cornell University.

Schlenker, Philippe. 2008. "Be Articulate: A Pragmatic Theory of Presupposition Projection." *Theoretical Linguistics* 34 (3): 157–212.

Sperber, Dan, and Deirdre Wilson. 1995. *Relevance: Communication and Cognition*. 2nd ed. Cambridge: Blackwell.

Szabolcsi, Anna. 1981. "The Semantics of Topic-Focus Articulation." In *Formal Methods in the Study of Language*, edited by J. A. G. Groenendijk, T. M. V. Janssen, and M. J. B. Stokhof. Amsterdam: Mathematisch Centrum.

Szabolcsi, Anna. 1994. "All Quantifiers Are Not Equal: The Case of Focus." *Acta Linguistica Hungarica* 42–43: 171–87.

Velleman, D., D. Beaver, E. Destruel, D. Bumford, E. Onea, and L. Coppock. 2012. "*It*-Clefts are IT (Inquiry Terminating) Constructions." *Semantics and Linguistic Theory* 22: 441–60.

Wedgwood, Daniel. 2003. "Predication and Information Structure: A Dynamic Account of Hungarian Pre-verbal Syntax." PhD diss., University of Edinburgh.

Gender, Number and Inflectional Class in Romance: Feminine/Plural *-a*

M. Rita Manzini[a] and Leonardo M. Savoia[b]

University of Florence, Italy

[a]rmanzini@unifi.it; [b]lsavoia@unifi.it

Abstract: We discuss the Romance nominal inflection *-a*, which surfaces both as a singular feminine exponent and as a lexicalization of "cohesive" plurals. The empirical focus is on Central Calabrian varieties, where *-a* plurals occur in the contexts of inflectional systems that do not differentiate masculine and feminine in the plural, as well as on Sursilvan Romansh varieties, where *-a* productively forms feminine singulars with an interpretation akin to that of *-a* plurals in Italian. On the basis of our case studies, we characterize the inflectional morphology in nouns as (sometimes) endowed with semantic content, specifically with the Class properties [aggregate] for mass and [⊆] for plural.

Keywords: gender; number; mass nouns; inflectional class; Romance

1. Introduction

In Italian, *-a* serves as a singular inflection, normally feminine, as in (1a); its plural is normally *-e*, as in (1b). The singular inflection *-o* is normally masculine, as in (1c), and its plural presents the plural inflection *-i*, as in (1d). However *-a* (apart from occurrences as masculine singular, not immediately relevant here) also introduces the plural of a set of nouns characterized by a distinctive semantics, denoting "a plurality of weakly differentiated parts" (Acquaviva 2008), as in (1e). The singular of these nouns is masculine and it sometimes displays a regular masculine plural with a pure count interpretation such as (1d). Romance languages have only two target genders, namely masculine and feminine—and the *-a* plural agrees in the feminine with determiners and adjectives in (1d). A comparison

can usefully be made with other language families that have genders, for instance the Semitic languages (Fassi Fehri 2016, Kramer 2015), which display the same syncretism between feminine singular and plural (non-gender specific), despite the fact that unrelated morphology is involved.

(1) (a) l-a cas-a bianc-a
the-F.SG house-F.SG white-F.SG
"the white house"

(b) l-e cas-e bianch-e
the-F.PL house-F.PL white-F.PL
"the white houses"

(c) il mur-o solid-o
the.M.SG wall-M.SG solid-M.SG
"the solid wall"

(d) i mur-i solid-i
the.M.PL wall-M.PL solid-M.PL
"the solid walls (e.g., of the house)"

(e) l-e mur-a solid-e
the-F.PL wall-PL(A) solid-F.PL
"the solid walls (e.g., of Rome)"

The potential theoretical interest of taking up the classical topic of the feminine/plural syncretism is that recent formal syntax and semantics studies revise the traditional distinctions between singular and plural, and between gender and number—yielding potential insights into their syncretism. First, underlying the standard number opposition singular/plural, there is an interpretive tripartition between mass nouns, count singulars and count plurals. More to the point mass singulars overlap in many respects quite closely with count plurals (Chierchia 2010); while in other respects the opposition is between count nouns, irrespective of number, and mass nouns (Borer [2005] for a syntactic model). So, it is expected that singular (at least mass singular) and plural may share a lexicalization.

Another stream of generative literature calls into question the traditional distinction between gender and number. The similarity between the genders of, say, Romance languages and the nominal classes of Bantu has been remarked more than once by the literature (Kihm 2005; Carstens 2008); genders and nominal classes are understood by this literature to be classification systems for nominal roots. Recall now that for Borer

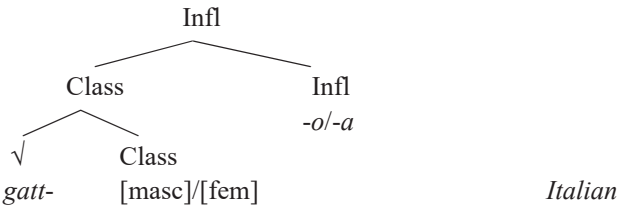
(2005) number (*qua* countability), as formally represented by her category Div, is also a classifier. In this perspective, gender and number (countability) are simply different facets of nominal classification (Déchaine et al. 2014). So, it is to be expected that the same exponent may lexicalize the apparently disjoint traditional categories of gender and number—conceived as superficial manifestations of nominal class. Indeed a well-known fact about Bantu nominal classes is that there are no specialized number morphemes; the same morphology forms the singular of one class, and the plural of another. The same holds for Italian *-a* and *-e*, which inflectional/gender markers of the singular and also form the plural (in different inflectional classes).

The empirical focus in this article will be on *-a* plurals in Central Calabrian varieties which lack gender distinctions in the plural—as well as on *-a* feminine singulars in Sursilvan Romansh varieties, with an interpretation similar to that of *-a* plurals in Italian.

Syntactically, this article is placed within the minimalist framework. Morphologically, we adopt a morpheme-based approach and we assume that the same basic computational mechanisms underlie syntax and morphology. The morphemic analysis of Indo-European nouns is fairly straightforward. The first component is a root; in consonance with Marantz (1997), we think of the root √ as category-less. Next to the root, a vocalic morpheme encodes properties that may include gender and/or number and/or declension class, depending on the language. A third slot is specialized for number and case (e.g., Latin) or just for number (e.g., Spanish). The consensus in the literature is that at least two functional projections are needed—corresponding roughly to gender and number. In homage to the cross-linguistic comparison with Bantu languages, the lower category is often labelled Class, the higher category is Num (Picallo 2008), i.e., [[√ Class] Num].

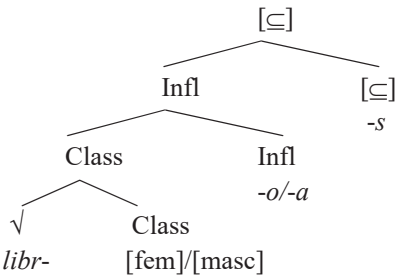
Extra complexity arises in Indo-European languages from the fact that there is no one-to-one mapping between the content of Class—which enters agreement with determiners and modifiers of N, and the inflections immediately following the root. The latter are instead sensitive to inflectional class, which we will henceforth call declension in order to avoid confusion with Class. As a first illustration of the structures that we will be using throughout, we exemplify Italian *gatt-o* "he-cat" and *gatt-a* "she-cat" in (2). In (2) the property "cat" is compatible with both a feminine and a masculine Class, depending on the sex denoted. We tentatively assign the inflectional vowel of Italian to an Infl Position—which embeds both the root and the Class node.

(2)



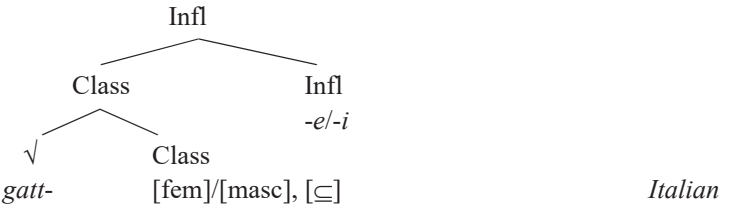
Languages like Spanish have an independent lexicalization for the plural, namely *-s*. Following Manzini and Savoia (2011) we formalize the content of the plural node as the category ⊆, which says that the denotatum of the predicate can be partitioned into subsets, as schematized for *libros*/*libras* "books/pounds" in (3).

(3)



In Italian however pluralization is obtained by a change of the inflectional vowel. In these terms we may suppose that the plural of *gatto*/*gatta* in (2), namely *gatti* "cats," *gatt-e* "she-cats" has the structure in (4), where the [⊆] property ("divisibility") is associated with the Class node.[1]

(4)



For reasons of space we cannot provide much further detail on the theoretical choices embodied by the structures in (2), (4), which we need in order to structure the data in later sections. We chose the category Class over the Distributed Morphology (DM) account of Class in terms of the nominalizing category *n* (Kihm 2005, Kramer 2015 a.o.)—essentially because it is less theory bound (Déchaine et al. 2014 use Asp). We also avoided the DM category Th, i.e., "thematic vowel" (Oltra-Massuet and Arregi

1 Lampitelli (2011) in a formal account of Italian, suggests that Italian has a structure similar to that of Spanish, though Gender and Number categories are lexicalized by elements, in the sense of Government Phonology. Specifically, the element -A lexicalizes the singular Number and -I the plural. Thus plural *-e* is the effect of the phonological combination of gender -A with number -I and so on. However the *-a* plurals of Italian (which are more productive than Lampitelli acknowledges) represent a problem for this approach—as are the *-e* plurals (of *-u* singulars) discussed by Loporcaro and Paciaroni (2011), Manzini and Savoia (forthcoming b).

2005), which is adjoined postsyntactically in order to externalize (via Late Insertion) the declension diacritics associated with the root. As we will see, our Infl, unlike Th, may be associated with interpreted content; in any event, in the present model we reject Late Insertion as unnecessarily costly (Manzini and Savoia 2005, 2011; Kayne 2010).

The main problem left open by the structures in (2), (4) has to do with the correct pairing of roots with their Class, when the latter is not semantically motivated, hence in practice with arbitrary gender—and with the correct pairing of [√, Class] substructures with their appropriate Infl.[2] Generative grammar has various mechanisms whereby the relevant matches could be implemented, including the standard syntactic mechanism of selection. Thus Kramer (2015, 54) explicitly endorses the view that gender she terms "arbitrary" is selected by the root; in present terminology this means that √ selects for Class when not determined by interpretive needs.

As for the Infl vowel, we just rejected the DM approach in terms of declension diacritics and Late Insertion of inflectional exponents. Rather, assuming that morphosyntax is projected from lexical terminals, as in the minimalist program, we adopt Kayne's (2010, 73–74) suggestion that inflectional vowels select (large sets of) roots. It is this selection that defines the descriptive notion of declension (not vice versa). In the same way nothing prevents us from assuming that it is really [fem] and [masc] Classes that select for (large sets of) root, when "arbitrary" (i.e., not corresponding to sexed interpretation). Our syntactic structures, with Class projecting, correspond in fact to this second option.

2. Central Calabrian *-a*

Our first case study concerns the Central Calabrian variety of Iacurso, which in the singular distinguishes the two genders [fem] and [masc] as well as the three inflectional classes *-a*, *-u*, *-ɛ*. At least *-ɛ* can combine with feminine or masculine bases. The plural has the gender-neutral realization *-i* on nouns, on adjectives and on functional categories of the noun, as illustrated in Fig. 1 just for definite determiners.

	-u, -i [masc]	-a, -i [fem]	-ɛ, -i [masc]	-ɛ, -i [fem]
sg	l-u fiɟɟuɐl-u "the son"	l-a rɔt-a "the wheel"	l-u mɛlun-ɛ "the melon"	l-a cav-ɛ "the key"
pl	l-i fiɟɟuɐl-i "the sons"	l-i ruɐt-i[3] "the wheels"	l-i mɛlun-i "the melons"	l-i cav-i "the keys"

Figure 1. Inflection classes of Iacurso (Central Calabria, Italy)

2 For a correct reading of the text it is necessary to remember that Class here refers to gender/number, as represented under the node Class.

3 In this dialect metaphony changes the stressed mid vowels /ɛ ɔ/ into the diphthongs [iɐ uɐ] in the context of a [+high] post-tonic vowel.

Iacurso also has *-a* plurals, illustrated in Fig. 2, for *-u* masculine singular bases; but while in Italian (1) *-a* plurals can be seen to switch to the feminine, in Iacurso in the absence of gender distinctions on adjectives and on functional categories of the noun no such switch is visible. In Iacurso, as in Italian, some Ns can be seen to alternate between the *-a* plural and the *-i* plural.

	-u [sg]	-i [pl]	-a [pl]
-u, -a	l-u jiðit-u “the finger”		l-i jiðit-a “the fingers”
-u, -i/-a	l-u kurtiɐʏ-u “the knife”	l-i kurtiɐʏ-i “the knives”	l-i kurtɛʏ-a “the knives (as a set)”

Figure 2. *-a* and *-i* plurals of Iacurso (Central Calabria, Italy)

As already mentioned, plural agreement on determiners and adjectives is systematically *-i*, independently of whether the singular is masculine or feminine (cf. Fig. 1), and whether the plural inflection is *-i* or *-a*, as further illustrated in (5)–(6). While (5) exemplifies the *-i* or *-a* plural of [masc] *-u* nouns, in (6) a [fem] *-a* noun is involved. Not only determiners, but also adjectives display the *-i* ending throughout.

(5) (a) du-i pum-a/ piɐrtsik-i matur-i
two apple-PL(A) peach-PL ripe-PL
“two ripe apples/peaches”

(b) kir-i dui kurtɛʏ-a/ kurtiɐʏ-i
that-PL two knife-PL(A) knife-PL
“those two knives”

(c) l-i jiðita/ diɐnt-i luɐŋg-i
the-PL finger-PL(A) tooth-PL long-PL
“the long fingers/teeth”

(d) l-i kurtɛʏ-a sunu lavat-i
the-PL knife-PL(A) are washed-PL
“The knives are washed.”

(6) (a) st-i buffiɐtt-i sunu luɐŋg-i
this-PL table-PL are long-PL
“These tables are long.”

(b) st-a buffɛtt-a ɛ llɔŋg-a
this-F.SG table-F.SG is long-F.SG
"This table is long."

Applying tests devised by Acquaviva (2008) we find that in partitive construction with a singular head of the type "one of . . . ," the gender of the noun on the numeral is determined by its singular form—regardless of whether an *-a* plural is involved. Hence, since *-a* plurals characterize nouns that in the singular are [masc], the ending on "one" in (7) is [masc] *-u* (cf. the first column of Fig. 1). We conclude that there is no evidence in Central Calabrian for the switch of gender that complicates the Italian picture in (1). In other words, we can study the alternation of two pure plurals, in *-i* and *-a*, in their simplest form.

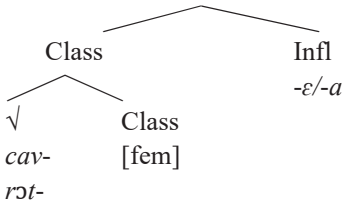
(7) (a) un-u dɛ kir-i ɔv-a
one-M.SG of that-PL egg-PL(A)
"one of those eggs"

(b) un-u dɛ kir-i lɛtt-a ɛ bbiɐcc-u
one-M.SG of that-PL bed-PL(A) is old-M.SG
"One of those beds is old."

Leaving aside for the moment the *-a* plural, the system illustrated in Fig. 1 can be accounted for on the basis of structures like Italian (2), (4), in which the root is associated with an Infl slot and a Class slot, as in (8a, b) for the singular and in (8c) for the plural. The Class slot can host three specifications, namely feminine, masculine and plural. We will say that in Calabrian, the *-i* Infl is associated with semantic content, namely [⊆], since it never appears but as a plural, along the lines of (8). The relation between the Class node and the Infl node in (8c) is one of agreement, specifically they agree with respect to the [⊆] property.[4]

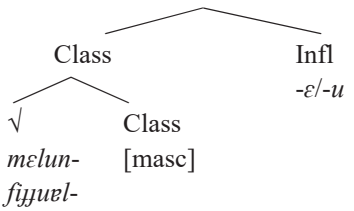
4 Technical issues arise concerning the exact operation of the rule of agreement, as pointed out by an anonymous reviewer. We assume that the [⊆] feature is independently introduced on both Class and Infl nodes and then the two features are matched under usual locality constraints—and interpreted as two occurrences of the same feature. This is unlike Chomsky's (2000) Agree, in that it does not present any interpretable/uninterpretable or valued/unvalued asymmetry—though it is identical to it in all other respects (Minimal Search, Identity, etc.).

(8) (a)



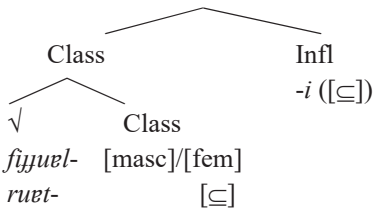
feminine

(b)



masculine

(c)



plural

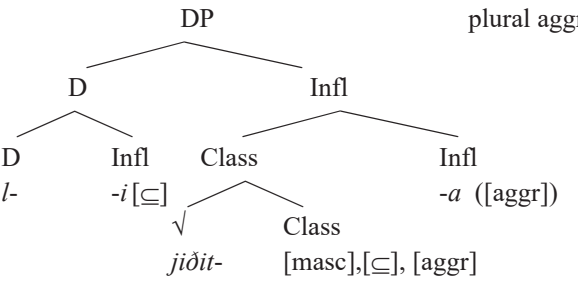
Acquaviva's (2008) semantic characterization of *-a* plurals as consisting of "weakly differentiated parts" appears to hold for Calabrian as well, witness the body part Ns present among *-a* plurals (*dinɔcc-a* "knees," *puddz-a* "wrists," *guvit-a* "elbows," *labbr-a* "lips," *jiðit-a* "fingers," *ɔss-a* "bones"). We take it that Acquaviva's characterization applies to body part Ns in Iacurso and to foodstuff with very much the same properties, such as *ɔv-a* "eggs," *pir-a* "pears," *pum-a* "apples." Other *-a* plurals attach to artifacts; in this respect, note that English also uses a collective singular for "cutlery" or speaks of a "knife set" (cf. *kurtɛʃ-a* "knives") and uses "fields" as a collective plural in "I went into the fields," etc. (cf. *ɔrt-a* "vegetable gardens"). Therefore the *-a* inflection corresponds to a set whose members are rather more like parts of whole than like individuated atoms. At the same time, of course, basic tests like the possibility of numeral quantifiers in (5a–b) or of partitive structures in (7) confirm that we are dealing with plurals.

The notion of an aggregate of parts is used by Chierchia (2010) to characterize mass denotation. Correspondingly, Manzini and Savoia (forthcoming a) use the feature [aggr] to characterize the content of the so-called neuter *de materia* (neuter with mass denotation) present in Central Italian varieties and which in effects configures

an agreement Class (together with [masc], [fem]) in at least some of them. Thus Manzini and Savoia eliminate the traditional class neuter (Loporcaro and Paciaroni [2011] for a recent approach) in favour of the class [aggr] (mass). At the same time their analysis is incompatible with Borer's (2005) idea that mass status depends on the mere absence of the Div category (see Kučerová and Moro (2011) on the Central Italian neuter). Rather mass has its own positively specified Class content, namely [aggr]—which is a conclusion consonant with that reached by Déchaine et al. (2014) for Bantu.

Assuming the existence of an [aggr] class in Romance or Indo-European, it is tempting to differentiate the *-a* plural from the *-i* plural by associating with the former the properties [⊆, aggregate]. This would yield structures of the type in (9) for *jiðita* "fingers." Note that despite having insisted on the non-availability of gender differences in the morphology of the Iacurso plural, we have nevertheless kept the [masculine] Class property in the representation in (9). This is because anaphoric material in the singular, e.g., "one of them" in (7) shows masculine gender. This confirms that *-a* plurals, though they happen to be connected to a gender change in Italian, have no necessary connection to it.

(9)



The structure of Class in (9) implies a very elementary ontology, consisting in the squaring of the two properties [⊆] [aggr]—each of which can be represented by specialized morphology in the natural languages we are considering. In fact, in consonance with the minimalist program, we hold that the syntax and the lexicon are relatively impoverished, albeit quick and efficient means, to restrict meaning, whose articulations are ultimately determined by contextual enrichment. Our claim therefore is syntactic, namely that [aggr]/zero crossed with [⊆]/zero is what is represented syntactically (in the type of languages we are considering).

Acquaviva (2008, 155–56) comments on "the dimness of some grammatical intuitions" going on to state that "the lack of individual distinctive properties is a matter of how the lexical predicates are conceptualized, and this often leads to variation among speakers and uncertain intuitions for one and the same speaker." This is consistent with what we are proposing here; rephrasing Acquaviva, the Iacurso speakers who indifferently

render Italian *coltell-i* "knives" with *kurtɛɽ-a* or *kurtiɐɽ-i* simply have two different ways of presenting the predicative content "knife"—namely as consisting of individuated atoms or as consisting of non-individuated atoms. In this sense the label proposed by Déchaine et al. (2014) for what we have called here Class, namely NAsp, seems particularly appropriate.

It remains for us to discuss our characterization of *-a* in (9) as endowed with the content [aggr].[5] Before doing so, however, we briefly discuss the proposal by Acquaviva (2008) that the *-a* plural is lexically fixed, at least in Italian. On the contrary, the approach that we have taken here is that *-a* plurals is a productive inflectional phenomenon (albeit a lexically conditioned one). Some support for our position come from diminutives. Simplifying a complex descriptive situation (Savoia et al. forthcoming), diminutives suffixes are transparent to the gender Class of the root, while at the same time determining their own choice of inflectional class; thus *-itʃiɐɽ-* in (10) is associated with the *-u* and *-a* inflectional classes in the [masc] and [fem] respectively. As expected on the basis of the inflectional classes of Fig. 1, diminutives ordinarily have an *-i* plural. However, if the content of the root (body part, foodstuff, etc.) warrants it, they can also have an *-a* plural, replicating the conditions described by Fig. 2. What shows that *-a* is not a "lexical plural" (not in Iacurso anyway) is that it can be associated with the diminutive of *pɛð-ɛ* in (10b), which cannot otherwise take the *-a* plural, like Nouns of the *-ɛ* inflectional class in general. In other words, our argument is that *-a* cannot be lexical because it may depend on derivational processes.

(10) (a) ɔss-itʃiɐɽ-u/ ɔss-itʃiɐɽ-i/ ɔss-itʃɛɽ-a
bone-DIM-M.SG bone-DIM-PL bone-DIM-PL(A)
"little bone/bones"

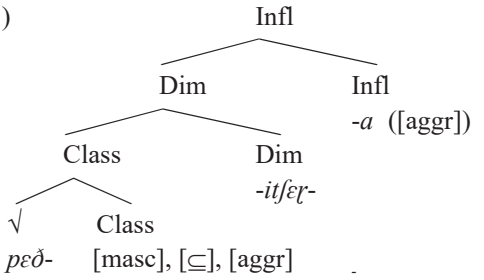
(b) dui pɛð-itʃiɐɽ-i/ pɛð-itʃɛɽ-a
two foot-DIM-PL foot-DIM-PL(A)
"two little feet"

Diminutives, and in general evaluative morphology, are beyond the scope of the present work. Nevertheless, following Savoia et al. (forthcoming), we will assume that diminutives have their own dedicated projections (Cinque 2015), immediately above Class.

5 Technical issues arise concerning the implementation of agreement, as pointed out by an anonymous reviewer. Recall that in fn. 3 we proposed that agreement identifies multiple occurrences of the same feature (under locality, etc.) so that it is interpreted just once. Evidently rather than identity, we must now invoke non-distinctness as the crucial property that allows agreement to take place, since the Infl in (9) picks up just the [aggr] Class feature.

Equivalently Class could be conceived as a field of (ordered) projections, including gender, size (diminutives, augmentatives), number (count/mass, but also singulative introduced in Romance as in other languages by diminutives), etc. As indicated in (11) the information concerning the gender selected by the root (specifically masculine) is preserved in the diminutive. A plural *-a* form can attach to the diminutive of Ns which admit of the [⊆, aggregate] mass plural, including body parts ("feet").

(11)



We are finally in a position to come back to the question that started the present investigation—concerning the syncretism of plural and gender in the *-a* ending. In the structure in (8c) we have embedded the assumption that the Infl element *-i* is associated with interpretive content, namely [⊆]. As discussed in the text, *-i* never turns up as nominal Infl except as a plural; this is made explicit in the lexical entry in (12a). In turn, from the point of view of number specifications, *-a* in (9), (11) is unambiguously associated with [aggr], so that we suggested that *-a* does in fact have an [aggr] content. Obviously, if we are to continue assuming that there is a single Infl item *-a* occurring in the (feminine) singular as well, [aggr] must be associated with *-a* only optionally, as in (12b). In the absence of other restrictions, we predict nevertheless that the property [aggr] may be present on *-a* in the singular as well. This is fairly trivially verified by the fact that the inflectional *-a* class will include mass nouns (e.g., *pɛtɣ-a* "stone"). When *-a* selects roots with individual content, like "wheel" in Fig. 1, it is not associated with the [aggr] content, because of its optionality.

(12) (a) *-i*: Infl, [⊆]
(b) *-a*: Infl, ([aggr])

Even with all the limitations noted, the lexical entry in (12b) provides an explanation of sorts for the syncretism of inflectional class and number morphology that we were seeking. Indeed, given the discussion that precedes, we can point to a positively specified property of *-a* that bridges between singular and plural namely [aggr]. In other words, it is in virtue of the property [aggr] that *-a* turns up both as a plural, and a singular inflectional

class marker. What escapes this analysis is the fact that *-a* happens to be feminine (at least by default). The latter is a matter to be learned by the child.[6]

3. Sursilvan *-a*

Sursilvan varieties of Romansh, such as that of Vattiz (Lumnezia Valley), differ from the Italian varieties considered in Sections 1–2 in that they have an *-s* plural, rather like Spanish in (3), which combines both with bare masculine bases, and with feminine bare or *-a* bases, i.e., [√, Class] ones, as summarized in Fig. 3. There are no other declensions in the relevant varieties, specifically not a masculine *-o* declension or a masculine/feminine *-e* declension.

	-∅, -s [masc]	-a, -as [fem]	-∅, -s [fem]
sg	iʎ meuŋ “the hand”	l-a rɔd-a “the wheel”	l-a nuʃ “the nut”
pl	iʎ-s meuŋ-s “the hands”	l-a-s rɔd-a-s “the wheels”	l-a-s nuʃ-s “the nuts”

Figure 3. Inflection classes of Vattiz (Lumnezia Valley, Switzerland)

The *-s* plurals have all of the relevant properties of count plurals, for instance that of being associated with numerical quantifiers (cf. the plural in Fig. 4). The same semantics that we have so far imputed to *-a* plurals appears to be associated in this language with singular *-a* forms, alternating with bare masculine bases, as in Fig. 4.[7] The similarity with the *-a* plurals of, say, Central Calabrian is confirmed by the fact that the *-a* singular of Vattiz, applies to the same roots, including notably body parts that come in a “cohesive” set (*bratʃ-a* “arm set,” *det-a* “finger set”) or “weakly differentiated” individuals such as foodstuffs (*mail-a* “apple set,” *per-a* “pear set”). Indeed the natural translation for these expressions in a language like English (or Italian) is a plural.

6 It is worth noting that the classical historical account of Indo-European feminine singular/neuter plural -a (see the summary in Clackson 2007, 107) is that a neuter/collective plural -a was extended to a new inflectional class for collective/abstract singulars—which only secondarily came to coincide with the default class for feminine animates. Viewed as a projection on the historical, external axis of an analysis motivated on internal grounds, this reconstruction appears to be quite compatible with the present discussion.

7 The changes undergone by the base *mail-* are phonological; thus [il] palatalizes to [ʎ], while the sequence [ls] is realized as [lts].

	[masc]	*-a* [fem]
sg	in maʎ “an apple”	l-a mail -a “the apples” (as a generic, etc.)
pl	du-s mailt-s “two apples”	

Figure 4. Count/aggregate alternations in Vattiz (Lumnezia Valley, Switzerland)

Needless to say, *-s* plurals agree in plurality with determiners and adjectives and also trigger plural agreement on the verb (the copula *ain*), as (13). The singular status of *-a*, even in alternations of the type in Fig. 4, is confirmed by agreement with the verb in (14), namely by the presence of the 3rd singular form of the copula *ai*; agreement with determiners and adjectives is in the feminine singular.

(13) (a) iʎ-s kɔrn-s ain liung-s
the-PL horn-PL are long-PL
“The horns are long.”

(b) kwɛl-s mailt-s ain marʃ-s
that-PL apple-PL are rotten-PL
“Those apples are rotten.”

(14) (a) si-a bratʃ-a ai kwərt-a
his-F.SG arm-F.SG is short-F.SG
“His arms are short.”

(b) l-a dɛt-a ai liuŋg-a
the-F.SG arm-F.SG is long-F.SG
“The fingers are long.”

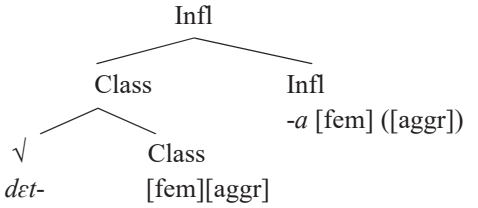
(c) l-a mail-a ai marʃ-a/ kurdad-a
the-F.SG apple-F.SG is rotten-F.SG fallen-F.SG
“The apples are rotten/have fallen.”

(d) kɔnt-a mail-a
how.much-F.SG apple-F.SG
“How many apples?”

According to Chierchia (2010), a mass singular is a plurality of sorts, namely a whole made up of parts. Thus a singular mass noun is like a plural count noun in that both include a multiplicity—namely a multiplicity of individuals, or a multiplicity of parts. In this perspective, we

are not surprised that the Romance *-a* morphology can turn up denoting both a "cohesive" plural of "weakly differentiated" parts, as in Italian (1) or Central Calabrian in Section 2, and a mass/collective singular, as in Sursilvan (14). Specifically, Sursilvan *-a* introduces a collective interpretation in combination with a subset of roots available to be interpreted as an aggregate of similar individuals, namely the same roots (body parts, foodstuff, etc.) that trigger *-a* plurals in Italian varieties. Suppose we characterize the mass/collective singular of Sursilvan as [aggr], stressing the continuity with the Italian *-a* plural. The difference between the two is simply that the Sursilvan *-a* [aggr] forms are singular, in other words no [⊆] properties are involved, so that the structure for *dɛta* "finger set" is as in (15).

(15)



The lack of plural [⊆] properties in (15) correctly predicts agreement in the singular between phrases such as *la dɛta* and the verb (cf. [14a–c]). On the other hand, *dɛta* in (15) triggers an *-a* agreement with its D; the same is true of Quantifiers, cf. (14d), and of adjectival and participial modifiers and predicates, cf. (14a–c). In (15) we have modelled this agreement as involving both [fem] gender and [aggr] properties. Thus the structure in (15) prospects the existence of a Class [aggr] property that enters into *-a* agreement. When the *-a* Infl is not [aggr], furthermore, it is univocally associated with [fem]. Indeed, as far as we can tell, the *-a* masculine class of standard Italian, Spanish, etc. (e.g., Italian *poet-a* "poet," *poet-i* "poets") is not found in Sursilvan (cf. *poet* "poet," *poet-s* "poets"). This suggests a relatively rich lexical entry for the Infl element *-a*, as an exponent of [fem] and optionally [aggr] in (16).

(16) *-a*: Infl, [fem], ([aggr])

The only other inflectional element present in Fig. 3–4 is the *-s* plural.[8] The natural conclusion to draw is that it is identified with the pure [⊆] morpheme and node, as in (17), yielding plural structures essentially like Spanish (3), cf. (18) below.

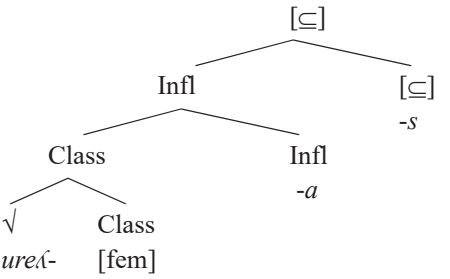
8 Perfect participles in the masculine plural present an *-i* Infl element, which evidently parallels Italian *-i*. This is exemplified in (i) from the variety of Vella (Lumnezia Valley).

(i) il-s tʃɔp-s lava-i
the-PL jacket-PL washed-M.PL
"the washed jackets"

(17) *-s*: [⊆]

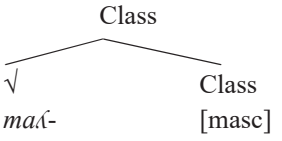
Nothing prevents the *-a* morphology in (16) and the *-s* morphology in (17) from combining if *-a* is simply [fem]. This is true also for root predicates referring to body parts, foodstuffs, etc., that might otherwise be compatible with the alternation in Fig. 4—as indicated in (18) for *ureʎ -a*(*-s*) "ear(s)." At the same time, [aggr] *-a* cannot combine with *-s*, despite the fact that the combination of features [aggr], [⊆] has been used in Section 2 to account for Central Calabrian. The obvious difference between the structure of Sursilvan in (18) and that of Central Calabrian in (9) has to do with the different position of the [⊆] content—and we may provisionally associate the incompatitibility of [⊆] with [aggr] to the position that [⊆] takes.

(18)



There is another difference between the [⊆] morphology of Sursilvan and that of Spanish. As indicated in Fig. 3, the Sursilvan masculine singular generally coincides with the bare lexical base; in present terms this means that Romance languages admit not only three-tiered structures like Sursilvan (18) and two-tiered structures like Italian (2), (4)—but also simple trees like consisting of a root and its class specification, as in (18') for Sursilvan *maʎ* "apple."

(18')



Despite the situation depicted in (19), a masculine singular *-s* ending survives in predicative contexts (Haiman and Benincà 1992, Manzini and Savoia 2005, 2012). In (19) we provide relevant examples from another variety of the Lumnezia Valley, Vella. The *-s* singular inflection characterizes both predicative adjectives (19a–c) and participles (19d). The subject of the predication can be animate, as in (19a–b) or inanimate, as in (19c–c') including mass nouns, e.g., "milk" in (19c'). Furthermore, case is irrelevant—witness the adjective in (19e), which is predicated of a direct object.

(19) (a) kwai om ai kwərt-s/ grɔnd-s
that man is short-M.SG tall-M.SG
"That man is short/tall."

(b) el ai meʎer-s ke jɛu
he is better-M.SG than I
"He is better than me."

(c) kwai rakwənt ai ver-s
that story is true-M.SG
"That story is true."

(c') kwai/ iʎ latʃ ai buŋ-s
that/the milk is good
"That/the milk is good."

(d) iʎ afɔn ai ɲiu-s
the boy is come-M.SG
"The boy has come."

(e) jɛu vai viu el kuntent-s/ grɔnd-s
I have seen him happy-M.SG tall-M.SG
"I saw him happy/tall."

Singular predicative adjectives/participles do not bear *-s* endings in contexts where the argument they are predicated of has a propositional content, as in (20a–b) and/or when the subject is an expletive with a DP/CP correlate, as in (20b–c). Perfect participles of transitive/unergative verbs, which do not agree with the subject but have an invariable ("default") inflection, also do not bear singular *-s*. Finally, the *-s* inflection, while possible, is not necessary with quantificational subjects, including wh-phrases in (21a), negatives and existentials in (21b).

(20) (a) kwai ai ver
that is true
"That is true."

(b) i(ʎ) ai meʎer da klama tai
It is better to call you
"It is better to call you."

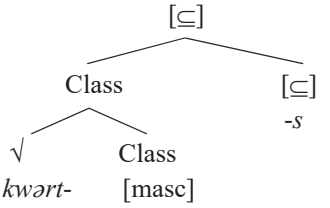
(c) i(ʌ) ai ɲiu afɔn-s
it is come child-PL
"There have come children."

(21) (a) tʃi ai ɲiu-s
the boy come-M.SG
"Who has come?"

(b) nidʒin/ tsitʃi ai ɲiu(-s)
nobody somebody is come(-M.SG)
"Nobody/somebody has come."

The characterization of plural as [⊆], seen in (17), is compatible with Borer's (2005) conclusion that plurality predicates divisibility of any given root property. In Borer's conception, Div does not attach specifically to plurals, but equally characterizes count singulars. It is therefore tempting to hypothesize that Sursilvan *-s* externalizes the [⊆]/Div property in predicative contexts, independently of singular or plural number, as schematized in (22) for *kwərts* "short-M.SG." For present purposes, we can simply restrict singular *-s* to [masc] bases by stipulation.

(22)



In their discussion of Sursilvan, Manzini and Savoia (2012) propose that the *-s* morpheme is quantificational in nature and is found only in environments where a quantificational closure is not provided by determiners and quantifiers of the noun. This is too strong in present terms, since [⊆] simply predicates divisibility of the root content, as we just saw. However, we may retain the conclusion that predicative adjectives must have a richer structure than modifier adjectives (embedded under the functional layers of the noun). Specifically, we propose that both plurals and singulars must be embedded under the morphology [⊆], introducing individuating properties, in order for predication to take place. This behavior is subject to parametrization—in other Romance languages predicative and modifier adjectives are treated alike, in German predicative adjectives are inflectionally poorer than categories within the DP (Haiman and Benincà 1992).

If the structure in (22) is to hold, on the other hand, the *-s* adjective cannot agree with the DP of which it is predicated with respect to the [⊆] feature, given that in (19c') a mass noun

is the subject of the adjectival predication. This state of affairs reminds us of linker structures, as seen for instance in Balkan languages, where (predicative) adjectives are obligatorily preceded by an article (Lekakou and Szendroi 2012, Franco et al. 2015 for recent accounts). As shown in the Albanian example in (22), though phi-features agreement holds, the preadjectival linker is represented by the definite article, even if the subject of the predication is indefinite. In other words, we propose that embedding under the individuating layer in (22) has the same formal role as embedding under the linkers layers in (23).[9]

(23)	dial-i/	një	dialë	i	bukur
	boy-the.M.SG	a.M.SG.	boy-M.SG	the.M.SG	nice
	"the/a nice boy"				

The fact that *-s* is excluded in (20) remains to be accounted for. In fairly traditional terms, one may say that in (20) predicative adjectives and participles agree in the neuter gender with the expletive *i(ʎ)*, which differs morphologically from both masculine and feminine singular, i.e., *El*/*Ela* (cf. [19b, e]); the same would also be true of the demonstrative in (20a). However, this does not quite account for the sensitivity of agreement to the presence of indefinite subjects, as in (21). The latter belong to the class of elements (wh-phrases, Negative Polarity Items, existentials) standardly modelled as free variables (Heim 1982) (closed by existential quantification, negation, etc.). Furthermore, Manzini and Savoia (2012) propose that expletive pronouns are variables, which are assigned a value at the C-I interface via predication—namely equated to the DP or sentential correlate in postverbal position. Thus we may want to substitute the generalization that the *-s* predicative singular inflection is incompatible with the neuter with a different generalization, namely that it is incompatible with free variable subjects.

Summarizing, Sursilvan varieties provide evidence for [fem], [aggr] Class externalized by *-a*. In addition the language has an *-s* ending specialized for the [⊆] content. The latter has a straightforward plural reading that combines with both [masc] and [fem]. Less straightforwardly *-s* has a singular masculine reading on predicative adjectives/participles. We have proposed that it is the same *-s* which forms the plural, hence the [⊆] content is independent of pluralization.

4. Conclusions

In Central Calabrian, the *-a* plural is a mass plural, resulting from the combination of the plural content [⊆] of and the [aggr] content. As for the *-a* Infl element, it has optional

9 An anonymous reviewer objects that by hypothesis [⊆] has an interpretable content, while linkers are ordinarily assumed to be uninterpretable. One could however argue that they are interpreted, namely as resumptive clitics are, as Ds bound by higher Ds/Qs (Franco et al. 2015, and see also Lekakou and Szendroi 2012).

[aggr] content. Vice versa the *-i* Infl specializes for [⊆] content. In short, Central Calabrian has [masc] and [fem] gender Classes—and what would be traditionally called a neuter plural, here understood to correspond to an [⊆, aggr] class.

In turn, Sursilvan provides evidence for a [fem], [aggr] Class externalized by *-a*. In addition the language has an *-s* ending specialized for the [⊆] content. The latter has a straightforward plural reading that combines with both [masc] and [fem], but not with [aggr]. Interestingly *-s* also has a singular masculine reading on predicative adjectives/participles. We have proposed that it is the same *-s* which forms the plural, hence the [⊆] content must be capable of expressing the count content independently of pluralization.

Works Cited

Acquaviva, Paolo. 2008. *Lexical Plurals*. Oxford: Oxford University Press.

Borer, Hagit. 2005. *In Name Only,* vol. 1 of *Structuring Sense.* Oxford: Oxford University Press.

Carstens, Vicki. 2008. "DP in Bantu and Romance." In *The Bantu-Romance Connection*, edited by Cécile De Cat and Katherine Demuth, 131–66. Amsterdam: John Benjamins.

Chierchia, Gennaro. 2010. "Mass Nouns, Vagueness and Semantic Variation." *Synthese* 174: 99–149.

Chomsky, Noam. 2000. "Minimalist Inquiries: The Framework." In *Step by Step, Essays on Minimalist Syntax in Honor of Howard Lasnik*, edited by Roger Martin, David Michaels, and Juan Uriagereka, 89–155. Cambridge, MA: MIT Press.

Cinque, Guglielmo. 2015. "Augmentative, Pejorative, Diminutive and Endearing Heads in the Extended Nominal Projection." In *Structures, Strategies and Beyond*, edited by Elisa Di Domenico, Cornelia Hamann, and Simona Matteini, 67–81. Amsterdam: John Benjamins.

Clackson, James. 2007. *Indo-European Linguistics*. Cambridge: Cambridge University Press.

Déchaine, Rose-Marie, Raphaël Girard, Calisto Mudzingwa, and Martina Wiltschko. 2014. "The Internal Syntax of Shona Class Prefixes." *Language Sciences* 43: 18–46.

Fassi Fehri, Abdelkader. 2016. "Semantic Gender Diversity and Its Architecture in the Grammar of Arabic." *Brill's Journal of Afroasiatic Languages and Linguistics* 8: 154–99.

Franco, Ludovico, M. Rita Manzini, and Leonardo M. Savoia. 2015. "Linkers and Agreement." *The Linguistic Review* 32: 277–332.

Haiman, John, and Paola Benincà. 1992. *The Rhaeto-Romance Languages*. London: Routledge.

Heim, Irene. 1982. "The Semantics of Definite and Indefinite Noun Phrases." PhD diss., University of Massachusetts at Amherst, Amherst, MA.

Kayne, Richard. 2010. *Comparisons and Contrasts*. New York: Oxford University Press.

Kihm, Alain. 2005. "Noun Class, Gender, and the Lexicon/Syntax/Morphology Interfaces: A Comparative Study of Niger-Congo and Romance languages." In *The Oxford Handbook of Comparative Syntax*, edited by Guglielmo Cinque and Richard Kayne, 459–512. Oxford: Oxford University Press.

Kramer, Ruth. 2015. *The Morphosyntax of Gender*. Oxford: Oxford University Press.

Kučerová, Ivona, and Anna Moro. 2011. "On Mass Nouns in Romance: Semantic Markedness and Structural Underspecification." In *Proceedings of the 2011 Annual Conference of the Canadian Linguistic Association.* http://homes.chass.utoronto.ca/~cla-acl/actes2011/actes2011.html.

Lampitelli, Nicola. 2011. "Forme phonologique, exposants morphologiques et structures nominales: étude comparée de l'italien, du bosnien et du somali." PhD diss., Université de Paris 7.

Lekakou, Marika, and Kriszta Szendrői. 2012. "Polydefinites in Greek: Ellipsis, Close Apposition and Expletive Determiners." *Journal of Linguistics* 48: 107–49.

Loporcaro, Michele, and Tania Paciaroni. 2011. "Four Gender-Systems in Indo-European." *Folia Linguistica* 45: 389–433.

Manzini, M. Rita, and Leonardo M. Savoia. 2005. *I dialetti italiani e romanci. Morfosintassi generativa*. Alessandria: Edizioni dell'Orso.

Manzini, M. Rita, and Leonardo M. Savoia. 2011. *Grammatical Categories*. Cambridge: Cambridge University Press.

Manzini, M. Rita, and Leonardo M. Savoia. 2012. "(Definite) Denotation and Case in Romance. History and Variation." In *Romance Languages and Linguistic Theory 2009: Selected Papers from "Going Romance" Nice 2009*, edited by Janine Berns, Haike Jacobs, and Tobias Scheer, 149–66. Amsterdam: John Benjamins.

Manzini, M. Rita, and Leonardo M. Savoia. Forthcoming a. "N Morphology and Its Interpretation: The Neuter in Italian and Albanian Varieties." In *Constraints on Language Structure* (*Proceedings of the LingBaw Conference, Lublin 2015*). Peter Lang.

Manzini, M. Rita, and Leonardo M. Savoia. Forthcoming b. "N Morphology and Its Interpretation: Neuter *-o* and Plural *-a* in Italian Varieties." In *Studies in Honor of Andrea Calabrese*, edited by Pietro Cerrone, Harry van der Hulst, and Roberto Petrosino. Berlin: de Gruyter.

Marantz, Alec. 1997. "No Escape from Syntax: Don't Try Morphological Analysis in the Privacy of Your Own Lexicon." *University of Pennsylvania Working Papers in Linguistics* 4: 201–25.

Oltra-Massuet, Isabel, and Karlos Arregi, 2005. "Stress-by-Structure in Spanish." *Linguistic Inquiry* 36: 43–84.

Picallo, Carme. 2008. "Gender and Number in Romance." *Lingue e linguaggio* VII: 47–66.

Savoia, Leonardo M., M. Rita Manzini, Ludovico Franco, and Benedetta Baldi. Forthcoming. "A Morpho-Syntactic Analysis of Evaluatives in Italian." *Studi Italiani di Linguistica Teorica ed Applicata*.

Locatives, Part and Whole in Uralic

Ludovico Franco,[a] Giulia Bellucci,[b] Lena Dal Pozzo,[c] and M. Rita Manzini[d]

[a]Nova University of Lisbon, Portugal; [b, c, d]University of Florence, Italy; [c]Pontifical Catholic University of Rio de Janeiro, Brazil

[a]franco.ludovico@gmail.com, [b]giulia.bellucci@unifi.it, [c]lena.dalpozzo@gmail.com, [d]mariarita.manzini@unifi.it

Abstract: In this paper we provide a characterization of the adpositional domain in Finnish, taking into account some comparative evidence from the Uralic family. We show that accounts put forth in the recent theoretical literature are not empirically adequate and we will provide a novel solution, assuming that inner case morphemes (both the genitive and the *-l*, *-s* morphemes in Finnish) are best characterized as part-whole/zonal inclusion relator (following Manzini and Savoia [2011] and Manzini and Franco [2016], among others), while what are traditionally labelled as adpositions in Finnish (and elsewhere in Uralic) are best characterized as Axial Parts (in the sense of Svenonius [2006] and following literature). We will leave the proper characterization of the outer case inflections for future research.

Keywords: Locative Cases; Adpositions; Axial Parts; Part-Whole, Uralic

1. Introduction

We aim at providing a novel characterization of the adpositional domain in Finnish, taking into account some comparative evidence from the Uralic family. We will show that recent theoretical accounts on the topic are not empirically adequate and we will provide a novel solution, assuming that inner case morphemes (both the genitive and the *-l*, *-s* morphemes in Finnish) instantiate a part-whole/zonal inclusion relator (Manzini and Savoia [2011] and Manzini and Franco [2016], among others). At the same time, we assume that what are traditionally labelled as adpositions in Finnish (and elsewhere in the Uralic Family) are best characterized as Axial Parts (Svenonius 2006). We will

leave the proper characterization of the outer case inflections of Finnish, as illustrated in Section 2, for future research.

1.1 The (Spatial) Adpositional Domain: Axial Part

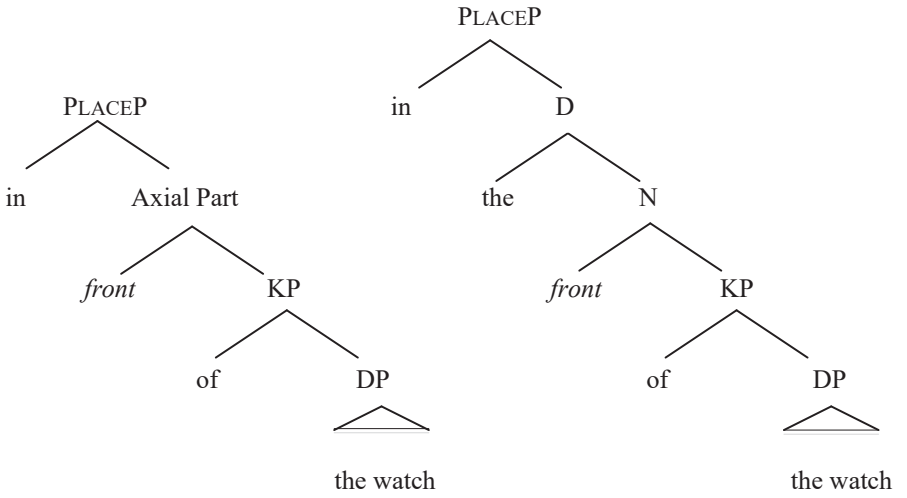
According to Svenonius (2006), the syntactic category Axial Part (AxP) identifies a part of the Ground (the reference landmark for the location), which can be taken as a spatial axis to locate the Figure (the object whose location is at issue; cf. Talmy [2000]). AxP has a mixed behaviour, sharing some properties with nouns. Indeed, often AxP is homophonous with Relation Nouns (Rel N) denoting body parts (Roy 2006) or other nominal items with spatial relevance.

Svenonius (2006), based on a set of diagnostics (e.g., AxPs contra homophonous Rel Ns commonly do not have articles, do not pluralize, do not take modifiers, can be specified by a measure phrase, etc.) argues against the idea that AxPs (items like *front*, *beside*, *behind* and so on), are a subclass of nouns, precisely Rel Ns (as opposed to Sortal nouns, e.g., *a child of someone* vs. **a person of someone*) (Hagège 2010, 162–65; Barker 1995).

It has been argued that AxP can be seen as an independent category, which is in between nouns and prepositions (Pantcheva [2011], Fábregas [2007], Roy and Svenonius [2009], Cinque [2010], and Franco [2016], among many others). In Svenonius (2006), the AxP projects a functional layer which is immediately dominated by a locative preposition (Place) and is above the DP that introduces the Ground. Svenonius uses the descriptive label K(ase) to indicate the item linking the AxP to the Ground. Svenonius assumes what Borer (2005) calls "Neo-constructivism," namely the working hypothesis that some dimensions of meaning are shaped by syntactic structure, while other dimensions come directly from the lexical content of the item introduced into the syntactic module.

The AxP *vs*. RelN dichotomy is a case in point. An item like English *front* is polysemous. Inserted under an N node in a syntactic derivation, *front* will express a noun, combining with plural morphology, determiners, etc. Inserted under an AxP node it will be part of the functional skeleton of the extended projection of P. Basically, Rel N and AxP enter different syntactic configurations, which following the representation given in Svenonius (2006) are illustrated in (1).

(1) (a) AxPart > *front* (b) RelNoun > (*the*) *front*



1.2 The Spatial Adposition Domain: PathP/PlaceP

The PP domain above AxP is assumed to be quite rich and articulated in the recent theoretical literature (Koopman [2000], Den Dikken [2010] on Dutch, Holmberg [2002] on Zina Kotoko, and Svenonius [2003; 2007] crosslinguistically), comprising at least what can be labelled PLACE (associated with stative locational meanings) and what is commonly called PATH (associated with directed motion) (cf. also Pantcheva [2011] and Romeu [2013] for more "layered" approaches).

Place elements give information on the relation between the Figure and the Ground (which is the reference landmark for the location of the Figure, as we have seen above). This is illustrated in (2a), where *the elephants* are the Figure and *the boat* is the Ground. On the contrary, Paths would provide information about a trajectory; Path elements may specify whether a Place is a Goal (2b) or a Source (2c), and may also give information on the orientation of a trajectory (2d) (cf. Pantcheva [2011]; examples are taken from Svenonius [2010, 127]).

(2) (a) The elephants remained in the boat.
(b) They cast a wistful glance to the shore.
(c) The boat drifted further from the beach.
(d) Their ears sank down several notches.

Path is commonly assumed to be higher than Place, as illustrated in (3) and (4), i.e., when they co-occur the Path layer can be assumed to be structurally more distant from

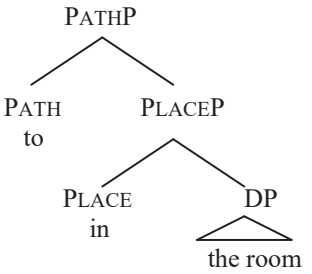
the Ground (cf. Baker [1988]; cf. also Koopman [2000] and Den Dikken [2010] for evidence from Dutch).

(3) na to gma
on to table
"onto the table" (Zina Kotoko, Holmberg [2002])

(4) sew-re-l-di
bear-erg-on-to
"onto the bear" (Lezgian, Haspelmath [1993])

In (5) below we provide a rough representation of a structure involving the Place and Path nodes. Please consider that AxP would be generated in a node below Place.

(5) "into the room"



Given this basic background, our aim is:

(a) to illustrate (and possibly solve) some puzzles regarding Finnish locatives;
(b) to show that oblique cases (K), as well as prepositions, are elementary predicates, endowed with a part-whole content, along the lines of Manzini and Franco (2016), Franco and Manzini (2017), based on cross-linguistic evidence from the Uralic family;
(c) to show that there's nothing special with locatives, namely UG does not provide us with a locative syntax. The same syntax is at work in non-spatial domains.

2. Finnish Locative Puzzles

In this section we will illustrate the most notable characteristics of the Finnish locative system and we will provide an overview of some previous theoretical accounts on the topic.

2.1 The *-l*/*-s* Series

Finnish has six productive morphological cases expressing location in a concrete sense (see Siro [1964, 29] for a full picture, including "relics" of non-productive locative

cases, which will not be taken into consideration in the present paper; cf. also Hakulinen [1979], Häkkinen [1996], and Itkonen [1997], among others). The Finnish spatial system is illustrated in (6) with the noun *talo* "house."

The spatial case morphemes of Finnish can be seen as compositional; on historical grounds, the illative -(*h*)*Vn* is related to -*s* though a phonological process occurred by which *s* gradually weakened into *h* and then disappeared in most instances (Huumo and Ojutkangas 2006; Lehtinen 2007). Comrie (1999) suggests that Finnish has two "series markers": -*s* "in" and -*l* "on." These series markers combine with the Locative endings to express states, motion-*to* meanings and motion-*from* meanings.

(6)

Series	"in/at"	"to"	"from"
-*s* (**internal**)	-*ssa* (*talo-ssa*)	-(*h*)*Vn* (*talo-on*)	-*sta* (*talo-sta*)
	inessive	**illative**	**elative**
-*l* (**external**)	-*lla* (*talo-lla*)	-*lle* (*talo-lle*)	-*lta* (*talo-lta*)
	adessive	**allative**	**ablative**

Spatial case system of Finnish (Pantcheva 2011; cf. Sulkala and Karjalainen 1992)

In Finnish the main spatial role of the so-called internal cases (-*s* series) is to designate containment, where one entity is situated within (or moves into or out of) a (usually three-dimensional) space.

(7) Tyttö on kirjastossa
girl be.PRS.3SG library.INE
"The girl is in the library."

(8) Tyttö meni kirjastoon
girl go.PST.3SG library.ILL
"The girl went to the library."

(9) Tyttö tuli kirjastosta
girl come.PST.3SG library.ELA
"The girl came from the library."

The external cases (-*l* series) designate a relation of "association," "contact," etc.

(10) Tyttö on kirjastolla
girl be.PST.3SG library.ADE
"The girl is at the library."

(11) Tyttö meni kirjastolle
girl go.PST.3SG library.ALL
"The girl went to the library."

(12) Tyttö tuli kirjastolta
girl come.PST.3SG library.ABL
"The girl came from the library."

2.2 Finnish Adpositions

Finnish makes extensive use of adpositions to express local relations. These adpositions are linked to the ground by a genitive (13) or partitive (14) case morpheme. The adposition is inflected by the spatial case series reviewed in (6). The partitive tends to encode complements of prepositions, as in (14), while genitives encode complements of postpositions, as in (13) though this does not represent a fixed behavior. Some postpositions take both series of spatial case (*-s*/*-l*) (e.g., *ede-* "above"), others take a series only (e.g., *perä-* "behind" > *-s* series).

(13) Järvi talo-n lähe-llä
lake house-GEN near-ADE
"The lake near the house."

(14) Järvi lähe-llä talo-a
lake near-ADE house-PART
"The lake near the house."

2.2.1 Previous Accounts: Pantcheva (2011)

According to Pantcheva (2011), the series markers *-l* and *-s* lexicalize the AxP head; the spatial case(s) above may lexicalize the Place head or the Path head, which she decomposes into Goal and Source. The Illative morpheme *-(h)Vn*, which does not overtly display *-s*, lexicalizes AxP, Place, and Path (Goal). Note that working within the nanosyntactic framework (Starke 2009; Caha 2009), she assumes that a single morpheme can lexicalize a series of syntactic nodes.

The model she proposes seems untenable on empirical grounds. In Finnish (and in the whole Uralic family), adpositions seem to instantiate prototypical AxP. In particular, they are often derived from body terms as in (15) (Suutari [2006], for full discussion) and they are linked to the Ground by means of a K device (in Finnish genitive and partitive case). Consider the examples reported above in (13)–(14). Adpositions clearly lexicalize the AxP node but the morphemes *-s*/*-l* are still present. What do they lexicalize in such cases?

There is a kind of "overabundance" issue here, if we assume that the two morphemes are both AxPs competing for the same slot. Namely, we would expect *-l* and *-s* to

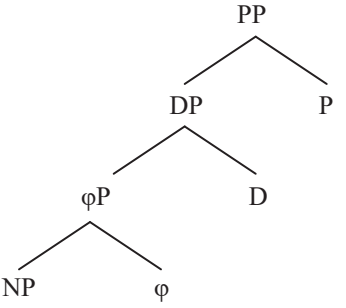
disappear whenever an axial adposition is present. The fact that this is not the case might lead to the conclusion that *-l* and *-s* lexicalize something else.

2.2.2 Previous Accounts: Asbury (2008)

Asbury (2008) assumes that spatial cases represent an outer layer (that she labels P) of the extended projection (Grimshaw 1990) of the noun phrase. She does not directly take into consideration the *-l* and *-s* morphemes, assuming that *-ltA*, *-stA* and so on are interpreted as a single unit. Namely, she does not assume that *-l*, *-s* host a projection on their own. This could be a problem once we assume that meaningful items must project their own phrase (and *-l*/*-s* conceptualize exterior/interior in Finnish).

However what seems to be the main problem with her account is that she assumes that the inner layer of the structure in (1) (i.e., genitive/partitive in Finnish) is occupied by determiner-like morphemes. Following Kayne (1994) on English *of*, she assumes that the Finnish genitive/partitive is a morpheme expressing definiteness (D). In (15) we provide a rough representation of her structure.

(15)



According to our own perspective, there is clear evidence that the morpheme *-n* is not a D item in Finnish. Just consider the "quirky" contexts in (16), (17) and (18) where the meaning associated with the genitive inflection *-n* is ergative-like (Franco and Reeve, in preparation; cf. Woolford 2005).[1] No Definiteness effects may be ascribed to the appearance of the genitive in such positions. Hence, we see no reason to link the genitive to D (analogous observations could be made for the partitive, which encodes aspect-like imperfective content; cf. Kiparsky [2001] and Kracht [2002]).

1 The standard characterization of the genitive as a structural case in Finnish (cf. Vainikka 1989; Kiparsky 2001) may be questioned on the basis of the same empirical evidence. Nevertheless, the issue is orthogonal to the topics introduced in this work and will not be addressed any further in what follows.

(16) Sinun kannattaa yrittää
you.GEN be.worth.PRS.3SG try.INF
"It's worth for you to try." *experiencer*

(17) Karin on lähdettävä
Kari.GEN be.PRS.3SG go.PTCL.PRS
"Kari has to go." *necessity construction*

(17') Sinun olisi hyvä soittaa huomenna
you.GEN be.COND.PRS.3SG good call.INF tomorrow
"It would be good if you would call tomorrow." *necessity construction*

(18) Kakku on äidin leipoma
cake be.PRS.3SG mom.GEN baked.PTCL.AGENT
"The cake was baked by the mom." *agentive participle*

Rather, as pointed out by Fábregas (2007) the relation between the Ground and the AxP is a part-whole/possessive one, namely a relation in which the Ground is the possessor of (an Axis) and the AxP is the possessum (of the Ground).

Further empirical evidence that a characterization in terms of definiteness of the genitive morpheme is inadequate is the fact that the genitive can also appear on determiners in Colloquial Finnish (Matthew Reeve, pers. comm.), as in (19).[2]

(19) sen koiran
the.GEN dog.GEN
"Of his dog."

2.2.3 Previous Accounts: Svenonius (2012)

Svenonius (2012), building on his (2006) work introduced in Section 1.1, argues that adpositions are AxPs, the series *-l*/*-s* "interior/exterior" lexicalize Place, while the spatial cases on top of them represent Path (cf. Section 1.2). The illative (which in any case, as we have already pointed out, is historically related to the *-s* series) would be a suppletive form, acting as a portmanteau for Path+Place.

In Svenonius's account, we see a kind of conceptual clash (Pantcheva [2011], which possibly, for this reason, analyses *-l*/*-s* as AxP). Take the inessive in (7), repeated below in (20) for ease of reference. How can *-sA* represent a Path? It seems here that both *-s* and *-sA* lexicalize Place (there are no conceptual hints of Path, neither Goal or Source).

2 In Finnish the demonstrative *se* is turning into a definite article (Laury 1997; Alexiadou et al. 2007).

(20) Tyttö on kirjastossa
girl be.PRS.3SG library.INE
"The girl is in the library."

Thus, the model seems to predict a wrong cartography, namely two distinct positions for morphemes expressing close (or the same) semantics (similar arguments can be provided for the *-l* series).

3. Finnish: Spatial Cases beyond Space

Before presenting our own account, we must note that in Finnish spatial cases are employed to express a range of non-spatial meaning, starting from possession. This fact can be seen as a cognitive tendency to express non-local relation by means of locative morphemes (Jackendoff 1983, Freeze 1992, Den Dikken 1995, and Kracht 2002, among others).

Nevertheless it could also point to the reverse, namely that a possession/inclusion relation could be transferred into the spatial domain (Manzini and Franco 2016; Franco and Manzini 2017). As shown in (21) and (22), both the external and internal cases are used to indicate possession, including concrete physical possession. Possession *stricto sensu* is not the only non-spatial relation expressible by the spatial cases of Finnish. The "instrumental" behaviour of the adessive in (23) is an example.

(21) Tytöllä on kirja
girl.ADE be.PRS.3SG book
"The girl has a/the book."

(22) Hän on flunsassa
s/he be.PRS.3SG flu.INE
"The girl has the flu."

(22') Talossa on iso ovi
house.INE be.PRS.3SG big door
"The house has a big door."

(23) Piirsin tämän lyijykynällä
draw.PRS.1SG this.GEN/ACC pencil.ADE
"I draw this with the pencil."

Furthermore the adessive case that we have seen is employed for instrumentals in (23) can be also used to encode the causee in causative constructions, as illustrated in (24).

(24) Keisari rakennutti orjilla temppelin
emperor.NOM make.build.3SG.PST slave.ADE.PL temple.GEN
"The emperor made the slaves build a temple."

Thus, a conceptualization in terms of PATH > PLACE and related hierarchy does not immediately explain the non-spatial occurrences of Finnish spatial cases.

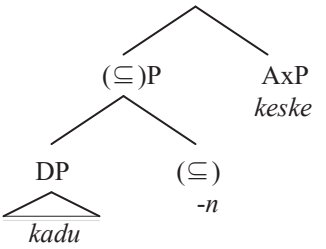
4. Our Proposal

Given the data we have presented so far, and given the problems outlined above for some contemporary approaches to the syntax of Finnish locatives, we try to advance an alternative proposal.

Consider an example including an AxP, such as (25) (we follow Svenonius in assuming that the adposition is an AxP). We assume that the Ground-complement is the possessor of the AxP. Following Manzini and Savoia (2011), Manzini and Franco (2016), we notate the genitive (and partitive) relating the possessor (Ground) and the possessum (AxP) as (⊆).The conceptual core of the relation between Grounds and AxP is that of part-whole.[3] A rough representation is in (26).

(25) Auto kadun keskellä
car.NOM street.GEN centre.ADE
"The car in the middle of the street."

(26)



There is a strict parallelism between the syntactic configuration/relation involving Axial Part and Ground and the proper Part-Whole relation, universally. Natural languages consistently employ the same strategy to lexically encode a Part-Whole and an Axial Part-Ground relation. Uralic data support this view, as shown in (27) and (28), for Skolt Saami and Nenets, respectively.

3 Following Belvin and Den Dikken (1997), we construe possessors as "zonally including" the possessee.

(27) to'b **ǩie′đj** **vue′lnn** leäi jõn pä'htträi'ğğ
there stone.GEN under be.PST.3SG big rock.hole
[koozz sää'm liâ piijjâm
[REL.SG.ILL Saami.PL be.PST.3SG put.PST.PTCP
kåå'dd **vuei'vid** čue'rveezvui'm]
reindeer.SG.GEN head.PL.ACC antler.PL.COM.3PL]
"There under the stone was a big hole in the rock, where the Saami used to put the heads of reindeer, with their antlers." *Skolt Saami* (Feist 2010, 351)

(28) (a) **wen'ako-h** **xawoda** lebtə°-q *Part-Whole*
dog-GEN ears.3SG hang-3PL
"Dog's ears are hanging."

(b) **tol°-h** **ŋil°na** kniga-q ŋaq *Figure-Ground*
table-GEN under book-PL be.3PL
"There are books under the table." *Nenets* (Nikolaeva 2014, 59, 181)

In Table 1, we report the result of our survey on Uralic languages showing the main devices by which Uralic languages encode the two relations (sometimes two different strategies may show up, cf. Finnish genitives alternating with partitives in encoding the relation between AxP and Ground). Notably, proper part-whole and AxP-Ground relations are expressed consistently by the same device.

Uralic Varieties	**Part-Whole**	**AxPart-Ground**
Eastern Khanty (Filchenko 2007)	POSSESSIVE INFL.	POSSESSIVE INFL.
Enets (Künnap 1999)	GENITIVE	GENITIVE
Erzya Mordvin (Van Pareren 2013)	GENITIVE	GENITIVE
Estonian (Viitso 1998)	GENITIVE	GENITIVE
Finnish (native knowledge)	GENITIVE	GENITIVE
Hungarian (Kenesei et al. 1998)	POSSESSIVE INFL.	POSSESSIVE INFL.
Kamassian (Simoncsics 1998)	GENITIVE	GENITIVE
Komi (Avril 2006)	JUXTAPOSITION	JUXTAPOSITION
Moksha Mordvin (Van Pareren 2013)	GENITIVE	GENITIVE
Nganasan (Chumakina 2011)	GENITIVE	GENITIVE
Ostyak (Nikolaeva 1999)	POSSESSIVE INFL.	POSSESSIVE INFL.
Skolt Saami (Feist 2010)	GENITIVE	GENITIVE
Tundra Nenets (Nikolaeva 2014)	GENITIVE	GENITIVE
Udmurt (Winkler 2001)	JUXTAPOSITION	JUXTAPOSITION
Vogul (Riese 2001)	JUXTAPOSITION	JUXTAPOSITION
Votic (Ariste 1997)	GENITIVE	GENITIVE

Table 1. Encoding Part-Whole and Axial Part relations in Uralic

We take the (⊆) part/relation to be very wide-ranging, potentially encompassing partitives, inalienable and alienable possession, light verbs, and also the notion of location—which is in competition with it as the true primitive underlying possession (Freeze 1992; Den Dikken 1995 vs. Levinson 2011; Franco and Manzini 2017). In other words, we assume that in natural languages, a locative may be construed as a specialization of the part-whole relation,[4] roughly:

(29) "x included by y, y location."

Therefore we propose that the series *-s* and *-l* are a locative specialization of this zonal inclusion (⊆). We may consider *-s* to be containment and *-l* to be contact/vicinity. In any event both morphemes contribute to the lexicalization of a (⊆) node. Similar considerations can be advanced for more familiar languages, for instance Italian (Romance) *a* vs. *in*. Consider the pair in (30). (30a) means that the sea defines a vicinity including me (the Italian dative preposition *a* hence matches the *-l* series of Finnish); on the contrary, (30b) says that I'm properly contained by the sea.

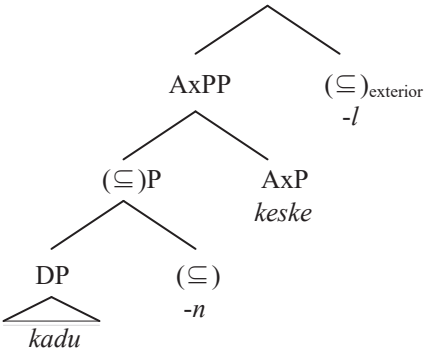
(30) (a) Sono **al** mare
"I'm *at* the sea."

(b) Sono **in** mare
"I'm *in* the sea."

To reiterate, we assume that both *-l* and *-s* (as well as Italian *a* and *in*) express different flavours of the (⊆) relation. The representation of the (growing) tree is, thus, as in (31).

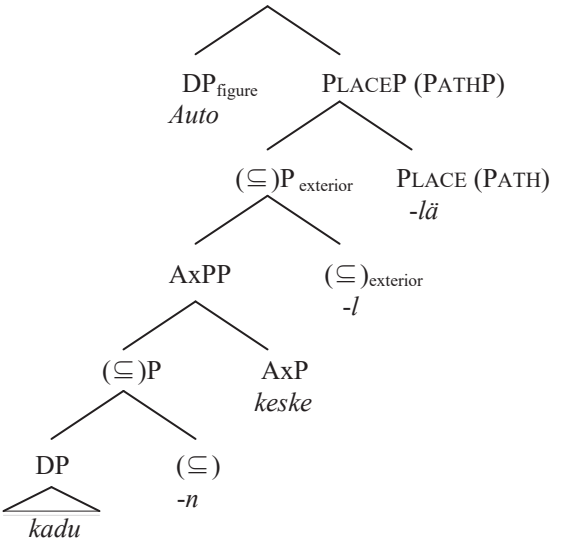
4 Manzini and Savoia (2011) argue in favour of the primitive nature of the part-whole relation on the basis of considerations regarding the morphological shape of Indo-European languages. Thus inflections alone suffice for the lexicalization of the more elementary possession/part-whole relation in languages where even the simplest of locative relations require the lexicalization of (complex) prepositions. Specifically, in discussing the syncretic lexicalizations of dative and locative in Albanian, Manzini and Savoia construe locative as a specialization of the part-whole relation, where different locatives introduce different locative restrictions on inclusion. This is compatible with the expression of (certain types of) possession as locations, for instance alienable possession in Palestinian Arabic, according to Boneh and Sichel (2010).

(31)



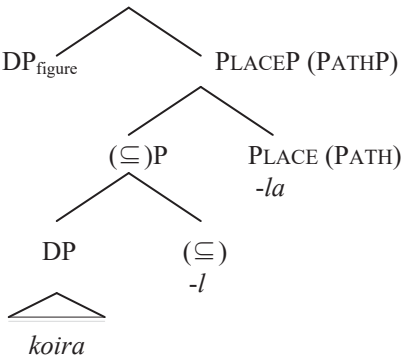
We now assume that the (⊆), specifically contact-inclusion/exterior (⊆) in (31) projects and merges with the "outer" inflection (*-lä*, *-ta*, etc.), that we provisionally notate as Place/Path in a way strictly consonant to their content (state, motion either towards or from) yielding the complete structure illustrated in (32), headed by the Figure DP (or eventually by an Event Figure; cf. Wood [2015]).

(32)



The same configuration is preserved when the suffixes *-s-sA*, *-l-lA* and so on appear on nouns not denoting locations, as in *koira-l-la* (dog-adessive) seen in example (21) above, for which we provide a representation in (33).

(33)



4.1 Back to AxP and Its Morpho-lexical Properties

The structure we propose for the AxP in (32) strongly implies that AxPs in Finnish are real relational nouns. The fact that Finnish adpositions are case inflected may suffice to determine this status, as case normally attaches to nouns. In many languages, AxPs have been shown to be indistinguishable from Rel N (failing Svenonius' discriminating tests). Johns and Thurgood (2011) provide evidence in this regard from Inuktitut and Uzbeki (cf. also Franco [2016] on the diachrony of Italian).

Below we provide some evidence from Uralic, focussing on Udmurt as recently described by Arkhangelskiy and Usacheva (2015). In Udmurt, axial postpositions are case inflected, just as in Finnish. They freely pluralize, depending on the plurality of either by the Ground (34) or of the Figure (35).

(34) Škafjos puš-jos-en kopo uka-śk-e
cupboard.PL inside-PL-LOC dust gather-PRS-3SG
"Dust is gathering inside cupboards."

(35) Milam d′erevna koter-jos-en lud′-jos
we.GEN village around-PL-LOC field-PL
"All fields around our village."

Inflected adpositions in Udmurt can appear in argumental position, namely inflected by core cases.

(36) Škaf puš-se miśk-ono
cupboard inside-POSS.3SG(ACC) wash-DEB
"The inside of the cupboard has to be washed."

Finally, possessive marking in Udmurt is possible on the dependent or on the head (Nichols 1986). Inflected adpositions pattern with ordinary nouns, as shown in (37)–(38).

(37) (a) Mə korka kośag-a-m šu-iśk-i-z zərgələ
I.GEN house window-ILL-POSS.1SG hit-DETR-PST-3SG sparrow
"A sparrow bumped into the window of my house."

(b) Mə korka-je-len kośag-a-z
I.GEN house-POSS.1SG-GEN window-ILL-POSS.3SG
šu-iśk-i-z zərgələ
hit-DETR-PST-3SG sparrow
"A sparrow bumped into the window of my house."

(38) So təb-i-z korka-je dor.e /
he go.up-PST-3SG house-POSS.1SG near.ILL
korka dor-am uža-nə
house near-ILL-POSS.1SG work-INF
"He went up to my house to work."

5. General Discussion and Conclusions

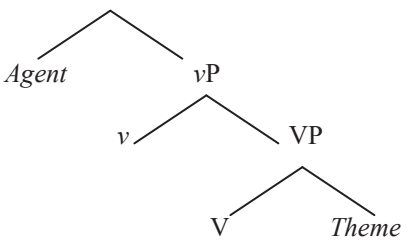
To summarize, we impute an interpretive content to the item which links Axial Parts and Grounds, which Svenonius (2006) descriptively characterizes as K (case). This content is predicative (not the D-like content of Asbury 2008), and it can be realized by prepositions (Italian, English), or by nominal inflections (Uralic). The inflectional realization of the (⊆) predicate is conventionally called case. But in present terms, case is definable at most as the crossing of the more elementary notions of atomic predicate and inflectional realization. As originally argued in Fillmore (1968), we see no other differences between (oblique) cases and adpositions (Manzini and Franco [2016] and Franco and Manzini [2017] for extended discussion and arguments against a post-syntactic approach to obliques).

Every account of natural language must address the proximity of dative/genitive/instrumental and locative specifications, corresponding to frequent syncretic lexicalizations: e.g., the instrumental/adessive in Finnish; the ergative/oblique = inessive in Caucasian languages (Comrie and Polinsky 1998), the genitive = inessive of Ossetic (an Iranian variety in contact with Caucasian languages; Kulikov [2009]). Possession is often identified with a location in the literature, in particular Freeze (1992), Lyons (1967). We reverse this perspective imputing a broad part-whole content to locatives. A locative may be construed as a specialization of the part-whole relation, roughly "x included by y, y location." The (⊆) content may correspond to several inflectional cases or adpositions. For instance, according to Manzini and Savoia (2011), the languages

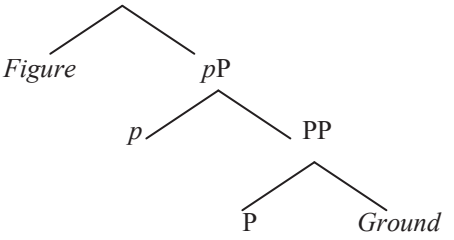
where dative is lexically different from genitive (including English *of* and *to*, Italian *di* "of" and *a* "to," etc.) display contextual sensitivity in the realization of the (⊆) category, which is externalized as dative "to" when attached to sentential projections, while it is externalized as genitive "of" when it is attached to nominal categories. Interestingly, the relation between AxPs and Grounds is instantiated in Italian by both *a* (e.g., *sotto al fiume* "below to-the river") and *di* (*sotto di te* "below of you"). This may be related to a sensitivity to the animacy hierarchy (i.e., *di* is the obligatory choice here with pronouns) (Fábregas 2015a, b). Uralic presents similar lexical variation in the realization of (⊆) (partitive, genitive, internal/external inclusion).

Svenonius (2003, 2007; cf. Franco 2016, Wood 2015) assumes that the figure has properties reminiscent of external arguments, while the ground has properties reminiscent of internal arguments. According to Svenonius (2003; 2007) figures are introduced by a functional head p, in a way that is analogous to the introduction of external arguments by v, along the lines of Hale and Keyser (1993) and Chomsky (1995); see the trees in (39). Here we simply do not take a position on this point, despite endorsing approaches that make a strict parallel between P and V (cf. also Van Riemsdijk 1978, Emonds 1985, and Demirdache and Uribe-Etxebarria 2000).

(39) (a)



(b)



What is most relevant for present purposes is that there is nothing special with space in syntax (roughly following Levinson [2011]).

Acknowledgements

We thank Matthew Reeve and an anonymous reviewer for her comments and suggestions. Ludovico Franco gratefully acknowledges the Portuguese National Science Foundation, Fundação para a Ciência e a Tecnologia (FCT), for supporting this work with the research grant IF/00846/2013.

Works Cited

Alexiadou, Artemis, Liliane Haegeman, and Melita Stavrou. 2007. *Noun Phrase in the Generative Perspective.* Berlin: Mouton De Gruyter.

Ariste, Paul. 1997. *A Grammar of the Votic Language.* The Hague: Mouton. First published 1968 by Indiana University Press.

Arkhangelskiy, Timofey, and Maria Usacheva. 2015. "Syntactic and Morphosyntactic Properties of Postpositional Phrases in Beserman Udmurt as Part-of-Speech Criteria." *SKY Journal of Linguistics* 28: 103–37.

Asbury, Anna. 2008. "The Morphosyntax of Case and Adpositions." PhD diss., Univesiteit Utrecht.

Avril, Yves. 2006. *Parlons Komi.* Paris: L'Harmattan.

Baker, Mark C. 1988. *Incorporation: A Theory of Grammatical Function Changing*. Chicago, IL: University of Chicago Press.

Barker, Chris. 1995. *Possessive Descriptions*. Stanford: CSLI Publications.

Belvin, Robert, and Marcel den Dikken. 1997. "*There*, happens, *to*, *be*, *have*." *Lingua* 101: 151–83.

Boneh, Nora, and Ivy Sichel. 2010. "Deconstructing Possession." *Natural Language and Linguistic Theory* 28: 1–40.

Borer, Hagit. 2005. *In Name Only,* vol. 1 of *Structuring Sense.* Oxford: Oxford University Press.

Caha, Pavel. 2009. "The Nanosyntax of Case." PhD diss., University of Tromsø.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chumakina, Marina. 2011. "Nominal Periphrasis. A Canonical Approach." *Studies in Language* 35: 247–74.

Cinque, Guglielmo. 2010. "Mapping Spatial PPs: An Introduction." In *Mapping Spatial PPs*, vol. 6 of *The Cartography of Syntactic Structures*, edited by Guglielmo Cinque and Luigi Rizzi, 3–25. New York: Oxford University Press.

Comrie, Bernard, and Maria Polinsky. 1998. "The Great Daghestanian Case Hoax." In *Case, Typology and Grammar: In Honor of Barry J. Blake*, edited by Anna Siewierska and Jae Jung Song, 95–114. Amsterdam: John Benjamins.

Comrie, Bernard. 1999. "Spatial Cases in Daghestanian Languages." *Sprachtypologie und Universalienforschung* 52: 108–17.

Dikken, Marcel den. 1995. *Particles*. New York: Oxford University Press.

Dikken, Marcel den. 2010. "On the Functional Structure of Locative and Directional PPs." In *Mapping Spatial PPs*, vol. 6 of *The Cartography of Syntactic Structures*, edited by Guglielmo Cinque and Luigi Rizzi, 74–126. New York: Oxford University Press.

Demirdache, Hamida, and Myriam Uribe-Etxebarria. 2000. "The Primitives of Temporal Relations." In *Step by Step. Essays on Minimalist Syntax in Honor of Howard Lasnik*, edited by Roger Martin, David Michaels, and Juan Uriagereka, 157–86. Cambridge, MA: MIT Press.

Emonds, Joseph. 1985. *A Unified Theory of Syntactic Categories*. Dordrecht: Foris.

Fábregas, Antonio. 2007. "(Axial) Parts and Wholes." *Nordlyd* 34: 1–32.

Fábregas, Antonio. 2015a. "Una nota sobre locativos y acusativos." *Archivum* LXV: 57–74.

Fábregas, Antonio. 2015b. "Direccionales con *con* y Marcado Diferencial de Objeto." *Revue Romane* 50: 163–90.

Feist, Timothy R. 2010. "A Grammar of Skolt Saami." PhD diss., University of Manchester.

Filchenko, Andrey Y. 2007. "A Grammar of Eastern Khanty." PhD diss., Rice University.

Fillmore, Charles J. 1968. "The Case for Case." In *Universals in Linguistic Theory*, edited by Emmon Bach and Robert T. Harms, 1–88. New York: Holt, Rinehart, and Winston.

Franco, Ludovico. 2016. "Axial Parts, Phi-Features and Degrammaticalization." *Transactions of the Philological Society* 114: 149–70.

Franco, Ludovico, and M. Rita Manzini. 2017. "Instrumental Prepositions and Case: Contexts of Occurrence and Alternations with Datives. *Glossa: A Journal of General Linguistics* 2(1): 8.1–37.

Franco, Ludovico, and Matthew Reeve. In preparation. "The Genitive/Ergative Connection: Evidence from Finnish." Ms. Lisboa.

Freeze, Ray. 1992. "Existentials and Other Locatives." *Language* 68: 553–95.

Grimshaw, Jane. 1990. "Extended Projection." Ms., Brandeis University.

Hagège, Claude. 2010. *Adpositions*. Oxford: Oxford University Press.

Häkkinen, Kaisa. 1996. *Suomalaisten esihistoria kielitieteen valossa*. *Tietolipas* 147. Helsinki: SKS.

Hakulinen, Lauri. 1979. *Suomen kielen rakenne ja kehitys*. 4th revised ed. Otavan korkeakoulukirjasto. Helsinki: Otava.

Hale, Kenneth, and S. Jay Keyser. 1993. "On Argument Structure and the Lexical Expression of Grammatical Relations." In *The View From Building* 20, edited by Kenneth Hale and S. Jay Keyser, 53–109. Cambridge, MA: MIT Press.

Haspelmath, Martin. 1993. *A Grammar of Lezgian*. Berlin: Mouton de Gruyter.

Holmberg, Anders. 2002. "Prepositions and PPs in Zina Kotoko." In *Some Aspects of the Grammar of Zina Kotoko*, edited by Bodil Kappel Schmidt, David Odden, and Anders Holmberg, 162–74. Munich: Lincom Europa.

Huumo, Tuomas, and Krista Ojutkangas. 2006. "An Introduction to Finnish Spatial Relations: Local Cases and Adpositions." In *Grammar from the Human Perspective: Case, Space and Person in Finnish*, edited by Marja-Liisa Helasvuo and Lyle Campbell, 11–20. Amsterdam: John Benjamins.

Itkonen, Terho. 1997. "Reflections on PreUralic and the 'Saami-Finnic Protolanguage'." *Finnisch-Ugrische Forschungen* 54: 229–66.

Jackendoff, Ray. 1983. *Semantics and Cognition*. Cambridge, MA: MIT Press.

Johns, Alana, and Brigid Thurgood. 2011. "Axial Parts in Inuktitut and Uzbeki." *Proceedings of the 2011 Annual Conference of the Canadian Linguistic Association*, edited by Lisa Armstrong. http://homes.chass.utoronto.ca/~cla-acl/actes2011/Johns_and_Thurgood_2011.pdf.

Kayne, Richard. 1994. *The Antisymmetry of Syntax*. Cambridge, MA: MIT Press.

Kenesei, István, Robert M. Vago, and Anna Fenyvesi. 1998. *Hungarian*. London: Routledge.

Kiparsky, Paul. 2001. "Structural Case in Finnish." *Lingua* 111: 315–76.

Koopman, Hilda. 2000. "Prepositions, Postpositions, Circumpositions, and Particles." In *The Syntax of Specifiers and Heads*, edited by Hilda Koopman, 204–60. London: Routledge.

Kracht, Marcus. 2002. "On the Semantics of Locatives." *Linguistics and Philosophy* 25: 157–232.

Kulikov, Leonid. 2009. "Evolution of Case Systems." In *The Oxford Handbook of Case*, edited by Andrej Malchukov and Andrew Spencer, 439–57. Oxford: Oxford University Press.

Künnap, Ago. 1999. *Enets*. München: Lincom Europa.

Laury, Ritva. 1997. *Demonstratives in Interaction: The Emergence of a Definite Article in Finnish.* Amsterdam: John Benjamins.

Lehtinen, Tapani. 2007. *Kielen vuosituhannet*. Helsinki: SKS.

Levinson, Lisa. 2011. "Possessive *with* in Germanic: *Have* and the Role of P." *Syntax* 14: 355–93.

Lyons, John. 1967. "A Note on Possessive, Existential and Locative Sentences." *Foundations of Language* 3: 390–96.

Manzini, M. Rita, and Ludovico Franco. 2016. "Goal and DOM Datives." *Natural Language and Linguistic Theory* 34: 197–240.

Manzini, M. Rita, and Leonardo M. Savoia. 2011. "Reducing 'Case' to Denotational Primitives: Nominal Inflections in Albanian." *Linguistic Variation* 11: 76–120.

Nichols, Johanna. 1986. "Head-Marking and Dependent-Marking Grammar." *Language* 62: 56–119.

Nikolaeva, Irina. 1999. *Ostyak*. München: Lincom Europa.

Nikolaeva, Irina. 2014. *A Grammar of Tundra Nenets*. Berlin: De Gruyter.

Pantcheva, Marina. 2011. “Decomposing Path: The Nanosyntax of Directional Expressions.” PhD diss., University of Tromsø.

Pareren, Remco van. 2013. “Body Part Terms as a Semantic Basis for Grammaticalization: A Mordvin Case Study into Spatial Reference and Beyond.” *Language Sciences* 36: 90–102.

Riemsdijk, Henk van. 1978. *A Case Study in Syntactic Markedness: The Binding Nature of Prepositional Phrases.* Lisse: The Peter de Ridder Press.

Riese, Timothy. 2001. *Vogul*. München: Lincom Europa.

Romeu, Juan. 2013. “The Nanosyntax of Path.” Ms. Madrid.

Roy, Isabelle. 2006. “Body Part Nouns in Expressions of Location in French.” *Nordlyd* 33: 98–119.

Roy, Isabelle, and Peter Svenonius. 2009. “Complex Prepositions.” In *Autour de la préposition, Actes du Colloque International de Caen (20–22 septembre 2007)*, edited by Francois Jacques, Eric Gilbert, Claude Guimier, and Maxi Krause, 105–16. Caen: Presses Universitaires de Caen.

Simoncsics, Peter. 1998. “Kamassian.” In *The Uralic Languages*, edited by Daniel Abondolo, 580–601. London: Routledge.

Siro, Paavo. 1964. *Suomen kielen lauseoppi*. Helsinki: Tietosanakirja.

Starke, Michal. 2009. “Nanosyntax: A Short Primer to a New Approach to Language.” *Nordlyd* 36: 2–6.

Sulkala, Helena, and Merja Karjalainen. 1992. *Finnish*. London: Routledge.

Suutari, Toni. 2006. “Body Part Names and Grammaticalization.” In *Grammar from the Human Perspective: Case, Space and Person in Finnish*, edited by Marja-Liisa Helasvuo and Lyle Campbell, 101–28. Amsterdam: John Benjamins.

Svenonius, Peter. 2003. “Limits on p: Filling in Holes vs. Falling in Holes.” *Nordlyd* 31: 431–45.

Svenonius, Peter. 2006. “The Emergence of Axial Parts.” *Nordlyd* 33: 1–22.

Svenonius, Peter. 2007. “Adpositions, Particles and the Arguments They Introduce.” In *Argument Structure*, edited by Eric Reuland, Tanmoy Bhattacharya, and Giorgos Spathas, 63–103. Amsterdam: John Benjamins.

Svenonius, Peter. 2008. “Projections of P.” In *Syntax and Semantics of Spatial P*, edited by Anna Asbury, Jakub Dotlačil, Berit Gehrke, and Rick Nouwen, 63–84. Amsterdam: John Benjamins.

Svenonius, Peter. 2010. “Spatial P in English.” In *Mapping Spatial PPs*, vol. 6 of *The Cartography of Syntactic Structures*, edited by Guglielmo Cinque and Luigi Rizzi, 127–60. New York: Oxford University Press.

Svenonius, Peter. 2012. “Structural Decomposition of Spatial Adpositions.” Presented at *The Meaning of P*, Universität Bochum, November 24, 2012. http://ling.auf.net/lingbuzz/001776.

Talmy, Leonard. 2000. *Concept Structuring Systems*, vol. 1 of *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.

Vainikka, Anne. 1989. "Deriving Syntactic Representations in Finnish." PhD diss., University of Massachusetts, Amherst.

Viitso, Tiit-Rein. 1998. "Estonian." In *The Uralic Languages*, edited by Daniel Abondolo, 115–48. London: Routledge.

Winkler, Eberhard. 2001. *Udmurt*. München: Lincom Europa.

Wood, Jim. 2015. *Icelandic Morphosyntax and Argument Structure*. Dordrecht: Springer.

Woolford, Ellen. 2005. "Lexical Case, Inherent Case, and Argument Structure." *Linguistic Inquiry* 37: 111–30.

On the New Expression *Bucuo V* in Taiwan Mandarin and Its Implications for Rule Borrowing

Jen Ting

National Taiwan Normal University, Taipei, Taiwan

ting@ntnu.edu.tw

Abstract: Studying the morphosyntactic properties of the new expression *bucuo*-V "good to V" in Taiwan Mandarin (TM), we first show that the *bucuo* V "good to V" sequence is a word and not generated in the syntax proper. Then we demonstrate that the morphological structure of *bucuo*-V "good to V" is unique for the grammar of TM in patterning more with that of its equivalent(s) in Taiwan Southern Min (TSM) than with that of any other existing word in TM. We move on to argue that a morphological rule that generates *bebai*/*bephai*-V "good to V" in TSM is responsible for deriving *bucuo*-V "good to V" in TM when the rule is borrowed or transferred from TSM to TM via language contact. The results of this study lend support to Thomason's (2006, forthcoming) theory of rule borrowing as well as provide evidence for the view that syntactic change may result from syntactic borrowing.

Keywords: morphosyntactic; *bucuo*-V; Taiwan Mandarin; language contact; rule borrowing

1. Introduction

Contact-induced structural change has long been an area of heated debate in contact linguistics. It is traditionally assumed (e.g., Lass 1980) that language change is caused by internal evolution and thus rule-governed and regular. Systematic changes induced by language contact are unexpected because externally-motivated changes would be unpredictable (Poplack and Levey 2010). Even if one adopts the view that contact-induced change may affect the structural (e.g., morphological and syntactic) domains, whether such change comes about as an indirect consequence of lexical borrowing (King

2000; Sankoff 2002) or results from application of morphosyntactic rules (Thomason 2006, forthcoming) or mechanisms (e.g., Harris and Campbell 1995) remains unsettled.

In this article, we use the new expression *bucuo* V "good to V" in Taiwan Mandarin discussed in Tseng (2003), Kao (2008), Lien (2008), and Cheng (2014), illustrated in (1), as a case study to show that it serves as an instance of rule borrowing, thus in support of direct involvement of morphosyntactic rules in contact-induced change.

(1) (a) Zhege xingren bing bucuo chi.
this.CL almond cake not.bad eat
"This almond cake tastes good." (Tseng 2003, 105)

(b) Zhengbu dianying zhende shi hen bucuo kan.
whole.CL movie really SHI very not.bad see
"The whole movie is indeed really very good to see." (Tseng 2003, 105)

(c) Zheben shu kanqilai bucuo du.
this.CL book read.QILAI not.bad read
"This book appears to be fun to read." (Kao 2008, 224)

2. *Bucuo* V "Good to V" as a Word

In this section, we argue that sequences like *bucuo* V "good to V" are not generated in the syntax proper, but rather are generated as a word in the morphology component. It will be further argued in Section 3 that the word *bucuo*-V "good to V" has a morphological template $[\text{adv-V}]_{\text{adj}}$ with the first component being disyllabic. A priori, there are at least three possible structures for the *bucuo* V "good to V" sequence. First, as adverbs often occur in the position immediately before the verb as shown in (2), it appears natural that *bucuo* "good," being in a preverbal position, is an adverb.

(2) (a) Ta xingfen-de pao-jin-lai.
he excited-DE run-enter-come
"He excitedly ran in." (Li and Thompson 1981, 323)

(b) Wo yanli-de zebei ta le.
I stern-DE reproach he LE
"I sternly reproached him." (Li and Thompson 1981, 323)

Under this approach, Tseng (2003, 105) claims that *bucuo* "good" in this construction is an adverb, but not clarifying whether *bucuo* "good" is a phrase or part of a word. Kao (2008, 227) points out that *bucuo* "good" is an adverbial in the *bucuo* V "good to V" sequence, just like its TSM counterpart *bebai* "good" in the *bebai* V "good to V" construction (see

Lien 2011 for a similar approach). One could further claim that with *bucuo* "good" being adverbial, the sentence can be analyzed as a middle sentence on a par with middle constructions in English (3) (e.g., Keyser and Roeper 1984; Stroik 1992, 1995, 1999):

(3) The bread cuts easily.

Another possible approach is that given that *bucuo* "good" alone can be an adjectival predicate in Chinese, the *bucuo* V "good to V" construction apparently resembles complement object deletion (COD) constructions in English (4a) (cf. Lasnik and Fiengo 1974), where an adjectival predicate takes a complement clause with an object gap. The object gap *e* in (4b) may be derived by null operator movement as proposed by Chomsky (1982) or by A'-binding as proposed by Cinque (1990) depending on what theory one adopts.

(4) (a) The article was too long for us to read *e*. (Cinque 1990, 98)

(b) The article was too long [O_i for [us to read e_i]] (Cinque 1990, 99)

Under this approach, the *bucuo* V "good to V" construction like (1) would have a structure as in (5) (with irrelevant details omitted).

(5) Zhege xingrenbing [VP bucuo [CP … chi e]]
this.CL almond.cake not.bad eat
"This almond cake tastes good."

Despite the initial plausibility of the above two syntactic approaches to the structure of *bucuo* V "good to V," we argue for an approach where the *bucuo* V "good to V" sequence is analyzed as a word that is generated in the morphology component and not in the syntax proper. For the sake of concreteness, we assume with Lieber and Scalise (2007) that the principles needed to construct complex words are distinct from principles needed to construct phrases and sentences. We further assume with researchers embracing a lexicalist theory (e.g., Chomsky 1995; Li 2005) that the former principles operate in the lexicon while the latter in the syntax proper.

To argue for the wordhood of *bucuo* V "good to V," we apply two tests that Wei (2005) provides in support of his analysis of *rongyi* V "easy to V" and the synonymous *hao* V "easy to V." He claims that the former has a structure where the "tough" predicate *rongyi* "easy" is a free morpheme taking a complement clause while the latter is a word.

The first test Wei (2005) provides is that parts of a word cannot be separated by a syntactic element such as a PP. According to him, this is because the intervention of the PP would induce violation of the Lexical Integrity Hypothesis (or LIH), according to which rules that apply in syntax to phrases cannot affect the internal structure of

words (Jackendoff 1972; Huang 1984). We find that independently words in Chinese are inaccessible to syntactic operations whatever theory this fact is captured by (e.g., the Limited Access Principle, together with the statement of Morphological Merge as suggested by Lieber and Scalise [2007]). To illustrate, a subordinative compound with an adv-V template in the sense of Chao (1968) such as *bei-ming* "sadly honk" from Liu (2010) does not allow a PP occurring between the two components of the lexical compound.

(6) *Da yan bei zai kong-zhong ming.
big wild.goose sadly at air-middle honk
Intended: "Big wild geese were honking in the air."

When we apply this PP intervention test to *bucuo*-V "good to V," we find that *bucuo* "good" and the following verb cannot be separated by prepositional phrases as shown by the contrast in (7).

(7) (a) Zhurou-xian zai tiaowei shang hai bucuo chi.
pork-filling at seasoning top rather not.bad eat
"The pork filling tastes good in terms of seasoning."

(b) *Zhurou-xian hai bucuo zai tiaowei shang chi.
pork-filling rather not.bad at seasoning top eat
"The pork filling tastes good in terms of seasoning."

This acceptability contrast would be left unexplained if *bucuo* "good" is an adverbial adjunct modifying the predicate as under a syntactic middle analysis, or is an adjectival predicate taking a clausal complement as under a COD analysis. On the other hand, facts like (7) follow from the word approach to *bucuo* V "good to V" given whatever approach that accounts for the LIH effects illustrated by (6).

Another test for wordhood Wei (2005) provides is based on the distribution of phase markers like *wan* "finish" and *hao* "good." Assuming with Tang (1992), phase markers are semi-affixes that lexically combine with morphemes (or roots) to form compounds or complex verbs. As shown in (8), the verb following *rongyi* "easy" is a proper host for the phase markers to attach to, but the verb following *hao* "easy" is not. This contrast is used by Wei to argue for the proposal that the verb following *rongyi* "easy" is a free morpheme, but that following *hao* "good" is part of a word.

(8) (a) Zhejian shi hen rongyi /*hao zuo-wan.
this.Cl matter very easy easy do-finish
"The thing is very easy to finish." (Wei 2005)

(b) Yingwen hen rongyi /*hao xue-hao.
English very easy easy learn-good
"English is very easy to learn well." (Wei 2005)

Turning to *bucuo* V "good to V," we find that the verb in the *bucuo* V "good to V" sequence cannot be followed by a phase marker such as *wan* "finish" in (9).

(9) *Zhepian wenzhang bucuo du-wan.
this.Cl article not.bad read-finish
"This article is good to finish reading."

If *bucuo* "good" is an adverbial adjunct or an adjectival predicate, it is not clear what renders (9) unacceptable (cf. acceptability of cases involving *rongyi* "easy" in [8]). On the other hand, *bucuo* V "good to V" patterns with *hao* V "easy to V" in being incompatible with a phase marker. Facts like (9) thus pose a serious challenge to the syntactic approaches whether it is a syntactic middle analysis or a COD analysis, but favor a lexical word approach to *bucuo* V "good to V."

In this section, we have argued that the sequence *bucuo* V "good to V" is a word, and is not generated in the syntax proper. In the next section, we will consider the issue of whether the emergence of this new expression is internally or externally motivated and reach the conclusion that the emergence of *bucuo*-V "good to V" is not likely to be internally caused.

3. The Rise of *Bucuo*-V "Good to V" Being Internally Caused?

In the literature on language change, a distinction is often drawn between internally and externally motivated change (Milroy 1992; Campbell 1998; Croft 2001). We argue that the emergence of *bucuo*-V "good to V" is not so likely to be internally motivated by showing that it has a unique morphological structure (more specifically the syllable structure) in comparison to the morphological structure of other words in TM in general.

We shall start with discussing the morphological template of *bucuo*-V "good to V." The word is formed by merging *bucuo* "good" with a V on its right to form an adjective. Judging from the relation between the first and second component of the word, *bucuo*-V "good to V" falls into the type of modifier-head compound under Chao's (1968) typology of compounds in Chinese. The morphological template of *bucuo*-V "good to V" can thus be represented as $[\text{adv-V}]_{\text{adj}}$.

The V part of *bucuo*-V "good to V" is often claimed to be subject to a monosyllabicity constraint in the literature (Tseng 2003; Kao 2008). This constraint, however, as pointed out by Cheng (2014), appears to be getting relaxed. According to Cheng (2014), based on corpus results obtained from a Google search, the verbs following *bucuo* "good" can now be disyllabic, including *wanle* "have fun," *wanshua* "play," *youwan* "play" and *chuli* "handle,"

etc. But still, if we consider the 260 tokens in her corpus, we find that 237 of them (91.1%) are monosyllabic and that only 23 of them (8.9%) are disyllabic, suggesting that although the monosyllabicity constraint on the syllable length of the verb part in *bucuo*-V "good to V" is getting relaxed, it still plays a role in regulating the pattern of *bucuo*-V "good to V."

Now we argue that the syllable structure of *bucuo*-V "good to V" is quite unique and that it is this uniqueness that renders it unlikely to be motivated by language-internal properties because there may be no other word in TM that has the same syllable structure. Recall that in this template $[\text{adv-V}]_{\text{adj}}$, the first component *bucuo* "good" is disyllabic and the second component is mainly monosyllabic but can be disyllabic; the resulting word is an adjective. For ease of discussion, we represent the syllable number of the morpheme by Arabic numbers and consider both the [2+1] and [2+2] combinations of morphemes within the word in turn.

Regarding the [2+1] combination, we compare a trisyllabic *bucuo*-V "good to V" with other trisyllabic compound words in TM. Dong (2014) points out that Mandarin compound words composed of a disyllabic and a monosyllabic morpheme are nouns, such as in (10).

(10) (a) youyong-chi
swim-pool
"swimming pools"

(b) dengshan-xie
climb.mountain-shoe
"mountain-climbing shoes"

(c) sushi-mian
speed.eat-noodle
"instant noodles"

Pan (2010) also observes that a trisyllabic adjectival compound with a modifier-head structure must be composed of a monosyllabic and a disyllabic morpheme, as illustrated in (11).

(11) (a) bu-mingyu
not-reputation
"infamous"

(b) bu-rendao
not-humane
"inhumane"

(c) da-wuwei
big-dauntless
"of great bravery"

Taken together, the generalizations from the above two works indicate that TM in general does not have trisyllabic [2+1] adjectives of a modifier-head structure with the trisyllabic *bucuo*-V "good to V" being an exception.

Turning to the [2+2] combinations, TM also does not have [2+2] adjectives of a modifier-head structure in general except for *bucuo*-V "good to V." According to Wei (2012, 58–60), among the 90 four-syllable words collected in a dictionary, only 12 of them have a [2+2] syllable structure and none of them have a modifier-head relation between the two components of the word. This finding clearly distinguishes a four-syllable *bucuo*-V "good to V" from other four-syllable words in TM.

Some remarks concerning *hao*-V "good to V" are necessary so as to cast doubt on its being considered a possible (major) motivation of the rise of *bucuo*-V "good to V" as occasionally suggested in the literature (e.g., Tseng 2003).

The word *hao*-V "good to V" may be a word that is the closest to *bucuo*-V "good to V" in meaning and structure in TM. Both *bucuo* "good" and *hao* "good" have similar meanings, only differing in that the latter expresses a high degree but the former a lower degree of appraisal (Shang 2006). Like *bucuo* "good," *hao* "good" can also occur in an attributive position (12a), a predicative position (12b) or a postverbal position (12c) in TM.

(12) (a) hen hao de yijian
very good DE opinion
"a very good opinion"

(b) Zhedao cai weidao hen hao.
this.CL dish taste very good
"The taste of this dish is good."

(c) Wo de zi xie de hao.
I DE word write DE good
"I write well."

Like *bucuo* "good," *hao* "good" can also merge with a V on its right to form a word. Abstracting away from the issue whether *hao* "good" is a prefix or the first morpheme in a word, it is clear that *hao*-V "good to V" is distinct from *bucuo*-V "good to V" in terms of syllable structure. The adjective *hao*-V "good to V" has a [1+1] syllable structure while *bucuo*-V "good to V" has a [2+1] or [2+2] structure. They do not have the

same syllable structure. This sole difference suffices to set them apart. Such constraints on syllable structure of words are commonly imposed on the morphology of Chinese. Interested readers are referred to the discussion of constraints on syllable count on NN compounds in TM as pointed out by Duanmu et al. (2015). We therefore claim that despite some close similarity in meaning and structure shared by *bucuo*-V "good to V" and *hao*-V "good to V," the latter is unlikely to be the motivation, or at best may not be the major motivation of the rise of the former.

4. The Rise of *Bucuo*-V "Good to V" Being Externally Caused

In this section, we start with showing that *bucuo*-V "good to V" shares the same morphological structure with its TSM equivalents *bebai*/*bephai*-V "good to V." There is evidence that the sequence *bebai*/*bephai* V "good to V" in TSM, like *bucuo*-V "good to V," is a word and is not generated in the syntax proper. Parallel to the behaviors of words such as *bei-ming* "sadly honk" in TM as in (6), *bebai*/*bephai* V "good to V" does not allow elements to be inserted between *bebai*/*bephai* "good" and V as in (13a); furthermore, *bebai*/*bephai* "good" cannot combine with a verb carrying a phase marker as in (13b) (cf. ex. [8]). Unacceptability of examples like (13) indicates that *bebai*/*bephai* V "good to V" is best analyzed as a word, rather than as a structure where *bebai*/*bephai* modifies a VP in a syntactic middle construction (cf. [3] in English) or takes a VP complement in a complement object deletion (COD) construction (cf. [4] in English).

(13) (a) *Tshaua te bebai/ bephai di tua juah tinn im.
herb tea not.bad not.bad on big hot day drink
"The herb tea is good to drink on a hot day."

(b) *Tsittiau kua bebai/ bephai tian liau.
this.Cl song not.bad not.bad listen finish
"This song is good to finish listening to."

We thus conclude that *bebai*/*bephai* V "good to V" in TSM, like *bucuo*-V "good to V" in TM, is also a word, with the morphological template $[\text{adv-V}]_{\text{Adj}}$.

If we consider the syllable structure of *bucuo*-V "good to V" in TM and *bebai*/*bephai* V "good to V" in TSM, we find that both have the same syllable structure. As pointed out by Kao (2008), the V part of *bebai*/*bephai*-V "good to V" can be monosyllabic or disyllabic as illustrated by *tautin* "be together" and *phue-png* "to go with rice." Although it is claimed by Tseng (2003) and Kao (2008) that the V part of *bucuo*-V "good to V" must be monosyllabic, as mentioned in Section 3, we assume with Cheng (2014) that the monosyllabicity requirement on the V part of *bucuo*-V "good to V" is getting relaxed because disyllabic verbs, despite a small portion (8.9%), can now

occur in the V part of *bucuo*-V "good," including *wanle* "have fun," *wanshua* "play," *youwan* "play" and *chuli* "handle," etc. We thus conclude that *bucuo*-V "good to V" in TM and *bebai*/*bephai* V "good to V" in TSM not only share the same morphological template [adv-V]$_{Adj}$, but also the same syllable structure.

In summary, given that *bucuo*-V "good to V" in TM and *bebai*/*bephai* V "good to V" in TSM share the same morphological template [adv-V]$_{Adj}$ and syllable structure, it is fairly plausible to claim that the emergence of the former is externally motivated by the latter via language contact.

5. The Mechanism of Emergence of *Bucuo*-V

The previous studies in the literature (Tseng 2003; Kao 2008; Lien 2008; Cheng 2014) all point to the influence of TSM on TM for the emergence of *bucuo*-V "good to V" in TM. While concurring with this view, we argue that the previous works do not fully explicate what mechanism underlies the emergence of the new expression in the context of language contact, and that the expression of *bucuo* V "good to V" arises from borrowing of a morphological rule from the source language, and therefore involves a process of rule borrowing as proposed by Thomason (2006, forthcoming).

5.1 Different Morphosyntactic Behavior of *Bucuo*-V "Good to V" and *Bebai*/*Bephai* V "Good to V"

To facilitate the evaluation of the previous proposals for the emergence mechanism of *bucuo*-V "good to V," it is instructive to consider the different behaviors of *bucuo*-V "good to V" in TM and *bebai*/*bephai*-V "good to V" in TSM. In Section 4, we pointed out that *bucuo*-V "good to V" in TM and *bebai*/*bephai* V "good to V" in TSM share the same morphological template [adv-V]$_{Adj}$ and syllable structure. Nevertheless, we would like to point out that it is not the case that they have identical morphosyntactic behaviors. In fact, *bucuo* "good" and *bebai*/*bephai* "good" allow different classes of verbs to merge with them to form an adjective. According to Cheng (2014), TM has examples like *bucuo-chuli* "good to handle" as in (14), but the TSM counterpart is not attested.

(14) (a)

Di	yin	jiu	yong	AiPlayer.	Ge	ge	yinyu
bass	sound	simply	use	AiPlayer	each	Cl	register
dou	hen	bucuo	chuli	de.			
all	very	not.bad	deal with	DE			

"Simply use AiPlayer to play the bass sound; all registers are very good to deal with."

(b) A: Yong dan qing jiu neng chu
use egg white simply able rid
diao PU jiao haiyo qingchu san
fall PU glue moreover clean.rid three
miao jiao yong bingtong.
second glue use acetone
"Simply use egg white to get rid of PU glue; moreover, use acetone to get rid of three second glue."

B: Xiexie ni. Man bucuo chuli de.
thank you very not.bad deal with DE
"Thank you. Very good to deal with."

Similarly, we find that acceptable instances in TSM such as *bebai*/*bephai-suihok* "good to convince," *bebai*/*bephai-tsingli* "good to tidy up," *bebai*/*bephai-hiangsiu* "good to enjoy," *bebai*/*bephai-tsohue* "good to be together with," *bebai-phuepng* "good to go with rice" (cf. Kao 2008), etc. cannot find equivalents in TM.

Given the above discussion, we conclude that *bucuo*-V "good to V" in TM and *bebai*/*bephai*-V "good to V" in TSM, though having the same morphological template and the same types of verbs as the V part in the compound, behave differently with respect to the inventories of verbs participating in the patterns. This conclusion will help us evaluate the previous proposals regarding the mechanism underlying the rise of *bucuo*-V "good to V."

5.2 The Calque Approach

Calques are defined by Thomason (2001, 260) as "a type of interference in which word or sentence structure is transferred without actual morphemes . . . typically a morpheme-by-morpheme translation of a word from another language." We find that examples of calques from TSM to TM can be illustrated by instances like *deng wu ren* "wait for no one" in TM, which is a borrowing of *tan bo lang* (lit. "wait for no person") in TSM (cf. Yen 2008). If *bucuo*-V "good to V" is coined by morpheme-by-morpheme translation of *bebai*/*bephai*-V "good to V," then we would expect that every instance of *bucuo*-V "good to V" would have a TSM *bebai*/*bephai*-V "good to V" counterpart, contra fact. This is evidenced by the fact discussed in Section 5.1; that is, instances of *bebai*/*bephai*-V in TSM such as *bebai*/*bephai-suihok* "good to convince," *bebai*/*bephai-tsingli* "good to tidy up," *bebai*/*bephai-hiangsiu* "good to enjoy," *bebai*/*bephai-tsohue* "good to be together with," *bebai-phuepng* "good to go with rice," etc. cannot find equivalents in TM.

5.3 The Pattern Replication Approach

Next, we argue against Cheng's (2014) claim that *bucuo*-V "good to V" in TM is an instance of what is labelled by Matras and Sakel (2007) as pattern replication borrowing from TSM.

According to them (2007, 841), pattern replication is "replication of usage patterns (organisation, distribution, and the mapping of grammatical or semantic meaning) from a model language" and does not involve transferring the phonological shape and morphological form from one language to another.

To illustrate, in the Macedonian dialects of Turkish, the infinitive in modal constructions has been replaced by a finite structure, just like its potential model languages (the contiguous languages of the Balkan):

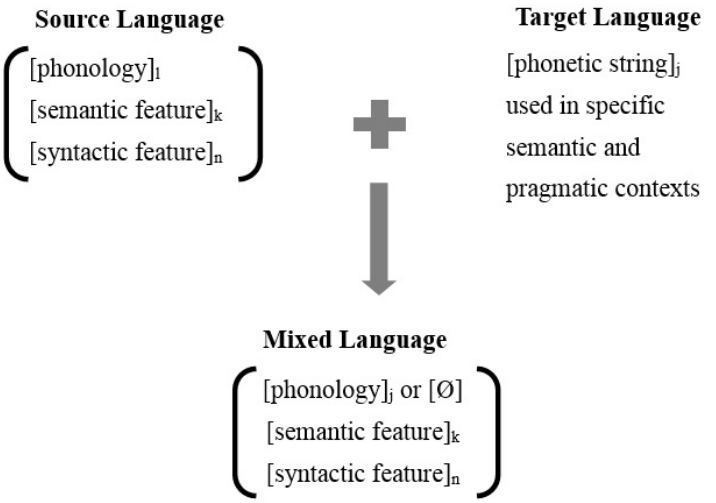
(15) istiyor git-sin
want.3SG go.3SG.SUBJ
"He wants to go."

Turning to the case of *bucuo*-V "good to V," if it is borrowed into TM by pattern replicating *bebai*/*bephai*-V "good to V" in TSM as argued by Cheng (2014), then we expect that when the pivot feature of the structure from TSM is mapped to TM, the result should respect various constraints in TM. However, as shown in Section 3, the syllable structure of *bucuo*-V "good to V" does not correspond to any other existing word in TM. This thus casts serious doubt on analyzing the rise of *bucuo*-V "good to V" along the lines of pattern replication.

5.4 The Relexification Approach

Another view of the emergence mechanism of *bucuo*-V "good to V" is provided by Lien (2008), who claims that this new expression is a result of relexification (cf. Lefebvre and Lumsden 1994; Lefebvre 2001) in the lexicon of TM. According to him, when two languages come into contact, new lexical items can be built by copying the semantic and syntactic features from the source language and the phonological features from the target language. The semantic and syntactic representations of source and target language entries need to only partially overlap. This is shown in the representation as in (16) (see Lien 2008, slide 8):

(16)



Regarding the case of *bucuo*-V "good to V," it is pointed out that, the lexical entry *bucuo* "good" in TM is selected to undergo "relexification," as a result of which *bucuo* "good" acquires the features of *bebai*/*bephai* "good" and the new expression *bucuo*-V "good to V" is coined. Under this approach, we may assume that first of all, the semantic and morphosyntactic features of the lexical entry *bebai*/*bephai* "good" are copied from TSM and the phonological features of *bucuo* "good" are copied from TM. The resulting new entry has semantic and morphosyntactic features from TSM and phonological features from TM. Now *bucuo* "good," endowed with the morphosyntactic features of *bebai*/*bephai* "good," is capable of merging with a verb to form an adjective.

To explain why this may not be a plausible analysis, we would first like to spell out how *bucuo* "good" in TM or *bebai*/*bephai* "good" in TSM merges with a verb to form an adjective. As shown in Section 5.1, it is a fact that neither *bucuo* "good" nor *bebai*/*bephai* "good" is allowed to merge with ANY verb, but rather there are constraints on what verb they can merge with. For example, while *bucuo-chi* "good to eat" (cf. *chi-qilai* "eat-QILAI") is acceptable, the nearly synonymous *bucuo-chang* "good to taste" (cf. *chang-qilai* "taste-QILAI") is not. We assume that *bucuo* "good" in TM or *bebai*/*bephai* "good" in TSM must have some sort of intrinsic requirement on what verb it is compatible with whatever the ultimate account will be. Presumably, such requirement is akin to some specification in a lexical entry, which ensures that *tsiah* "eat" in TSM can take an NP/DP complement *te* "tea," *tsiu* "liquor," *khitsui* "soda water" and *kapi* "coffee," but *chi* "eat" in TM cannot.

Coming back to the relexification account of the rise of *bucuo*-V "good to V," recall that, as reviewed earlier, for a new entry in the mixed language, the semantic and

syntactic features are copied from the source language and the phonological features from the target language. Since the semantic and morphosyntactic features of the new *bucuo* "good" solely come from *bebai*/*bephai* "good" in TSM, the new entry *bucuo* "good" should be subject to the same intrinsic selectional requirement on the V part of the word as *bebai*/*bephai* "good" in TSM, and is expected to be able to merge with the TM counterparts of the TSM verbs that can be merged with *bebai*/*bephai* "good" in TSM. This prediction is not borne out. As discussed in Section 5.1, instances of *bebai*/*bephai*-V "good to V" in TSM such as *bebai*/*bephai-suihok* "good to convince," *bebai*/*bephai-tsingli* "good to tidy up," *bebai*/*bephai-hiangsiu* "good to enjoy," *bebai*/*bephai-tsohue* "good to be together with," *bebai-phuepng* "good to go with rice," etc. are acceptable but their *bucuo*-V "good to V" counterparts are not in TM. We thus conclude that the relexification approach cannot satisfactorily account for the mechanism of the rise of *bucuo*-V "good to V" in TM.

5.5 The Rule Borrowing Approach

Having argued against the approaches of calques, pattern replication and relexification, we now argue for an approach of rule borrowing to account for the mechanism underlying the emergence of *bucuo*-V "good to V." Thomason (2006, forthcoming) argues against a traditional view that rules cannot be borrowed; she further argues for the proposal that grammatical rules can be transferred from one language to another. Following Trask (1993, 245), a rule is defined as "any statement expressing a linguistically significant generalization about the grammatical facts of a particular language, especially when formulated within the formalism of some particular formal description." Clear examples of rule borrowing, she argues, involve "a contact-induced change in which an innovative generalization in the receiving language A matches a pre-existing rule in the source language B, but in which no morphemes have been transferred from B to A" (Thomason forthcoming, 12). Crucially, there is no transfer of actual morphemes from TSM to TM and therefore *bucuo*-V "good to V" quite perfectly fits Thomason's description of a good candidate of rule borrowing.

If *bucuo*-V "good to V" is indeed an instance of rule borrowing, then what is the rule that gets transferred from TSM to TM? We propose that it is a morphological rule that merges a disyllabic modifier with a verb on its right to form an adjectival compound. The disyllabic modifiers that can undergo this rule include *bebai*/*bephai* "good" in TSM and its equivalent *bucuo* "good" in TM when the rule has been borrowed from TSM into TM. Under this approach, *bebai*/*bephai* "good" in TSM and *bucuo* "good" in TM would merge with verbs that they are compatible with as intrinsically licensed by individual languages. We claim that when rules are transferred from one language to another, they may still be subject to different constraints imposed by the individual languages.

This claim is not at all outrageous if we consider other instances of syntactic borrowing cross-linguistically. Take the NV compounds in Spanish as an example. Varela

and Felíu (2003) discuss the new compounds in Spanish with an NV structure as in (17a), which contrast with native Spanish compounds with a VN structure as in (17b).

(17) (a) *ruidofabricante* "noise maker," *euroconversor* "euroconverter"

(b) *Escurr eplatos* "dish rack" (lit. "drains dishes")

According to Varela and Felíu (2003), the new compounds are coined by structural borrowing of an order manifested in English synthetic compounds such as *taxi driver*. Such N + V compounds must have an overt suffix, such as *-or* as in *euroconversor* "euro-converter" and *-ente* in *radioyente* "radio listener" in (18a). If the compound noun does not carry an affix, then the internal order for this type of compound is V + N as in (18b).

(18) (a) $N+Vsuf_N$: *euroconversor* "euroconverter" vs. **conversoreuros* (lit. "convertereuros"); *radioyente* "radio listener" vs. **oyenterradios* (lit. "listenerradios")

(b) $V\emptyset_N$+N: cubrecama "bedspread," (lit. "covers bed") vs. **camacubre* (lit. "bedcover"); *guardabosques* "forest ranger," (lit. "guards woods") vs. **bosqueguarda* (lit. "wood guards")

We can take these facts as indicating that Spanish borrows a rule of forming synthetic compounds from English as in (19) (see, e.g., Fabb 1984; Lieber 1983):

(19) [X V affix], where X is interpreted as an argument of V.

Varela and Felíu (2003) point out that most of these English-style compounds have a disyllabic noun as the first component or contain the binding or concatenating vowel [o] characteristic of learned compounds in Spanish. Given this restriction, we know that when the rule of forming synthetic compounds in (19) is borrowed into Spanish, it is not the case that a verb can merge with any element interpreted as its complement to form a compound; rather, the compound formation must be subject to constraints specific to Spanish, which do not apply to English.

Just like what we see in the constraints on syllable length of the second component of the compound *bucuo*-V "good to V" in TM, Spanish, when making English-style synthetic compounds, tends to use a disyllabic noun as the first component of such compounds. Similarly, just as *bucuo*-V "good to V" is coined by a rule that derives *bebai*/*bephai*-V "good to V" in TSM but the rule may not apply to the equivalent verbs in both languages, we find that the new compounds in Spanish, which are derived by the rule borrowed from English, may not have English equivalents. One such example

can be illustrated by *digitpuntor* "massager" (lit. "finger pointer") (as seen in Varela and Felíu 2003), which has no English counterpart. In other words, the rule applies to *digit* "finger" and *puntor* "pointer" in Spanish, but not to their equivalents in English. We take these parallel behaviors of *bucuo*-V "good to V" and the English-style synthetic compounds in Spanish as supporting evidence for our claim that rules transferred from one language to another may be subject to language-specific constraints.

Summarizing, in this section, we have argued that the new expression *bucuo*-V "good to V" in TM, emerges as a result of borrowing of a word formation rule from TSM, and that the results of rule application may be subject to language-particular constraints.

6. Concluding Remarks

The results of this study lend support to Thomason's (2006, forthcoming) theory of rule borrowing. Given that no actual lexemes are transferred from TSM to TM, we can be sure that the shared morphosyntactic properties of *bucuo*-V "good to V" in the recipient language and *bebai*/*bephai*-V "good to V" in the source language do not result from lexical borrowing, which one may argue enables the speakers to abstract a rule from the enriched lexicon. The findings of this study also provide evidence for the view that syntactic change may result from syntactic borrowing (Harris and Campbell 1995; Thomason 2006, forthcoming; Bowern 2008), contra the view that interference should be excluded as a possible explanation for syntactic change (Longobardi 2001, 278; cf. Chomsky and Halle 1968).

Funding Acknowledgement

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant No. 102-2410-H-003-022-MY2.

Works Cited

Bowern, Claire. 2008. "Syntactic Change and Syntactic Borrowing in Generative Grammar." In *Principles of Syntactic Reconstruction*, edited by Gisella Ferraresi and Maria Goldbach, 187–216. Amsterdam: John Benjamins.

Campbell, Lyle. 1998. *Historical Linguistics: An Introduction*. Cambridge, MA: MIT Press.

Chao, Yuen-Ren. 1968. *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Cheng, Yi-Hsin. 2014. "On the Sequence of *Bucuo* V in Taiwan Mandarin." Paper presented at the 15th National Conference on Linguistics (NCL-15), Tunghai University, Taichung, Taiwan, May 23–24.

Chomsky, Noam. 1982. *Some Concepts and Consequences of the Theory of Government and Binding*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English.* New York: Harper and Row.

Cinque, Guglielmo. 1990. *Types of A'-Dependencies*. Cambridge, MA: MIT Press.

Croft, William. 2001. *Explaining Language Change: An Evolutionary Approach.* London: Longman.

Dong, Xiufang. 2014. "2 + 1 Shi San Yinjie Fuheci Goucheng Zhong de Yixie Wenti" [Some Issues in Constructing Trisyllabic Compounds in Chinese]. *Hanyu Xuexi* 2014 (6): 3–10.

Duanmu, San, Yahong Xue, and Feng Qi. 2015. "The Phonology and Morpho-syntax of AN in Chinese." Paper presented at MidPhon 20, Indiana University, September 11–13.

Fabb, Nigel Alexander John. 1984. "Syntactic Affixation." PhD diss., Cambridge, MA: MIT.

Harris, Alice C., and Lyle Campbell. 1995. *Historical Syntax in Cross-Linguistic Perspectiv*e. Cambridge, MA: Cambridge University Press.

Huang, C.-T. James. 1984. "Phrase Structure, Lexical Integrity, and Chinese Compounds." *Journal of the Chinese Language Teachers Association* 19: 53–78.

Jackendoff, Ray. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.

Kao, Wan-yu. 2008. "Lun Xinxing Jiegou 'Bucuo V'" [On New Structure 'Bucuo V']. *Cheng Ta Zhongwen Xuebao* 9: 213–34.

Keyser, Samuel Jay, and Thomas Roeper. 1984. "On the Middle and Ergative Constructions in English." *Linguistic Inquiry* 15: 381–416.

King, Ruth. 2000. *The Lexical Basis of Grammatical Borrowing*. Amsterdam: John Benjamins.

Lasnik, Howard and Robert Fiengo. 1974. "Complement Object Deletion." *Linguistic Inquiry* 5: 535–71.

Lass, Roger. 1980. *On Explaining Language Change*. Cambridge, MA: Cambridge University Press.

Lefebvre, Claire. 2001. "Relexification: A Process Available to Human Cognition." In *The Proceedings of the 27th Annual Meeting of Berkeley Linguistic Society*, edited by Charles Chang, Michael J. Houser, Yuni Kim, David Mortensen, Mischa Park-Doob, and Marzia Toosarvandani: 125–39. Berkeley: University of California.

Lefebvre, Claire, and John S. Lumsden. 1994. "Relexification in Creole Genesis." Paper presented at the MIT Symposium on the Role of Relexification in Creole Genesis: The Case of Haitian Creole.

Li, Yafei. 2005. *X^0: A Theory of the Morphology-Syntax Interface*. Cambridge, MA: MIT Press.

Li, Charles N., and Sandra A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

Lieber, Rochelle. 1983. "Argument Linking and Compounds in English." *Linguistic Inquiry* 14: 251–85.

Lieber, Rochelle, and Sergio Scalise. 2007. "The Lexical Integrity Hypothesis in a New Theoretical Universe." In *On-line Proceedings of the Fifth Mediterranean Morphology Meeting*, edited by Geert Booij, Bernard Fradin, Angela Ralli, and Sergio Scalise. University of Bologna. http://mmm.lingue.unibo.it/proc-mmm5.php.

Lien, Chin-fa. 2008. "Yuyan Benneng de Meili: Zhi Nan Yan Yi: Cong *Bucuo Chi* Tan Qi" [The Attraction of Language Instinct: Easier to Speak than to Know: To Begin with *Bucuo Chi* 'good to eat']. Paper presented at the NSC lecture Series on Mass Science Education, Kaohsiung, Taiwan, Feb 24, 2008. http://science.nchc.org.tw/science/science2007/master/ 07mas. htm.

Lien, Chin-fa. 2011. "Aspects of Contact-Induced Changes: Interaction between TM and TSM." Paper presented at Yuan Zhi University Workshop on Language Structure and Language Learning, December 3–4. http://fl.hs.yzu.edu.tw/WLSLL/ppt/10.pdf.

Liu, Li. 2010. "Xiandai Hanyu Sanyinjie Zuhe Duojiaodu Kaocha" [Multi-perspective Study on Modern Chinese Three-syllable Combinations]. MA thesis, Wuhan, China: Central China Normal University.

Longobardi, Giuseppe. 2001. "Formal Syntax, Diachronic Minimalism, and Etymology: The History of French *Chez*." *Linguistic Inquiry* 32: 275–302.

Matras, Yaron, and Jeanette Sakel. 2007. "Investigating the Mechanisms of Pattern Replication in Language Convergence." *Studies in Language* 31: 829–65.

Milroy, James. 1992. *Linguistic Variation and Change.* Oxford: Blackwell.

Pan, Yan. 2010. "Xiandai Hanyu San Yijie Ci Jiegou Fenxi" [The Analysis of the Structure of Trisyllabic Words in Modern Chinese]. *Xiandai Yuwen (Yuyan Yanjiu Ban)* 2010 (11): 48–50.

Poplack, Shana, and Stephen Levey. 2010. "Contact-Induced Grammatical Change: A Cautionary Tale." In *Theories and Methods*, vol. 1 of *Language and Space: An International Handbook of Linguistic Variation*, edited by Peter Auer and Jürgen Erich Schmidt, 391–419. Berlin: Mouton de Gruyter.

Sankoff, Gillian. 2002. "Linguistic Outcomes of Language Contact." In *The Handbook of Language Variation and Change*, edited by Jack Chambers, Peter Trudgill, and Natalie Schilling-Estes, 638–68. Oxford: Blackwell.

Shang, Guowen. 2006. "'*Hao*' yu '*Bucuo*'" ['*Hao*' and '*Bucuo*']. *Yuwen Xuekan 2006* 5: 102–3.

Stroik, Thomas. 1992. "Middles and Movement." *Linguistic Inquiry* 23: 127–37.

Stroik, Thomas. 1995. "On Middle Formation: A Reply to Zuibi-Hertz." *Linguistic Inquiry* 26: 165–71.

Stroik, Thomas. 1999. "Middles and Reflexivity." *Linguistic Inquiry* 30: 119–31.

Tang, Ting-Chi. 1992. *Yuyan Fenxi yu Yingyu Jiaoxue* [*Linguistic Analysis and English Instruction*]. Ms. National Tsing Hua University.

Thomason, Sarah G. 2001. *Language Contact: An Introduction.* Edinburgh and Washington, DC: Edinburgh University Press and Georgetown University Press.

Thomason, Sarah G. 2006. "Rule Borrowing." In *Encyclopedia of Language and Linguistics.* 2nd ed. Volume 10, edited by Keith Brown, 671–77. Oxford: Elsevier.

Thomason, Sarah G. Forthcoming. "Can Rules Be Borrowed?" In *Festschrift for Terrence Kaufman*, edited by Thomas Smith-Stark and Roberto Zavala.

Trask, R. L. 1993. *A Dictionary of Grammatical Terms in Linguistics*. London: Routledge.

Tseng, Hsing-Yi. 2003. "Dandai Taiwan Guoyu de Jufa Jiegou" [The Syntax Structures of Contemporary Taiwanese Mandarin]. MA thesis, Taipei, Taiwan: National Taiwan Normal University.

Varela, Soledad, and Elena Felíu. 2003. "Internally Motivated Structural Borrowing in Spanish Morphology." In *Theory, Practice, and Acquisition: Papers from the 6th Hispanic Linguistics Symposium and the 5th Conference on the Acquisition of Spanish and Portuguese*, edited by Paula Kempchinsky and Carlos-Eduardo Piñeros, 83–101. Somerville, MA: Cascadilla.

Wei, Minghua. 2012. "'Xiandai Hanyu Cidian' (Diwu Ban) Sanzi Ge yu Size Ge Yanjiu" [The Research on the Three-word and Four-word Form in Modern Chinese Dictionary (The Fifth Edition)]. MA thesis, Hangzhou, China: Zhejiang University.

Wei, Ting-Chi. 2005. "Middle *Hao* as the Lexicalization of Tough *Rongyi*." *TELL Journal*: *Teaching of Languages, Linguistics, and Literature* 2: 17–34.

Yen, Hsiu-shan. 2008. "Taiwan Huayu Zhong de Minnan Fangyan Ci Chu Tan" [A Study of Southern Min Words in Taiwan Mandarin]. *Xinzhu Jiaoyu Daxue Renwen Shehui Xuebao* 1: 49–68.

Definiteness and Specificity in Two Types of Polish Relative Clauses

Wojciech Guz

The John Paul II Catholic University of Lublin, Poland

wguz@o2.pl

Abstract: This study contrasts Polish *który* and *co* relative clauses in terms of the definiteness and specificity of their relativized heads. As is shown with corpus data, *co* relatives are strongly associated with definite (especially demonstrative-headed) and specific NPs, while the majority of *który* relatives tend towards indefinites, half of which are also non-specific. Consequently, unlike *wh*-pronoun relatives, complementizer relatives exhibit restrictions such that the [-def] and/or [-spec] values (or their combinations) may be infelicitous in some contexts. Relative acceptability of sentence variants is compared by means of constructed examples, which complement the corpus material. The study also draws a parallel between nominal (in)definiteness and clausal (ir)realis mood in that prototypical *co* relatives involve definite specific NPs grounded in the context of actual (realis) events, rather than irrealis events.

Keywords: complementizer and *wh*-pronoun relatives; definiteness; specificity

1. Introduction

The goal of this study is to contrast two types of Polish relative clauses in terms of the definiteness and specificity of their relativized heads. The two types are illustrated in (1); (1a) is the standard construction with the inflected relative pronoun *któr-y*/*-a*/*-e*/etc. "who/which," (1b) employs the uninflected relative marker (complementizer) *co* "that" and is characteristic of colloquial informal style.

(1)	(a)	Te	jabłka,	które	masz	tu	na	stole.
		these	apples	which	have-2SG	here	on	table
		"These apples which you have here on the table."						

(b) Te jabłka, co masz tu na stole.
these apples CO have-2SG here on table
"These apples that you have here on the table."

The discussion in this paper relies on the distinction between definiteness and specificity, two properties of noun phrases (NPs) which are interconnected but may also operate independently. Namely, indefinites may be specific or non-specific, and so can definites, as in (2) and (3) from Lyons (1999, 165, 172).

(2) (a) I haven't started the class yet; I'm missing **a student**—there should be fifteen and I only count fourteen.

(b) I haven't started the class yet; I'm missing **a student**—Mary is always late.

(3) (a) We can't start the seminar, because **the student** who's giving the presentation is absent—typical of Bill, he's so unreliable.

(b) We can't start the seminar, because **the student** who's giving the presentation is absent—I'd go and find whoever it is, but no-one can remember.

In (2a) *a student* is indefinite and non-specific [-def-spec] because the referent of the NP is known neither to speaker nor hearer. In (2b) the NP is still indefinite but the reference is specific [-def+spec] in that the referent is known or familiar to the speaker (although still not identifiable to the hearer without the final part of the sentence). Specificity then is based on the knowledge of the speaker only. On the other hand, definiteness relates to the shared knowledge of both speaker and hearer (Fodor and Sag 1982). The same specificity distinction can be made for definites in (3): *the student* is [±spec] depending on the speaker's knowledge of the referent.

2. Previous Research

Previous research on the contrasts between Slavic relative clauses introduced by complementizers and relative pronouns has focused on a number of topics including: (i) inflected vs. uninflected relativizers and the use of resumptive pronouns in complementizer relatives (Gołąb and Friedman 1972; Rudin 1986; Bondaruk 1995; Szczegielniak 2006; Bošković 2009; Fried 2010; Hladnik 2015), (ii) the disputed categorial status of the uninflected relativizer (Minlos 2012), (iii) semantic and functional types and preferences (Fried 2011), (iv) standard vs. non-standard relative constructions (Murelli 2011).

As for definiteness, Fried (2010) and (2011) examines its relevance (but not of specificity) in Czech complementizer relatives. She finds explicit definiteness/deixis

marked by demonstratives heading NPs in approximately half of the relatives examined. Also, she notes that animacy and number are especially relevant here in that the prototypical constellation that attracts demonstratives involves a singular animate NP, and the frequency of demonstratives in relativized heads decreases along the scale: anim. sg. > inanim. sg. > (in)anim. pl. Fried concludes that, in the context of *co* clauses, the demonstratives observed are to be seen as an issue of referentiality or individuation rather than simple deixis. The preferred head referents in Czech *co* clauses tend to be entities relatively high in referentiality or individuation.

This study further explores the questions raised by Fried (2010; 2011) and contributes an examination of definiteness and specificity in Polish *który* and *co* relative clauses. Then, in Section 8, the discussion expands beyond the properties of the NP and we turn to indicate symmetry between the definiteness and specificity at the level of the NP and the category of realis/irrealis mood at the level of the clause.

3. Corpus and Data

Since *co* relatives represent informal colloquial style, the data come from Spokes, a corpus of conversational spoken Polish (Pęzik 2015). Much of the corpus's transcribed material is aligned with audio data and it is only this section of the corpus that was used in the present study. The reason for this is that the audio material was used to verify that the transcripts are accurate and that only relevant tokens of *co* and *który* clauses were taken into account. In sum, approximately 77% of the corpus data were used, which translates into approximately 1.6 million words.

1,729 *który* relatives and 679 *co* relatives were collected from Spokes by an exhaustive search of all occurrences of the words *co* and *któr-y*/*-a*/*-e*/etc. Each occurrence was manually inspected so that only relevant tokens were collected. Included in the sample were subject and object clauses (direct and oblique). As *co* clauses are prototypically restrictive, *który* nonrestrictives were excluded and the entire sample consists of restrictive relatives only. Examples from Spokes are marked "Spokes." When sentence variants are compared for their relative acceptability, such examples are marked "constructed" or "modified" (i.e., modified versions of preceding examples). Original spelling and punctuation is preserved.

4. Definiteness and Specificity in the Spokes Corpus

The starting point of the discussion is the observation that *co* relatives are frequently accompanied by demonstratives in the head NPs. This is immediately apparent in analysis of the Spokes data as 80% of *co* clauses involve either a relativized NP introduced by a demonstrative or a self-standing pronominal demonstrative, as in (4) and (5) respectively:

(4) prawdopodobnie to **ten** koleś co się zwolnił od (Spokes)
probably it this bloke CO REFL resigned from
nas nie?
us no
"It's probably the bloke that resigned at our (company), right?"

(5) a to jest **ta** co te zdjęcia robiła? (Spokes)
and it is this-F CO these photos made
"It's the one (girl) that was taking those photos?"

It is worth noting here that both subject and object *co* relatives are similar in this respect as 77% of subjects and 85% of objects co-occur with demonstratives in the head. Examples (4) and (5) are subject relatives while example (6) is an object relative.

(6) wzięłyście **to** wino co wam przyniosłem? (Spokes)
took-2PL-F that wine CO you-DAT brought-1SG
"Did you take the wine I brought for you?"

On the other hand, only 25% of *który* clauses feature demonstrative-headed NPs, as shown in Table 1. This suggests a strong preference for *co* relatives to be associated with overtly definite NPs, compared to *który* relatives.

	***Co* clauses**	***Który* clauses**
Demonstrative	548 (80.7%)	447 (25.8%)
No demonstrative	131 (19.2%)	1,282 (74.1%)
Total	679 (100%)	1,729 (100%)

Table 1. Demonstratives in *co* and *który* relatives

With this in mind, consider the following examples from Kardela (1986, 90) and McDaniel and Lech (2003, 70), both marked with question marks by the respective authors.

(7) (Kardela's question mark)
?Przeczytałem gazetę, co kupiłem wczoraj.
read-1SG newspaper-ACC CO bought-1SG yesterday
"I have read a/the paper (that) I bought yesterday."

(8) (McDaniel and Lech's question mark)
?To jest kredka, co chłopiec nadepnął na nią.
this is crayon CO boy stepped on her
"This is a/the crayon that a/the boy has stepped on."

Kardela, and McDaniel and Lech find examples (7) and (8) awkward for the absence of a resumptive pronoun, which they assume to be required or at least welcome in object clauses. However, the same examples improve—are in fact the norm in spoken Polish—when the referents of the NPs are given definiteness, and even without the use of resumptives, as in (9) and (10).

(9) Przeczytałem **tę** **gazetę**, co kupiłem wczoraj. (modified)
read-1SG this newspaper CO bought-1SG yesterday
"I have read the paper (that) I bought yesterday."

(10) To jest **ta** **kredka**, co (modified)
this is this crayon CO
ten chłopiec/Adam/mój brat na nią nadepnął.
this boy/Adam/my brother on her stepped
"This is the crayon that the boy/Adam/my brother has stepped on."

In (9) and (10), the demonstratives mark the NPs as explicitly definite: "the newspaper" as opposed to "a newspaper," and "the boy" as opposed to "a boy." The referent of the NP is then further specified in the *co* clause: "the newspaper that I bought," "the boy that stepped on the crayon." As Polish has no definite or indefinite articles, demonstratives can provide the required explicit definiteness. In (10), replacing *ten chłopiec* with a proper noun (*Adam*) or a possessive phrase (*mój brat*) would have a similar effect of adding definiteness or uniqueness of reference.

The improved acceptability of (9)–(10) over (7)–(8) suggests that *co* relatives are better suited for definite rather than indefinite reference. Having said that, it comes as no surprise that example (9) is awkward when the demonstrative is replaced with the explicitly indefinite/non-specific determiner *jakiś*/*jakaś*/*jakieś*/etc. "some," as in (11).

(11) ?Przeczytałem **jakąś** gazetę, co kupiłem wczoraj. (modified)
read-1SG some newspaper CO bought-1SG yesterday

Also, genuinely non-restrictive *co* relative clauses are rare and/or substandard (12a) (cf. Fried 2011), and the reason is clear: the job of a *co* relative is to specify the referent,

not to give additional information about it. Note that the use of a standard *który* relative clause in non-restrictives eliminates the problem, as in (12b).[1]

(12) (a) ??Kupiłem jakąś gazetę, co (ją) (modified)
bought-1SG some newspaper CO (her)
przeczytałem w całości.
read-1SG in entirety
"I bought some newspaper that I read in its entirety."

(b) Kupiłem jakąś gazetę, którą przeczytałem (modified)
bought-1SG some paper which read-1SG
w całości.
in entirety
"I bought some paper, which I read in its entirety."

Thus, *co* relatives work better with definite NPs, while indefiniteness can be readily signalled by *który* relatives.[2] Consider the distinct (in)definiteness effects produced by the *co* relative and *który* relative in (13a) and (13b) (the difference is well captured by the English glosses). In (13a), the postposition of the NP *chłopiec* relative to the verb *nadepnął* additionally invites an indefinite reading.

(13) (a) To jest kredka, na **którą** (modified; indefinite)
this is crayon on which
nadepnął chłopiec.
stepped boy
"This is a crayon that was stepped on by a boy."

(b) To jest **ta** kredka, **co** (modified; definite)
this is this crayon CO
ten chłopiec/Adam na nią nadepnął.
this boy/Adam on her stepped
"This is the crayon that the boy/Adam has stepped on."

1 This also has parallels in the availability of definite/indefinite referents in English *wh*-relatives vs. *that* relatives (cf. *John gave me a book, which I read* vs. ?*John gave me a book that I read*).

2 Definiteness interacts with relative clause types in similar ways cross-linguistically. See Bošković (2009) for similar definiteness effects in Serbo-Croatian *što* relatives, *što* being a relative complementizer. Also, parallels can be seen in *deto* relatives in Bulgarian (Rudin 1986) and *co* relatives in Czech (Fried 2010).

Given the 80 percent rate of occurrence of demonstratives in *co* clauses, what types of NPs are found in the remaining 20 percent? Is it possible to establish any consistent pattern in the totality of the NPs? Consider four examples with definite and indefinite NPs including a possessive (14), a bare noun (15), and two determiner-headed NPs (16)–(17).

(14) gdzie jest **mój klej** co kupiłam (Spokes; +def+spec)
where is my glue CO bought-1SG
"Where is the (my) glue I bought?"

(15) są **ludzie** co wiesz nie mają (Spokes; -def-spec)
are people CO know-2SG not have-3PL
nic do garnka włożyć
nothing into pot put-INF
"There are people that have nothing to put in their cooking pot."

(16) poszliśmy chyba do **takiego** **pubu** (Spokes; -def+spec)
went-2PL probably to such pub
co się nazywa przechowalnia
CO REFL called storage room
"I think we went to this pub that's called Storage Room."

(17) **tą** **ziemię** co Zdzisiek kupuje przywozi (Spokes; +def-spec)
this soil CO Zdzisiek buys brings
i przywozi
and brings
"This soil that Zdzisiek buys and brings and brings."

Each referent of the NPs in (14)–(17) may be considered on the two planes of definiteness and specificity, with four different value combinations: [+def+spec] in (14), [-def-spec] in (15), [-def+spec] in (16) and [+def-spec] in (17). *Mój klej* is both definite (through the use of the possessive pronoun) and specific (a particular item), by contrast *ludzie* is indefinite and non-specific, *takiego pubu* is indefinite[3] (not identifiable to the addressee)

3 Following a reviewer's comment, one point is worth noting here. Namely, *taki* "such/this" has two uses—one definite, the other indefinite. This is paralleled in a similar dualism in English definite/indefinite *this* (Lyons 1999, 176–78), for example: *Spotkałem dziś* ***takiego faceta****. Nazywał się Jacek.* [-def+spec] "I met **this guy** today. His name was Jacek." vs. ***Takie zachowanie*** *jest karygodne.* [+def+spec] "**Such behaviour** is inexcusable." In its definite use, *taki* may combine with other definite determiners, such as *każdy* "each" (e.g., *każde takie zachowanie* [+def+spec] "each

but specific (known/familiar to the speaker), while *tą ziemię* is definite (identifiable to the addressee through the use of a demonstrative) but non-specific. As will be shown below, the interplay of the two categories of definiteness and specificity is what distinguishes prototypical *co* and *który* relatives. Specifically, in *co* relatives, indefinite referents such as that in (16) and non-specific referents such as that in (17) are relatively rare compared to definite and specific ones such as those in (13b) and (14); more to the point, NPs which are simultaneously indefinite and non-specific—such as that in (15)—are even more uncommon in *co* relatives and are normally handled by *który* relatives.

Tables 2 and 3 show the distribution of the def/spec value combinations in *co* clauses and *który* clauses respectively. As can be seen, *co* relatives have a preference for definite NPs and especially for definite NPs marking specific referents. On the other hand, *który* relatives are more readily used for indefinite NPs referring to both specific and non-specific referents (the latter are slightly more frequent). The quantitative information from Tables 1 and 2 are discussed in more detail in the following sections.

679 *co* clauses					
definites (demonstratives)		**definites (non-demonstratives)**		**indefinites**[4]	
487 (71%) +def+spec	61 (8%) +def-spec	15 (2%) +def+spec	1 (0.1%) +def-spec	77 (11%) -def+spec	38 (5%) -def-spec

Table 2. *Co* clauses in Spokes

instance of such behaviour"), but not in its indefinite use (**Spotkałem dziś każdego takiego faceta*.). Conversely, indefinite *taki* may combine with other indefinite determiners, such as *jakiś* "some" (e.g., *Spotkałem dziś jakiegoś takiego faceta.*[-def+spec] "I met this guy today."). In other words, definite *taki* is compatible with other definite determiners, while indefinite *taki* is compatible with other indefinite determiners. The two uses of *taki* were treated separately in the quantitative counts below.

4 In this study we distinguish definite and indefinite NPs in a somewhat restricted fashion. *Co* clauses are frequently introduced by demonstratives, and thus are made overtly definite. Other markers of definiteness include possessives and universal quantifiers such as *każdy* "each" and *wszyscy* "all." All these are taken as marking definiteness (after Lindvall 1996 and Lyons 1999, Section 1.2). In contrast, NPs without such overt markers of definiteness will be referred to as indefinite. They may be bare NPs or may be introduced by indefinite determiners (e.g., *taki* "such/this" [only the indefinite use parallel to English indefinite *this*], *jeden* "one," *pewien* "certain," *jakiś* "some"). In other words, "indefinite" here means not overtly definite through use of definite determiners—this is slightly different to what some scholars say about bare NPs whose referents can receive a definite interpretation through, e.g., word order. The purpose of such a distinction of (in)definiteness is to highlight the contrast between NPs headed by demonstratives (overtly definite) and NPs without demonstratives (not overtly definite). As is shown, the contrast based on the presence/absence of demonstratives is vital to our comparison of *który* and *co* relatives.

1,729 *który* clauses					
definites (demonstratives)		**definites (non-demonstratives)**		**indefinites**	
303 (17%) +def+spec	144 (8%) +def-spec	42 (2%) +def+spec	30 (1%) +def-spec	581 (33%) -def+spec	629 (36%) -def-spec

Table 3. *Który* clauses in Spokes

5. A Closer Look at *co* Clauses in Spokes

When *co* relatives are considered in more detail, based on Table 2, one observes two general quantitative tendencies:

Tendency 1

Definites (83%) are far more common than indefinites (16%); definiteness is typically marked by demonstratives and rarely by other means (possessives, universal quantifiers *każdy* "each," *wszyscy* "all," proper nouns). Specificity is another key factor: 89% of definites are associated with specific referents, and 10% with non-specifics. Thus a typical *co* relative is [+def+spec], as in (18):

(18) to są te co dla babci Jasi są wybrane (Spokes)
this are these CO for grandma Jasia-DAT are selected
"These are the ones selected for grandma Jasia."

Tendency 2

When indefinites are involved, specific referents are again more frequent (66%) than non-specific (33%). Some of the Spokes *co* relatives with indefinite non-specifics sound infelicitous and they improve when specificity is added. Consider examples (19a) and (20a) from Spokes with their improved [+spec] versions in (19b) and (20b).

(19) (a) poczekaj, może mam **jakieś** (Spokes, -def-spec)
wait-IMP maybe have-1SG some/any
zdjęcia **co** ci mogę pokazać
photos CO you-DAT can-1SG show
"Wait, I might have some photos that I can show you."

(b) poczekaj, mam tu **dwa** **takie** (modified, -def+spec)
wait-IMP have-1SG here two such
zdjęcia **co** ci mogę pokazać
photos CO you-DAT can-1SG show
"Wait, I have these two photos that I can show you."

(20) (a) chciałbym **kartę** **co** GTA 5 (Spokes, -def-spec)
would like-1SG card CO GTA 5
obsłuży na pełnych detalach
handle-FUT on full details
"I'd like a card that will be able to handle GTA 5 on full graphics."

(b) mam **taką** **kartę** **co** GTA 5 (modified, -def+spec)
have-1SG such card CO GTA 5
obsłuży na pełnych detalach
handle-FUT on full details
"I have this card that will be able to handle GTA 5 on full graphics."

By way of comparison, many *który* clauses with indefinite and non-specific referents sound awkward when the relative pronoun (21a) is replaced with *co* (21b). However, when definiteness and/or specificity are added, as in (21c) and (21d)[5] the results are much better.

(21) (a) ale chodzi o **piosenki** **które** ludzie (Spokes, -def-spec)
but be about about songs which people
chcą jeszcze raz usłyszeć na żywo
want more once hear on live
"But it's about songs which people want to hear live once more."

(b) ?ale chodzi o **piosenki** **co** (modified, -def-spec)
but be about about songs CO
ludzie chcą (je) jeszcze raz
people want (them) more once
usłyszeć na żywo
hear on live

(c) chodzi o **właśnie** **te** **piosenki** **co** (modified, +def-spec)
be about about just these songs CO
ludzie chcą jeszcze raz usłyszeć na
people want more once hear on
żywo
live
"It's about precisely those songs that people want to hear live once more."

5 Also, example (21d) is slightly better than (21c) for its realis mood in the *co* clause VP (see Section 8 for discussion). The proposition in the *co* clause has been modified to accommodate realis mood (also below in [23c] and [23d]).

(d) chodzi o **właśnie** **te** **piosenki** (modified, +def+spec)
be about about just these songs
co ostatnio słyszeliśmy na żywo
CO recently heard-1PL on live
"It's about precisely those songs that we recently heard live."

Let us consider the reverse situation, i.e., when the original *co* relative has a specific referent and the modified version with a non-specific referent sounds less felicitous. This is the case in (22a) and (22b); the original from Spokes in (22a) refers to a specific item; the reference to a non-specific item in (22b) is infelicitous and would work better with *który*.

(22) (a) pożyczyłam sobie **taką** **kaczkę** (Spokes, -def+spec)
borrowed-1SG oneself **such** **duck**
co się wkłada rękę
CO REFL put in hand
i ona gada niby
and she talks as if
"I borrowed this (toy) duck that you put your hand in and it talks, as if."

(b) ?**chciałabym** **kaczkę** **co** się wkłada (modified, -def-spec)
would like-1SG duck CO REFL put in
rękę i ona gada niby
hand and she talks as if
"I'd like a duck that you put your hand in and it talks, as if."

In sum, definiteness and specificity are important factors in *co* clauses in that the NPs involved tend to be definite, and the entities referred to tend to be specific. The most frequent def/spec value combinations are the following (in descending order): +def+spec (73%), -def+spec (11%), +def-spec (9%), -def-spec (5%). Very much in line with Fried's (2011) account of Czech material, Polish *co* relatives preferentially co-occur with heads of relatively high individuation in terms of not only definiteness, but also specificity.

6. A Closer Look at *który* Clauses in Spokes

When *który* relatives are considered in more detail, based on Table 3, one observes two general quantitative tendencies:

Tendency 1
Indefinite NPs are more frequent (70%) than definite (30%). In indefinites, non-specifics are slightly more frequent than specifics (51% and 48% respectively), which—compared

to indefinites in *co* relatives (non-specific 33%, specific 66%)—constitutes a notable increase in the rate of non-specific referents.

Tendency 2
When definites are involved, specific referents are more frequent (66%) than non-specific (33%), which—compared to definites in *co* relatives (specific 89%, non-specific 10%)—constitutes an increase in the frequency of non-specific referents.

Thus *który* clauses readily and frequently allow [-def/-spec] referents, as in (23a). In such cases, *co* often cannot replace *który* without sounding awkward, as in (23b), unless the referent of the relativized head is given specificity, as in (23c), or both definiteness and specificity, as in (23d).

(23) (a)

Austriak	**który**	ma	pracę	(Spokes, -def-spec)
Austrian	**who**	has	job	
nigdy,	dokładnie,	nie	kupi,	
never	exactly	not	buy-3SG-FUT	
nie	pójdzie	na	stragan	
not	go-3SG-FUT	on	stall	
i	nie	kupi		
and	not	buy-3SG-FUT		

"An Austrian who has a job will never, exactly, won't buy, won't go to a market stall and buy (it)."

(b)

?Austriak	**co**	ma	pracę	(modified; -def-spec)
Austrian	CO	has	job	
nigdy	nie	pójdzie	na	
never	not	go-3SG-FUT	on	
stragan	i	nie	kupi	
stall	and	not	buy-3SG-FUT	

(c)

jeden	**taki**	**Austriak**	**co**	ma	(modified; -def+spec)
one	such	Austrian	CO	has	
pracę	poszedł	na	stragan	i	kupił
job	went-3SG	on	stall	and	bought-3SG

"This Austrian [-def+spec], who has a job, went to a market stall and bought (it)."

(d)	**ten**	**Austriak**	**co**	ma	(modified; +def+spec)	
	this	Austrian	CO	has		
	pracę	poszedł	na	stragan	i	kupił
	job	went-3SG	on	stall	and	bought-3SG

"That Austrian [+def+spec] that has a job went to a market stall and bought (it)."

In sum, indefiniteness is an important factor in *który* relatives in that the NPs involved tend to be indefinite. This contrasts sharply with *co* relatives. The distribution of specific and non-specific referents is more evenly distributed than in *co* relatives. The most frequent def/spec value combinations are the following (in descending order): -def-spec (36%), -def+spec (33%), +def+spec (19%), +def-spec (10%).

7. Complementizer vs. *Wh*-Pronoun Relatives: Summary

When all the quantitative patterns for *co* and *który* clauses are compared, one notes, in *co* relatives, the predominance of definite NPs (83%) and specific referents (85%); on the other hand, in *który* relatives, one notes the predominance of indefinite NPs (70%) and a fairly even contribution of specific and non-specific referents. These proportions are represented graphically in Figures 1 and 2.

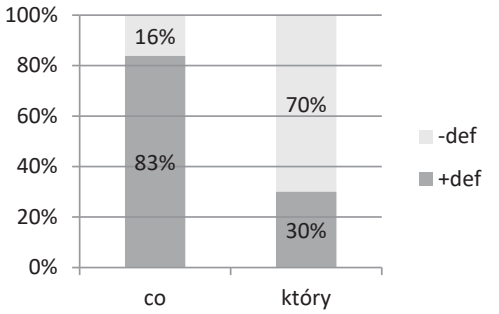


Figure 1. Definiteness in *co* and *który* relatives

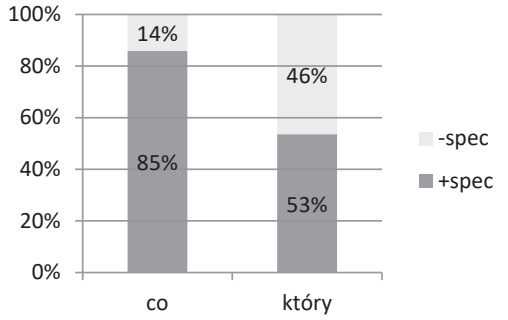


Figure 2. Specificity in *co* and *który* relatives

The preferential association of *co* relatives with definiteness (and specificity) and the relatively infrequent explicit definiteness in *który* clauses may be related to the distinct statuses of the two relativizers. *Który*, as a true *wh*-pronoun, carries φ features and case to unambiguously relate the head with whatever proposition is expressed in the relative clause; on the other hand, *co*, as a complementizer, carries neither φ features nor case and thus needs to employ other means of indicating a subordination link between the head and the relative clause. As resumptive pronouns are often absent (cf. Hladnik 2015), this is achieved by means of demonstratives, which point forward to the *co* clause, thus consolidating the subordination connection between the two parts of the construction. Compared to *wh*-pronoun relatives, this connection is gapless, more linear, and somehow looser for the lack of the dense network of agreement features present in *wh*-pronoun relatives.

8. Definiteness, Specificity, and (Ir)realis Mood

Given the preference of *co* relatives to be used with definite and specific NPs, one would expect few *co* relatives with the indefinite pronoun *jakiś* "some." In Spokes there are fifteen (2%). Also, given that *jakiś* may have specific or non-specific readings, (as in [24] and [25] respectively) we would expect non-specific readings to be the less frequent possibility. This is the case in Spokes: there are nine examples of *jakiś X* with a [-def+spec] reading (24) and six with a [-def-spec] reading (25).

(24) **jakieś** **ciecie** kurde ze szkół (Spokes; -def+spec)
some lowlifes EXPL from schools
średnich **co** nas naciągnęli na
secondary CO us tricked-3PL on
drogie wino
expensive wine
"Some lowlifes from secondary schools that tricked us into getting expensive wine."

(25) Widziałeś **jakichś** **młodych** **co** dookoła (Spokes, -def-spec)
saw-2SG any young-ACC-PL CO around
mają ciotki?
have-3PL aunts
"Have you seen any young people that have aunts around them?"
(speakers discuss seating arrangements at a reception)

In this section we take a closer look at some restrictions in the use of determiners in *co* relatives and note that some of these restrictions do not apply in *który* relatives. As it turns out, this is related to the realis/irrealis mood distinction in the relative clause VP.

Consider the three sentence templates (i)–(iii), each imposing different def/spec interpretations depending on the referent's identifiability or lack thereof.

(i) . . . książka, którą/co kupiłem wczoraj
". . . book which/that I bought yesterday"
(identifiable to speaker, and possibly to addressee; thus ±def+spec)

(ii) . . . książka, którą/co mi dałeś wczoraj
". . . book which/that you gave me yesterday"
(identifiable to speaker and addressee; thus +def+spec)

(iii) . . . książka, którą/co będziesz mógł czytać w pociągu
". . . book which/that you will be able to read on the train"
(neutral as to identifiability; thus ±def±spec)

We will use these templates to establish the extent to which three different determiners (*ten* "this," *taki* "such/this [-def]," *jakiś* "some") are permissible with *który* and *co*. As can be seen in (ia)–(iiia), when the *wh*-pronoun is used, the [±def+spec] template in (ia) is not compatible with the [-def-spec] determiner, and the [+def+spec] template in (iia) is not compatible with the [-def+spec] and [-def-spec] determiners. This is because it would simply be illogical to use [-def] determiners in a context which presupposes the familiarity and identifiability of the referent.

(ia)	Przeczytałem	tą/	taką/	*jakąś[6]	książkę,	którą	(constructed)
	read-1SG	this	such	some	book	which	
	kupiłem	wczoraj.					
	bought-1SG	yesterday					

"I read that [+def+def]/this [-def+spec]/*some book which I bought yesterday."

(iia)	Przeczytałem	tą/	*taką/	*jakąś	książkę,	(constructed)
	read-1SG	this	such	some	book	
	którą	mi	dałeś	wczoraj.		
	which	me	gave-2SG	yesterday		

"I read that [+def+def]/*this [-def+spec]/*some book which you gave me yesterday."

6 *Jakąś* is ungrammatical and indeed illogical here as a truly [-spec] determiner. However, it is acceptable in a slightly different use meaning "a particular book, but I will not go into the details."

(iiia)	Weźmy	tą/	taką/	jakąś	książkę,		(constructed)
	take-1PL-IMP	this	such	some	book		
	którą	będziesz	mógł				
	which	be-2SG-FUT	can				
	czytać	w	pociągu.				
	read	in	train				

"Let's take that [+def+spec]/this [-def+spec]/some book which you will be able to read on the train."

The same illogicality holds for *co*, restricting the use of [-def] determiners in (ib) and (iib). However, with *co* relatives, there are even more restrictions, and the use of demonstratives produces the best results. In (iiib) all three determiners sound awkward in combination with *co* and the trigger seems to be that the *co* clause itself (*co będziesz mógł czytać w pociągu*) lacks the definiteness and specificity that constitutes the prototypical environment for *co* relatives.

(ib)	Przeczytałem	tą/	taką/	*jakąś[7]	książkę,	co	(constructed)
	read-1SG	this	such	some	book	CO	
	kupiłem	wczoraj.					
	bought-1SG	yesterday					

"I read that [+def+def]/this [-def+spec]/*some book (that) I bought yesterday."

(iib)	Przeczytałem	tą/	*taką/	*jakąś	książkę,	(constructed)
	read-1SG	this	such	some	book	
	co	mi	dałeś	wczoraj.		
	CO	me	gave-2SG	yesterday		

"I read that [+def+def]/*this [-def+spec]/*some book (that) you gave me yesterday."

(iiib)	Weźmy	?tą/	?taką/	?jakąś	(constructed)
	take-1SG-IMP	this	such	some	
	książkę,	co	będziesz	*mógł*	
	book	CO	be-2SG-FUT	can	
	(ją)	*czytać*	*w*	*pociągu.*	
	(her)	read	in	train	

"Let's take that [+def+spec]/this [-def+spec]/some book which you will be able to read on the train."

7 See fn. 6.

Even with the demonstrative *tą* in (iiib), the sentence is infelicitous because the *co* clause itself does not provide the required "concreteness." This indicates that the presence of a demonstrative is not the sole component of a prototypical *co* relative; another one is the "actualness" of an event—the realis of it—expressed in the VP. Compare templates (i)–(ii) to template (iii). Templates (i) and (ii) refer to actual past actions in which the speaker—and the hearer in (ii)—were involved; real events that actually happened and may be recounted, as opposed to the mere possibility of an event in (iii). In cases of non-actual (irrealis) events such as (iiib), *który* seems to work better. Consider another example in (26a–d).

(26) (a) to jest dom **który** **ma** **wyglądać** jak statek (Spokes)
this is house which is to look like ship
"This is a house which is to look like a ship."

(b) ?to jest dom **co** **ma** **wyglądać** jak statek (modified)
this is house CO is to look like ship

(c) to jest dom **co** **wygląda** jak statek (modified)
this is house CO looks like ship
"This is a house that looks like a ship."

(d) **kupiłem** dom **co** **wygląda** jak statek (modified)
bought-1SG house CO looks like ship
"I bought a house that looks like a ship."

Note that in each case the head is a bare noun so that the use of determiners is irrelevant: the referent is -def+spec in each case. What makes the difference is that the different contexts supply varying levels of whether the proposition is actually valid. Note that (26c) sounds better than (26b) because it is less tentative and vague (i.e., more realis). In turn, example (26d) sounds better than (26c) because the existential *to jest* "this is" is replaced with a more concrete event of buying a house—an event that actually happened. The more real or factual the event, the better a *co* relative sounds. In sum, *co* seems better suited for definite and specific referents involved in actual real events rather than situations whose occurrence is not certain or hypothetical, or propositions whose validity is uncertain.

Following Rijkhoff and Seibt (2005), we assume here that there is a parallel between the definite/indefinite and specific/non-specific distinctions at the level of NPs and the realis/irrealis distinction at the level of the clause. For example, the (in)definiteness and (ir)realis of referents/events rely on the use of similar lexical and grammatical modifiers that specify the properties of referents/events: localizing modifiers (demonstratives and

relative clauses for NPs, adverbials and tense markers for VPs), quantifying modifiers (numerals for NPs, iterative aspect for VPs), qualifying modifiers (adjectives for NPs, verbal aspect and adverbs of manner for VPs) (Rijkhoff and Seibt 2005, 88). Given this parallel, the correlation in (26a–d) between nominal definiteness and verbal realis is not surprising.

Consider another example in which the realis/irrealis mood distinction is relevant. In (27a) the original NP *ewentualnych szkód* "potential damage, GEN PL" is [-def-spec] and suitably paired with an irrealis event in a *który* clause. The same combination of [-def-spec] and irrealis is incompatible with *co* in (27b). In (27c), the original NP is replaced with *ostatnich szkód* "recent damage, GEN PL," which is [-def+spec] and paired with a realis clause, thus providing a more suitable environment for a *co* relative. Note that (27c) sounds better than (27b) even without the optional demonstrative, which indicates that it is the introduction of a realis event that saves the *co* relative.

(27) (a) ubezpieczenie akademika od (Spokes)
insurance dorm-GEN from
ewentualnych **szkód** **które**
potential damage which
poczynisz w akademiku
cause-2SG-FUT in dorm
"Insurance of the dorm covering potential damage which you will/might cause in the dorm."

(b) ?ubezpieczenie akademika od **ewentualnych** (modified)
insurance dorm-GEN from potential
szkód **co** **poczynisz** w akademiku
damage CO cause-2SG-FUT in dorm

(c) ubezpieczenie akademika od (tych) **ostatnich** (modified)
insurance dorm-GEN from (these) recent
szkód **co** **poczyniłeś** w akademiku
damage CO caused-2SG in dorm
"Insurance of the dorm covering (that) recent damage that you caused in the dorm."

9. Conclusions

Using quantitative and qualitative analysis of authentic and constructed data, the study shows the relevance of definite/indefinite and specific/non-specific reference in *wh*-pronoun and complementizer relative clauses in Polish. Specifically, in *co* relatives, one notes the predominance of definite NPs (83%) and specific referents (85%), thus

indicating a strong preference for *co* relatives to be associated with heads of relatively high individuation or referentiality. The most frequent marker of definiteness—a demonstrative—may be seen as performing another function: to point forward to the *co* clause, thus indicating the subordination link between the head and the relative clause (recall that, unlike *który*, *co* carries neither φ features nor case, and that resumptives are often absent). As has been illustrated, as a result of the preference of *co* relatives for definite and specific heads, the use of the construction may be infelicitous in contexts where indefinite and/or non-specific heads are involved.

On the other hand, in *który* relatives, one notes the predominance of indefinite NPs (70%) and a fairly even contribution of specific and non-specific referents. This contrasts sharply with *co* relatives. While *co* works better with definite and specific NPs, *który* can be readily used with indefinite and non-specific NPs, although it is also perfectly compatible with definite and specific heads, and may always be used as a *wh*-pronoun replacement for *co*.

We have also established a link between (in)definiteness/(non-)specificity and (ir)realis mood. Nominal definiteness/specificity patterns with clausal realis, while indefiniteness/non-specificity patterns with clausal irrealis. The prototypical environment for *co* relatives is thus realis events referred to in the relative clause, while *który* may be readily used with both realis and irrealis events.

Works Cited

Bondaruk, Anna. 1995. "Resumptive Pronouns in English and Polish." In *Licensing in Syntax and Phonology. PASE Studies and Monographs*, vol. 1, edited by Edmund Gussmann, 27–55. Lublin: Folium.

Bošković, Željko. 2009. "On Relativization Strategies and Resumptive Pronouns." In *Studies in Formal Slavic Phonology, Morphology, Syntax, Semantics and Information Structure*. *Proceedings of FDSL 7*, edited by Gerhild Zybatow, Uwe Junghanns, Denisa Lenertova, and Petr Biskup, 79–92. Peter Lang.

Fodor, Janet, and Ivan Sag. 1982. "Referential and Quantificational Indefinites." *Linguistics and Philosophy* 5 (3): 355–98.

Fried, Mirjam. 2010. "Accusative Resumptive Pronoun in Czech Relative Clauses with Absolutive Relativizer *Co*." *Korpus, Gramatika, Axiologie* 1 (1): 16–29.

Fried, Mirjam. 2011. "Grammatical Analysis and Corpus Evidence." In *Grammar and Corpora* 3, edited by Marek Konopka, Jacqueline Kubczak, Christian Mair, František Štícha, and Ulrich H. Waßner, 63–86. Mannheim: Narr Verlag.

Gołąb, Zbigniew, and Victor A. Friedman. 1972. "The Relative Clause in Slavic." In *The Chicago Which Hunt: Papers from the Relative Clause Festival*, edited by Paul M. Peranteau, Judith N. Levi, and Gloria C. Phares, 30–46. Chicago Linguistic Society.

Hladnik, Marko. 2015. “Mind the Gap. Resumption in Slavic Relative Clauses.” PhD diss., LOT Dissertation Series, 390. Utrecht University.

Kardela, Henryk. 1986. *Wh-Movement in English and Polish. Theoretical Implications.* Lublin: UMCS.

Lindvall, Ann. 1996. “Definite Marking and Referential Status in Greek, Swedish and Polish.” *Working Papers* 45: 113–32. Lund University, Dept. of Linguistics.

Lyons, Christopher. 1999. *Definiteness.* Cambridge: Cambridge University Press.

McDaniel, Dana, and Dorota Lech. 2003. “The Production System’s Formulation of Relative Clause Structures: Evidence from Polish.” *Language Acquisition* 11 (2): 63–97.

Minlos, Philip R. 2012. “Slavic Relative *Čto*/*Co*: Between Pronouns and Conjunctions.” *Slověne. International Journal of Slavic Studies* 1 (1): 74–91.

Murelli, Adriano. 2011. *Relative Constructions in European Non-Standard Varieties.* Berlin/New York: Mouton de Gruyter [Empirical Approaches to Language Typology 50].

Rijkhoff, Jan, and Johanna Seibt. 2005. “Mood, Definiteness and Specificity: A Linguistic and a Philosophical Account of Their Similarities and Differences.” *Tidsskrift for Sprogforskning* 3 (2): 85–132.

Rudin, Catherine. 1986. *Aspects of Bulgarian Syntax: Complementizers and WH Constructions*. Columbus: Slavica.

Szczegielniak, Adam. 2006. “Two Types of Resumptive Pronouns in Polish Relative Clauses.” In *Linguistic Variation Yearbook 2005*, edited by Pierre Pica, Johan Rooryck, and Jeroen van Craenenbroeck, 165–85. John Benjamins.

Corpora

Pęzik, Piotr. 2015. “Spokes—A Search and Exploration Service for Conversational Corpus Data.” In *Selected Papers from the CLARIN 2014 Conference*, October 24–25, 2014, Soesterberg, The Netherlands, 99–109. Linköping University Electronic Press. Available online at http://spokes.clarin-pl.eu.

Word Study and the Lexicon: Phonological Approaches

Where's the Contrast? Discovering Underlying Representations with a Language Game

Joanna Zaleska

University of Leipzig, Germany

joanna.zaleska@uni-leipzig.de

Abstract: The aim of this article is to take a fresh look at the long-standing controversy regarding the phonemic status of Polish high unrounded vowels, front [i] and central [ɨ]. The previous literature suggests two competing hypotheses, based on the distribution of the vowels and their relation to palatalized and velarized consonants. One hypothesis holds that [i] and [ɨ] are derived from the same underlying segment, while the other one views these vowels as underlyingly distinct. I show that the two approaches predict different outcomes if a sequence of a consonant followed by a high unrounded vowel is split up. I report the results of a pilot study that tests these predictions using a version of Pig Latin. These results provide support for the hypothesis that [ɨ] and [i] are underlyingly distinct.

Keywords: high central vowel; artificial language game; surface palatalization; Polish

1. Introduction

For over a hundred years, the number of contrastive vowel segments in Polish has been the subject of continuing controversy among scholars working on Polish phonology. The debate has centred on the representation of two vowels in the high region, front [i] and central [ɨ]. Since the two segments are in full complementary distribution, some linguists have argued that they are positional variants of a single underlying vowel, whose quality depends on the left-hand context. Others have pointed to the role that [i] and [ɨ] play in palatalization processes, arguing that the two vowels must be viewed as underlyingly distinct and that the quality of the consonants occurring to the left of these vowels is therefore the effect, and not the source, of the underlying quality of the vowel.

Thus far, arguments for either of the two views have been based on internal evidence, such as phonotactics or morphophonological processes. In this article, I suggest a way to

discriminate between the two positions using an artificial language game. I report on the results of a pilot study providing preliminary evidence for the hypothesis that the two vowels are underlyingly distinct. Additionally, they suggest that the lack of word-initial [ɨ]'s may be an accidental gap and as such should not be reflected in models of Polish speakers' phonological knowledge.

The article is organized as follows. In Section 2, I present the background information about the phonetics and distribution of [i] and [ɨ], as well as of "soft" and "hard" consonants, which have been argued to govern the quality of the following high unrounded vowel. In Section 3, I provide an overview of selected previous studies that have addressed this issue, focussing on their approach to the underlying representation of morpheme internal sequences of consonants followed by a high unrounded vowel. On that basis, I formulate the predictions they make for the results of a language game in which these sequences are split. The following sections describe an experiment I conducted to test these predictions. Section 4 outlines the method and Section 5 presents and discusses the results. Section 6 briefly concludes and presents directions for future research.

2. Background: Polish [+back] and [−back] Segments

2.1 Front and Back Vowels

The surface vowel inventory of Polish contains six oral monophthongs: three high vowels, [i, ɨ, u], two mid vowels, [ɛ, ɔ], and one low vowel, [a]. The focus of this article is on the two unrounded vowels in the high region, [i] and [ɨ], spelled *i* and *y*, respectively. The [i] vowel is articulated with the body of the tongue strongly raised and pushed forward, whereas [ɨ] is described as a fronted central vowel (Biedrzycki 1974, 61). Although phonetically fronted, [ɨ] patterns phonologically with back vowels. Consequently, the distinction between [i] and [ɨ] is normally expressed in terms of the feature [±back], with [i] classified as [−back] and [ɨ] classified as [+back].

Of most interest to the present discussion are the contexts in which the two vowels appear. As can be seen in Table 1, [i] and [ɨ] stand in complementary distribution. Word-initially, only [i] is permitted. Within words, [i] and [ɨ] follow disparate classes of sounds. The front [i] vowel may appear after "soft" (i.e., palatalized) consonants as well as prepalatals, whereas [ɨ] may only appear after "hard" consonants.

	[i]	**[ɨ]**
Word-initially	*ił* [**i**w] "loam"	—
After soft consonants	*miła* [**mʲi**wa] "nice" (FEM)	—
After prepalatals	*siła* [**ɕi**wa] "strength"	—
After hard consonants	—	*myła* [**mɨ**wa] "she washed"

Table 1. The distribution of [i] and [ɨ] in Polish

This complementary distribution of [i] and [ɨ] has led some scholars to conclude that the two vowels are in fact surface manifestations of the same underlying segment, whose quality depends on the preceding context. It is this hypothesis that forms the first point of departure for the present study. The second one is related to the distribution of Polish plain and palatalized consonants, detailed in the following section.

2.2 "Soft" and "Hard" Consonants

Polish has a rich surface inventory of consonants, shown in Table 2. The observation relevant to the present discussion is that all Polish consonants except prepalatals have two variants that differ in terms of secondary articulation.[1] In Slavic linguistics, these are traditionally referred to as "hard" (on the left of each pair below) and "soft" (on the right) consonants.

	Labial	**Dental**	**Post-alveolar**	**Pre-palatal**	**Velar**
Stop	p–pʲ b–bʲ	t–tʲ d–dʲ			k–kʲ g–gʲ
Affricate		ʦ–ʦʲ ʣ–ʣʲ	ʧ–ʧʲ ʤ–ʤʲ	ʨ ʥ	
Fricative	f–fʲ v–vʲ	s–sʲ z–zʲ	ʃ–ʃʲ ʒ–ʒʲ	ɕ ʑ	x–xʲ
Nasal	m–mʲ	n–nʲ		ɲ	
Approximant		l–lʲ r–rʲ			

Table 2. Polish surface consonant inventory (simplified)

The "hard" series is pronounced with a weak velarization gesture, in which the body of the tongue is retracted to the same position as for the vowel /a/ (Wierzchowska 1963). Consonants in the "soft" series are produced with the body of the tongue moved forwards and raised, as with the vowel /i/ (Wierzchowska 1971). [ɕ ʑ ʨ ʥ ɲ] do not have palatalized correspondents since they are inherently soft themselves: They are articulated in the prepalatal region, with the body of the tongue raised towards the hard palate.

1 The soft variants of postalveolar consonants only occur in unassimilated borrowings, e.g., *Chicago* [ʧʲikagɔ] "Chicago," *dżinsy* [ʤʲinsɨ] "denims," *szisza* [ʃʲiʃa] "sheesha," *reżim* [rɛʒʲim] "regime," and across word boundaries, e.g., *kosz Janka* [kɔʃʲ janka] "Janek's basket."

The distinction between hard and soft consonants is usually made in terms of the feature [±back].[2] Hard consonants are specified as [+back], whereas soft consonants (both those with secondary palatalization and prepalatals) are [−back] and additionally [+high].[3] This means that there are no consonants in Polish which are neutral with respect to [±back].

What is most pertinent to the present discussion is again the distribution of the two classes of consonants. The soft and hard equivalents occur in mutually exclusive contexts. As shown in Table 3, soft consonants may appear only before high front vocoids, [i] and [j].[4] Hard consonants occur elsewhere, that is, before other vowels ([a ɛ ɔ u ɨ]), before consonants and word-finally.

	Soft consonants	**Hard consonants**
Before *i, j*	*sinus* [**sʲi**nus] "sine" *pasja* [pa**sʲj**a] "passion"	—
Before other vowels	—	*synuś* [**si**nuɕ] "son" (DIM) *sosen* [**sɔsɛ**n] "pines" (GEN)
Before consonants	—	*smok* [**sm**ɔk] "dragon" *pasta* [pa**st**a] "paste"
Word-finally	—	*lis* [lʲi**s**] "fox" *czas* [ʧa**s**] "time"

Table 3. The distribution of soft and hard consonants in Polish

Note that this observation does not hold for prepalatals, which may occur in all the contexts in which hard and other soft and hard consonants do: word-finally, before consonants and most vowels.[5] Since prepalatals behave differently from the remaining

2 The Clements-Hume model (Clements 1989; Hume 1992, 1996; Clements and Hume 1995), where privative [coronal] and [dorsal] features under the V-place node are used to express secondary articulations and the frontness/backness of vowels, is considered and ultimately rejected by Rubach (2007), who argues that the model is unable to account for Polish palatalization facts.

3 Since the specification for [±high] in every consonant can be inferred from its specification for [±back], the following discussion focusses on the feature [±back], treating the value of [±high] as derivable.

4 As noted by Rubach (2003a, 604), both hard and soft velars may appear before vowels in the morpheme internal position. This contrast does not arise before either of the high unrounded vowels, the focus of this article. Consequently, I set this complication aside here but return to it briefly in Section 3.1.

5 The only restriction on the appearance of prepalatals is that they do not occur before the high central vowel [ɨ]. Since prepalatals are uncontroversially assumed to be contrastive, the studies that view [ɨ] and [i] as underlyingly separate uniformly attribute this gap to a fronting rule that

soft consonants, in the remainder of this article, the term "soft" will be used to refer to consonants with a secondary palatalization gesture only. While prepalatals are clearly contrastive, the complementary distribution of the remaining soft and hard consonants may suggest that the quality of palatalized and nonpalatalized consonants in a sequence of a soft/hard consonant followed by [i]/[ɨ] (CI sequence, henceforth) depends on the quality of the vowel. Probing this hypothesis is the second aim of this article.

3. Research Questions

3.1 Possible Underlying Specifications of [+back] and [−back] Segments

The complementary distribution of [i] and [ɨ] as well as of soft and hard consonants might be viewed as a reason to conclude that the pairs of segments are positional variants of a single underlying vowel, or consonant, respectively. It is clear, however, that this assumption cannot be made for *both* the vowels and the consonants at the same time. To see this, consider the pair of words in (1).

(1) (a) *miła* [**mʲi**wa] "nice" (FEM)

(b) *myła* [**mɨ**wa] "she washed"

If [mʲ] and [m] in (1) were derived from the same consonant and if [i] and [ɨ] were derived from the same vowel, the two words in (1) would be underlyingly identical. This would make it impossible to derive the two distinct surface forms. This example shows that if a [+back]–[−back] consonant pair is derived from the same segment, then the [+back]–[−back] vowel pair has to be underlyingly distinct. If, vice versa, a [+back]–[−back] vowel pair is derived from the same sound, the [+back]–[−back] consonants have to be specified as such at the underlying level.

After the possibility of deriving both the high unrounded vowels and hard/soft consonants each from a single underlying segment has been eliminated, three other options remain. These are (i) assuming that the two vowels are underlyingly the same but that hard and soft consonants are distinct, (ii) assuming, conversely, that it is the consonants that are derived from a single segment while the vowels are separate, and (iii) assuming that both the [+back]–[−back] vowels and [+back]–[−back] consonants are distinct units. As I will show below, each of these hypotheses has been entertained in the literature.

neutralizes the contrast. Because of this neutralization, sequences of prepalatals followed by high unrounded vowels were not investigated in the experiment reported here. As noted in Section 6, however, such forms might constitute a valuable diagnostic tool for identifying participants who exhibit orthographic bias. If prepalatals are included in future adaptations of this study, the restriction will have to be taken into consideration.

The idea that [i] and [ɨ] are realizational variants of the same segment can be traced back as early as the 19th century. For Baudouin de Courtenay, the two vowels (referred to as "i_1" and "i_2" are "general phonetic divergents" (1893, 27) of the same phoneme, *i mutabile*. Closer to the present day, Czaykowska-Higgins (1988, 149–50) argues that [ɨ] is a surface rendition of an underlying /i/, citing the complementary distribution of the two segments, and crucially, the lack of [ɨ] word-initially and after vowels, as the most important argument for this view. In accordance with the reasoning presented above, this leads her to postulate an underlying contrast between soft and hard consonants, at least for those exhibiting a surface [Cʲi]–[Cɨ] distinction, such as those in (1). These include other labials and (marginally) the velar fricative, on the basis of pairs such as *hymn* [xɨmn] "hymn" versus *Chiny* [xʲinɨ] "China" (1988, 152). The remaining velars, as well as dentals and postalveolars are not specified as [−back] since according to Czaykowska-Higgins, the quality of these consonants (and of the following high unrounded vowel) can always be predicted: velars are uniformly soft and followed by [i], whereas dentals and postalveolars are hard and followed by [ɨ].[6] She proposes that the quality of front and back segments is derived by means of the following rules (Czaykowska-Higgins 1988, 137, 145):

- Spreading the dorsal node (with the [−back] feature) from the vowel onto the preceding velar stop, driven by a phonotactic constraint against *[ki] and *[gi].
- Delinking the [−back] feature of the vowel if it follows a hard labial, a dental, a postalveolar or [x], driven by a constraint against the appearance of [i] in a syllable in which the onset is not [−back].
- Delinking the [−back] feature of a soft labial preceding other consonants and word-finally. Note that in this analysis, the soft-hard contrast is not neutralized before vowels. However, as noted above, soft consonants are not permitted before segments other than [i] and [j]. Czaykowska-Higgins (1988, 137) solves this problem by arguing that in the course of the derivation, a [j] glide is inserted between front consonants and vowels other than [i].

Thus, under this view, the words in (1) are underlyingly //mʲiw+a//,[7] with a soft labial for (1a) and //miw+a//, with a hard labial, for (1b). The underlying segments surface unchanged in *miła* [**mʲi**wa] but in *myła* [**mɨ**wa], the [−back] feature of the vowel is delinked, resulting in retraction of the vowel.

6 Czaykowska-Higgins (1988) does not address the question how the quality of initial consonant-vowel sequences should be derived in borrowings such as *tir* [tʲir] "heavy goods vehicle," *czipsy* [ʧʲipsɨ] "crisps" or *kynologia* [kɨnɔlɔgʲja] "cynology." Consequently, I sidestep this issue in my brief summary of her work.

7 I ignore here the underlying representation of the labiovelar glide, which is usually assumed to be //ɬ//.

The opposite view is espoused by Gussmann (1980a, b) and Rubach (1984), who recognize /ɨ/ as a separate underlying segment and treat the soft consonants as allophones of their plain equivalents, derived by a palatalization rule spreading [−back] from the following high vocoid. More recently, Rydzewski (2014, 2016, 2017) argues that despite [ɨ] and [i]'s fully complementary distribution, an explanatorily adequate analysis of the processes of Coronal and Velar Palatalization in Polish requires the assumption that the two segments are underlyingly distinct. Under this view, the underlying representations of the words in (1) are //miw+a// for (1a) and //mɨw+a// for (1b). The difference between the two forms lies in the quality of the high unrounded vowel. Both words begin with a hard consonant, but in (1a), the following front vowel triggers softening. It must be noted that as it stands, this analysis fails to account for the limited distribution of [ɨ], for example the lack of [ɨ] after prepalatals, velar stops and at the beginning of words. Gussmann (1980a, 89) explains the first two distributional patterns by postulating a fronting rule that changes /ɨ/ into [i] after palatals and velar plosives. Rubach (1984, 152–57) argues that the rule is cyclic and thus does not apply morpheme-internally. Consequently, he treats the lack of [kɨ], [gɨ] sequences as an accidental, though historically motivated, gap. No phonological rules proposed by Gussmann or Rubach account for the lack of word-initial [ɨ]'s, however. It is not clear whether the authors treat this gap as accidental or whether the omission is related to the thematic focus of their studies.

Finally, some recent analyses of Polish combine the two assumptions mentioned above. Rubach (2003a; 2003b; 2007) treats [ɨ] and [i] as underlyingly distinct segments. However, he additionally argues that backness has to be contrastive in the class of labials and velar stops, while the remaining consonants (save for prepalatals) are underlyingly hard.[8] As far as velars are concerned, he reaches this conclusion on the basis of pairs such as *kiedy* [kʲɛdɨ] "when" versus *kelner* [kɛlnɛr] "waiter" and *giermek* [gʲɛrmɛk] "henchman" versus *gest* [gɛst] "gesture" (Rubach 2003a, 604). The distinction between soft and hard labials, on the other hand, is important for morphology, as it determines allomorph selection. As shown in (2), the nominative plural suffix can take one of two forms, [ɨ] and [ɛ].

(2) Phonologically driven allomorph selection in Polish (Rubach 2007, 109)
(a) *kot* [kɔt] "cat" (NOM SG) – *koty* [kɔtɨ] (NOM PL)

(b) *struś* [struɕ] "ostrich" (NOM SG) – *strusie* [struɕɛ] (NOM PL)

8 Rubach (2003b) additionally assumes that prepalatals are derived from [−back] alveolars. Since prepalatals are not discussed in this study, I ignore this complication.

With stems ending in a labial consonant, the [ɨ] suffix is selected for some words (3a) and the [ɛ] suffix for others (3b).

(3) Soft and hard labial stems in Polish (Rubach 2003a, 617)
(a) *trup* [trup] "corpse" (NOM SG) – *trupy* [trupɨ] (NOM PL)

(b) *karp* [karp] "carp" (NOM SG) – *karpie* [karpʲjɛ] (NOM PL)

Rubach (2003a, 2007) accounts for this by assuming that stems such as those in (3b) end in a labial specified as [−back]. Soft labials depalatalize word-finally, so the contrast is neutralized in the nominative singular forms. Like in Czaykowska-Higgins's (1988) analysis, the soft–hard contrast is not neutralized if a vocalic ending is added. However, before vowels, soft labials split into a sequence of a labial followed by a front glide [j] (with the frontness spreading onto the preceding consonant at a later derivational stage). Consequently, on the surface, the distinction is realized as one between a hard consonant versus a soft consonant followed by a front glide. Some examples are given in (4).

(4) Surface realization of the soft–hard labial contrast before vowels
(a) *jedwab* [jɛdvap] "silk" (NOM SG) – *jedwabiu* [jɛdva**bʲju**] (GEN SG), VS
grab [grap] "hornbread" (NOM SG) – *grabu* [gra**b**u] (GEN SG),

(b) *szczaw* [ʃʧaf] "sorrel" (NOM SG) – *szczawiowi* [ʃʧa**vʲjɔ**vʲi] (DAT SG), VS
staw [staf] "pond" (NOM SG) – *stawowi* [sta**v**ɔvʲi] (DAT SG)

(c) *modrzew* [mɔdʒɛf] "larch" (NOM SG) – *modrzewiem* [mɔdʒɛ**vʲj**ɛm] (INST SG), VS
krzew [kʃɛf] "shrub" (NOM SG) – *krzewem* [kʃɛvɛm] (INST SG)[9]

Assuming that not only [ɨ] and [i] but also hard and soft consonants are underlyingly distinct means that there are now four possible ways to represent the initial CI sequence in the words in (1): //mʲi . . .//, with two [−back] segments, //mɨ . . .//, with two [+back] segments, //mi . . .// and //mʲɨ . . .//, with different combinations of segments disagreeing in terms of backness. The greater number of possible underlying forms leads to some indeterminacy of underlying representations, in the sense that each of the forms in (1) may be represented in more than one way. It does not, however, create any ambiguity, as none of the underlying forms could be successfully mapped onto more than one of the two surface forms. The underlying forms in which the initial segments agree in

9 Note that the contrast is neutralized before the locative [ɛ] suffix, which triggers palatalization of hard labials, e.g., *grabie* [gra**bʲj**ɛ] "hornbread" (LOC SG), *stawie* [sta**vʲj**ɛ] "pond" (LOC SG), *krzewie* [kʃɛ**vʲj**ɛ] "shrub" (LOC SG).

terms of backness correspond to identical surface forms. In Rubach's (2003a) optimality theoretic analysis, the underlying form in which a hard consonant is followed by a front vowel must be mapped onto a [−back] sequence, whereas the one in which a soft consonant is followed by [ɨ] must be mapped onto a [+back] sequence, due to a high-ranked IDENT-V([−bk]) constraint, which, as Rubach shows, is independently necessary to account for a range of palatalization effects in Polish.[10] As far as the forms in (1) are concerned, the analysis that assumes contrastive hard and soft consonants in addition to distinct [ɨ] and [i] turns out to be equivalent to the one that only assumes [+back]–[−back] in the vowels. As can be seen in (5), the quality of the initial sequence depends on the underlying quality of the high unrounded vowel, while the quality of the consonant is immaterial.

(5) Possible underlying representations for the forms in (1) in Rubach's (2003a) system
(a) *miła* [mʲiwa]: //mʲiwa// or //miwa//

(b) *myła* [mɨwa]: //mʲɨwa// or //mɨwa//

To sum up, what really distinguishes the analyses reviewed here is whether they adopt the assumption that [ɨ] and [i] are underlyingly distinct. If they do, then it is the underlying quality of the vowel that governs the surface [+back]–[−back] quality of a CI sequence, irrespective of whether hard and soft consonants are also contrastive. If, on the other hand, the [ɨ] vowel is viewed as a positional variant of /i/, the quality of the sequence is dictated by the underlying quality of the consonant. Thus far, arguments for either of the two hypotheses have been based on internal evidence, such as phonotactics or morphophonological processes. If language-internal evidence is inconclusive, external evidence (such as word games, speech errors or psycholinguistic experiments; see Ohala 1986 for an overview) may be brought to bear on competing analyses. In the following section, I discuss an experimental study based on a language game, which may help discriminate between the two views.

3.2 Testing the Representation of [i] and [ɨ] Using Pig Latin

Under two of the hypotheses discussed above, one pair of sounds is treated as realizational variants of a single underlying segment, whose quality depends on the context. If the context is removed, the segment is expected to lose that quality. However, such a

10 Recapping Rubach's (2003a) Optimality Theory analysis and extending it to the forms containing a CI sequence falls beyond the scope of the present paper. The reader is encouraged to verify the evaluation of these forms, noting that the labial will become soft at Level 3 in Rubach system, and that fission must be blocked by a high-ranked constraint absent from Rubach's ranking, possibly one that bans the [jɨ] sequence (which is indeed illicit in Polish).

test is not possible for Polish CI sequences, since the context for the putative allophonic variant is a feature in an adjacent (preceding or following) segment. In order to remove the context, it would be necessary to split up the sequence by deleting or displacing one of the segments. No morphological process has this effect in Polish. As a result, morpheme-internal CI sequences surface in the same form in all occurrences of the morpheme.

The third hypothesis assumes that both classes of segments carry the [±back] contrast but the surface quality of CI sequences is governed by the underlying quality of the high unrounded vowel. Consequently, if the consonant is removed from a CI sequence, the vowel is expected to retain its [±back] specification. The expectations concerning the quality of the consonant in a CI sequence after the removal of the vowel are less clear. As noted above, the underlying representations of these consonants are non-unique. If speakers have access to both potential input representations, variation is expected. A more likely scenario, however, assumes that speakers postulate a unique underlying representation for each form. Here, two approaches are possible. One is to follow McCarthy (2005) in assuming that learners allow some nonalternating segments to take a "free ride" on a phonological rule that results in alternations elsewhere in the language. If this is the case, the segments that could be the effect of that rule are expected to change their value when the context is removed. Another approach assumes the optimality-theoretic principle of Lexicon Optimization (Prince and Smolensky 1993, 209), which states that in the absence of evidence to the contrary, speakers posit underlying forms that are as close to the input as possible. Under Lexicon Optimization, we predict that removing the vowel will have no effect on the quality of the preceding consonant in a CI sequence.

I propose to test these two hypotheses by teaching native speakers of Polish a language game that moves word-initial consonants away from the following vowel and then observing their responses when they are asked to modify a word beginning with a CI sequence. Transformational language games (or *ludlings*; Laycock 1972), which delete, replace or invert segments, have been employed as evidence in generative phonology since its inception in the 1960s (Chomsky and Halle 1968, 43; Bertinetto 1987; Derwing et al. 1988; Treiman 1983; Treiman and Danis 1988; Pierrehumbert and Nair 1995). As noted by Guimarães and Nevins (2013, 157), they offer the possibility to disrupt a phonological string in a way that may reveal the underlying representation of segments that make up that string. Here, the appropriate tool is Pig Latin, a game which involves moving the word-initial onset to the end of the word and suffixing a vocalic ending. For example, the game changes the Polish word *droga* [drɔga] into [ɔgadru].[11]

11 In the original, English, version of Pig Latin, the vocalic ending added to the modified word is [eɪ]. This was changed to [u] in the Polish version to make the result sound more natural to Polish speakers and to ensure that potential softness of the preceding consonant has a clear source, in the sense that it cannot be attributed to the frontness of the suffix vowel (cf. footnote 9).

Pig Latin can help discriminate between the hypotheses in two ways. First, it could be applied to words beginning with a hard consonant followed by [ɨ]. Looking at the words in (1) again, the possible outcomes of applying the game to *myła* [mɨwa] "she washed" are as in (6).

(6) Possible outcomes for *myła* [mɨwa] "she washed"
(a) [ɨwamu]: inconsistent with the hypothesis that [ɨ] is derived from //i//

(b) [iwamu]: consistent with all hypotheses

Under the hypothesis that [ɨ] is a positional variant of /i/ that only appears after hard consonants, we expect the vowel to return to its underlying quality if the context, i.e., the hard consonant, is taken away. Thus, if the underlying representation of *myła* is //miw+a//, as in Czaykowska-Higgins's (1988) analysis, we expect the game to transform the word into [iwamu], (6b). If the speakers pronounce the word as [ɨwamu], (6a), the hypothesis would be falsified. An opposite result, however, is in line with all the hypotheses. It is, of course, the expected result in Czaykowska-Higgins's (1988) analysis. Under the hypothesis assuming that the underlying representation of the word is //mɨw+a//,[12] on the other hand, we may expect that the speakers produce [ɨwamu]. However, recall that [ɨ] never occurs word-initially in Polish and that this gap is not addressed in the analyses reported here. Recent experimental studies (e.g., Becker et al. 2011; Dawdy-Hesterberg 2014) indicate that the fact that a phonotactic regularity can be found in the lexicon does not necessarily mean that speakers have tacit knowledge of that regularity. Thus, the lack of word-initial [ɨ]'s in Polish could be an accidental gap. If this is the case, speakers are expected to produce [ɨwamu] as the output of the game. However, if the gap forms part of the speakers' phonological knowledge, they might try to repair the [ɨ]-initial outputs, possibly by mapping the first vowel onto one that is allowed at the beginning of the word. Two phonetically near candidates are the vowels [ɛ] and [i]. If the latter is used, the result of the game will be undistinguishable from the one predicted by Czaykowska-Higgins's (1988) analysis.

The game can also be applied to words that begin with a soft consonant followed by [i]. Here, again two different results can be expected, shown in (7).

12 The potentially available underlying representation containing a soft consonant, i.e., //mʲɨw+a// is excluded both under the free-ride principle (because no alternations exist that would justify postulating a rule deriving [m] from //mʲ// before //ɨ//) and under Lexicon Optimization (because given the lack of alternations, the speaker is expected to posit an input that is identical to the output).

(7) Possible outcomes for *miła* [mʲ**i**wa] "nice" (FEM)
   (a) [iwamʲju]: inconsistent with the hypothesis that [−back] consonants are realizational variants of [+back] ones and with the hypothesis that both classes carry the [±back] contrast, assuming the free-ride principle

   (b) [iwamu]: inconsistent with the hypothesis that consonants bear the [±back] contrast (and that Lexicon Optimization holds)

If soft consonants in a Cʲi sequence are underlyingly [−back] (irrespective of whether [ɨ] and [i] are underlyingly distinct or not), we expect their quality to be retained in the outcome of the game (and, additionally, spawn a front glide after the consonant and before the affix *u*), as in (7a),[13] in two cases: (i) if [ɨ] and [i] are not underlyingly distinct, and (ii) if [ɨ] and [i] are underlyingly contrastive and Lexicon Optimization holds. If, on the other hand, the consonants are underlyingly hard with their softness coming from the following vowel, they should be pronounced as [+back], as in (7b), when followed by the back [u] suffix. The same result is expected if both classes carry the [±back] contrast and the free-ride principle applies. This is because in Polish, there exists a process that derives soft consonants from hard ones, causing alternations. Some examples are given in (8).

(8) Surface palatalization in Polish (Rubach 2003a, 611)
   (a) *krzew* [kʃɛ**f**] "shrub" (NOM SG) – *krzewić* [kʃɛ**v**ʲitɕ] "to promulgate"

   (b) *tom* [tɔ**m**] "volume" (NOM SG) – *tomik* [tɔ**m**ʲik] (DIM)

With the free-ride principle, the speakers assume that the rule applies in words with non-alternating soft CI sequences. They undo its effect when postulating the underlying forms, arriving at a hard underlying consonant.

4. Method

4.1 Participants

The participants were 20 volunteers, 10 female and 10 male, aged between 22 and 52 (mean age 31.2). They were all native speakers of Polish.

13 This is true for sequences containing soft labials or a soft velar fricative, but not necessarily for those that contain the soft velar stop. If, as assumed by Czaykowska-Higgins (1988), the softness of [kʲ] and [gʲ] is the result of spreading from the high front vowel, the expected result of the game in this case contains a hard consonant before the suffix.

4.2 Stimuli

The stimulus set consisted of nineteen Polish disyllabic words of the shape (C)CV.(C)CV. All of them were singular or plural nouns in nominative case. Ten of the experimental items began with one or two hard consonants followed by the high back vowel [ɨ]; nine began with a soft consonant followed by the high front vowel [i].[14] All experimental items are listed in Table 4.

Word	Transcription	Gloss	Word	Transcription	Gloss
wiza	[vʲiza]	"visa"	pyza	[pɨza]	"dumpling"
wiśnia	[vʲiɕɲa]	"cherry"	wydra	[vɨdra]	"otter"
misie	[mʲiɕɛ]	"teddy bears"	ptysie	[ptɨɕɛ]	"pastry puffs"
pikle	[pʲiklɛ]	"pickles"	życie	[ʒɨtɕɛ]	"life"
kiwi	[kʲivʲi]	"kiwi"	łyżki	[wɨʃkʲi]	"spoons"
figi	[fʲigʲi]	"figs"	bryki	[brɨkʲi]	"cars" (colloq.)
piwo	[pʲivɔ]	"beer"	pysio	[pɨɕɔ]	"muzzle" (DIM)
lisy	[lʲisɨ]	"foxes"	mydło	[mɨdwɔ]	"soap"
			cyfry	[tsɨfrɨ]	"digits"
bitwy	[bʲitfɨ]	"battles"	ryby	[rɨbɨ]	"fish" (PL)

Table 4. Experimental items

The 19 experimental items were randomly interspersed with 60 fillers (plus one item beginning with a prepalatal fricative, initially included in the list of words beginning with "soft" consonants). These words contained vowels other than [i] and [ɨ] in the initial syllable, but otherwise they had the same characteristics as the experimental items (disyllabic plural and singular nouns in the nominative, [C]CV.[C]CV shape). The list of all 80 items was randomized. The full list of items used in the experiment in the order in which they were presented is shown in the Appendix.

All the items were digitally recorded by a phonetically-trained female native speaker of Polish using Tascam DR-40 linear PCM recorder, with 44.1 kHz and 16 bit (mono).

4.3 Procedure

The experiment began with a training phase, in which the participants were trained to achieve fluency in the game. The phase consisted of several stages of increasing difficulty. In each stage, participants heard words provided verbally by the experimenter. These

14 Due to a mistake in the experimental design, the list of stimuli with "soft" consonants included one word beginning with a prepalatal fricative, *sito* [ɕitɔ] "sieve," item 37 in the Appendix). The item was removed from analysis.

were selected at random from a list containing words of different lengths in which the first syllable was headed by a vowel other than [ɨ] or [i]. Each stage was continued until the participants could respond correctly and without hesitation five times in a row. The first task, aimed to familiarize the participants with the term "syllable" and putting them at ease with the experimental setting, was dividing the heard words into syllables and clapping for each syllable. In the following stage, the participants continued to clap for each syllable in a given word but this time they only repeated the vowels. This was done to ensure that the participants were familiar with the term "vowel." In the third stage, participants were asked to repeat each word beginning at the first vowel it contained up to the end of that word. This stage was only one step away from the Pig Latin game, which was practiced in the fourth, and final, stage of the training phase.

The training session was followed by a production phase, in which the participants had to produce a verbal response to the list of 80 stimuli presented acoustically over headphones (in the same order for each participant). At the beginning, the participants listened to a pre-recorded instruction, in which the rules of the game were repeated and illustrated using the word *droga* [drɔga] "road." This was followed by 80 trials, which included the 19 items with a CI sequence. There was a self-paced break after 40 trials. The participants had three seconds to produce their response to each item. After that time, a warning signal alerted them for the next stimulus. The responses were audio recorded and then transcribed phonetically and coded by the experimenter. The duration of the test phase was approximately 6 minutes. After completing the experiment, the participants filled in a short demographics questionnaire, indicating their age, gender, education and the place of origin.

5. Results and Discussion

5.1 Words with [Cɨ] Sequences

Figure 1 shows the results for words that began with a hard consonant (or a sequence of consonants) followed by [ɨ]. When the [+back] consonant was moved to the end of the word, the now word-initial [ɨ] vowel was pronounced as [+back] in 163 trials. In 32 trials, it was modified.

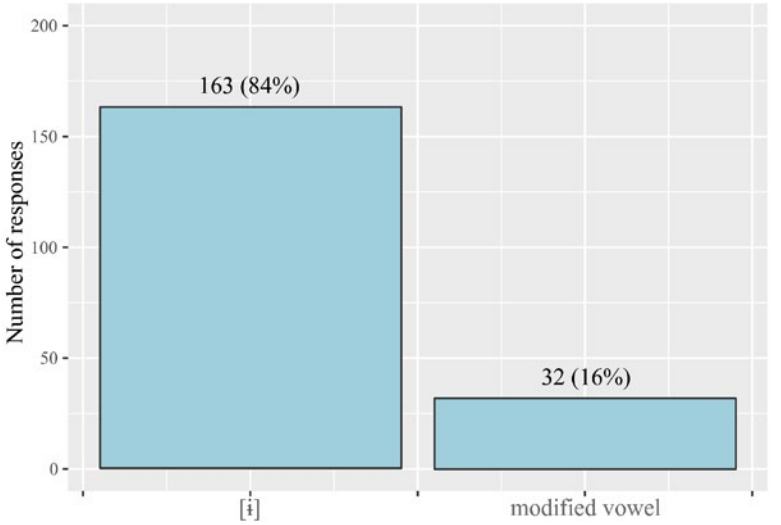


Figure 1. Results for words with [Cɨ] sequences

In most trials, then, moving the [+back] consonant(s) away from the vowel did not affect the [±back] value of that vowel. These results cast some doubt on the hypothesis that [ɨ] is a positional variant of //i//. The speakers did not seem to associate [ɨ] with [i], which is consistent with the hypothesis that the two vowels are underlyingly distinct.

As far as the putative ban on word-initial [ɨ]'s is concerned, the results are somewhat inconclusive. It is true that some of the newly-formed words were pronounced with a vowel other than [ɨ]. The vowel was never fronted and raised to [i]. Rather, it was lowered to the position approximating (but not quite as low as) the vowel [ɛ], with F1 on average 100 Hz higher than in unmodified [ɨ] vowels. There are some additional confounding factors, however. First, the modifications were more frequent with items at the beginning of the list and less frequent at the end. Since the items were presented to each participant in the same order, an ordering effect may be at play. Additionally, there was some degree of inter-speaker variation, with some participants pronouncing all initial vowels as [ɨ] and others modifying more than half of the items. These results could suggest that for some speakers, the ban on word-initial [ɨ]'s does form part of their phonological knowledge. Before we can conclude this, however, we would have to verify whether the "modified" articulations really constitute lowering with respect to the target vowel: as observed by Gonet (1993), Polish [ɨ] and [ɛ] tend to overlap in acoustic space. This may have been true for some of the participants, too.

5.2 Words with [Cʲi] Sequences

Figure 2 shows the results for words that began with a soft consonant followed by [i]. When the [−back] consonant was moved away from the [i] vowel, it was depalatalized in 173 trials and remained soft in 5 trials (four of these performed by the same participant).

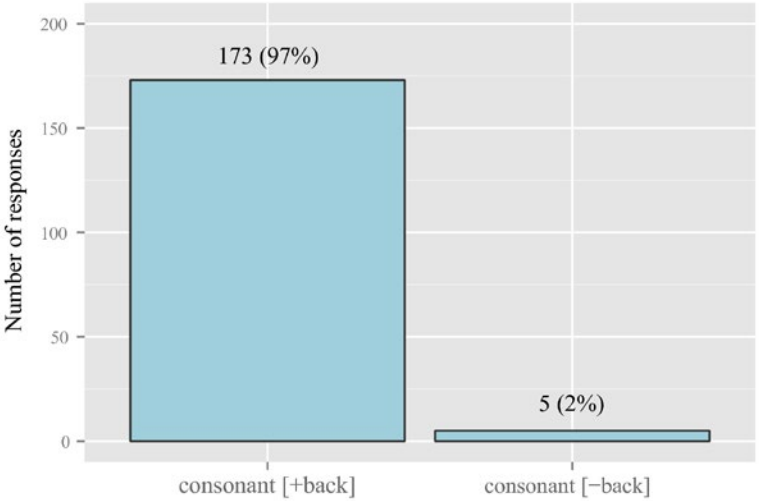


Figure 2. Results for words with [Cʲi] sequences

These results are inconsistent with the hypothesis that soft consonants are underlyingly specified as [−back] and that [ɨ] and [i] are realizational variants of the same underlying segment whose quality depends on the backness of the preceding consonant. Furthermore, the results are at odds with the hypothesis that both classes are contrastive for the feature [±back] and that the principle of Lexicon Optimization holds. The results are in line with two hypotheses, both of which state that [ɨ] and [i] are underlyingly distinct. The first hypothesis treats soft and hard consonants as allophonic variants, selected on the basis of the following context. The second hypothesis is that the [±back] contrast is also present in the consonants but the underlying representation of CI sequences includes hard consonants due to the free-ride principle.

One line of defense of the hypothesis that [ɨ] is a surface variant of //i// would be to argue that the results could be explained by reference to the orthography. Since consonant softness is not marked in spelling, if the participants performed the game by thinking in terms of letters, rather than sounds, the effect would be identical to the above. There are reasons to doubt strong orthographic influence, however. First, the stimuli were only presented auditorily. Second, the clapping activities in the training phase emphasized the aural nature of the game. Finally, the full list of items included a word beginning with a prepalatal consonant followed by [i], *sito* [ɕitɔ] "sieve." If the participants manipulated

letters, the expected outcome of the game would be [itɔsu]. However, 13 of the 20 participants produced [itɔɕu], retaining the quality of the initial consonant that is usually assumed to be underlying. Nevertheless, since only one such item was included, no firm conclusions can be drawn regarding the possibility of orthographic influence.

6. Summary and Outlook

Taken together, the results of this pilot study provide initial support for the hypothesis that [ɨ] and [i] are underlyingly distinct. The results are compelling enough to warrant a larger-scale experiment, but the study also indicates that the experimental design requires some modification. First, to control for orthographic influence, the set of stimuli should include a greater number of items beginning with prepalatals followed by [i]. Since such items are pronounced differently depending on whether letters or sounds are moved to the end of the word, they make it possible to identify the participants who think of the task in terms of spelling.

Additional changes need to be made to shed light on the question whether speakers have active knowledge of the lack on word-initial [ɨ]'s. To ensure that the "lowering" reported above is indeed the result of removing the initial consonant, it would be necessary to compare the production of the vowels in the transformed words to the pronunciation of the original input to the game. Additionally, to control for possible ordering effect, the list of stimuli should be presented in randomized order for each participant. Even with these modifications, there is a possibility that some participants will use a strategy that does not involve underlying sound categories. Consequently, the results of the study cannot be considered in isolation. However, used in conjunction with corroborating evidence from other experimental studies, they may constitute a valuable contribution to the still unresolved debate about the locus of the [±back] contrast in Polish.

Acknowledgements

I would like to thank the anonymous Olinco reviewer, whose insightful remarks have led to a much improved version of the paper. I am also grateful to Andrew Murphy and to the audience at Olinco 2016 for helpful comments and suggestions.

Works Cited

Baudouin de Courtenay, Jan Niecisław. 1893. *Próba teorji alternacyj fonetycznych. Cz. 1, Ogólna*. Kraków: Akademia Umiejętności.

Becker, Michael, Nihan Ketrez, and Andrew Nevins. 2011. "The Surfeit of the Stimulus: Analytic Biases Filter Lexical Statistics in Turkish Laryngeal Alternations." *Language* 87: 54–125.

Bertinetto, Pier Marco. 1987. "Lingue segrete e segreti delle lingue. Alcuni problemi di fonologia italiana studiati attraverso un gioco linguistico." *Annali della Scuola Normale Superiore di Pisa, Classe di lettere e filosofia, Serie III* 17: 889–920.

Biedrzycki, Leszek. 1974. *Abriß der polnischen Phonetik*. Warszawa: Wiedza Powszechna.
Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.
Clements, George N. 1989. "A Unified Set of Features for Consonants and Vowels." Ms., Cornell University, Ithaca, NY.
Clements, George N., and Elizabeth Hume. 1995. "The Internal Organization of Speech Sounds." In *The Handbook of Phonological Theory*, edited by John Goldsmith, 245–306. Cambridge, MA: Blackwell.
Czaykowska-Higgins, Ewa. 1988. "Investigations into Polish Morphology and Phonology." PhD diss., Massachusetts Institute of Technology.
Dawdy-Hesterberg, Lisa. 2014. "The Structural and Statistical Basis of Morphological Generalization in Arabic." PhD diss., Northwestern University.
Derwing, Bruce, Maureen Dow, and Terrance Nearey. 1988. "Experimenting with Syllable Structure. In *Proceedings of the Eastern States Conference on Linguistics*, vol. 5, edited by Joyce Powers and Kenneth de Jong, 83–94. Columbus: Ohio State University.
Gonet, Wiktor. 1993. "Próba określenia normy wymowy polskich samogłosek ustnych." In *Opuscula Logopaedica in honorem Leonis Kaczmarek*, edited by Jerzy Bartmiński et al., 232–52. Lublin: Wydawnictwo UMCS.
Guimarães, Maximiliano, and Andrew Nevins. 2013. "Probing the Representation of Nasal Vowels in Brazilian." *Oragon* 28: 155–78.
Gussmann, Edmund. 1980a. *Introduction to Phonological Analysis*. Warszawa: Państwowe Wydawnictwo Naukowe.
Gussmann, Edmund. 1980b. *Studies in Abstract Phonology*. Cambridge, MA: MIT Press.
Hume, Elizabeth. 1992. "Front Vowels, Coronal Consonants and Their Interaction in Nonlinear Phonology." PhD diss., Cornell University.
Hume, Elizabeth. 1996. "Coronal Consonant, Front Vowel Parallels in Maltese." *Natural Language and Linguistic Theory* 14: 163–203.
Laycock, Don. 1972. "Towards a Typology of Ludlings, or Play Languages." *Linguistic Communications* 6: 61–113.
McCarthy, John J. 2005. "Taking a Free Ride in Morphophonemic Learning." *Catalan Journal of Linguistics* 4: 19–55.
Ohala, John J. 1986. "Consumer's Guide to Evidence in Phonology." *Phonology Yearbook* 3: 3–26.
Pierrehumbert, Janet, and Rami Nair. 1995. "Word Games and Syllable Structure." *Language and Speech* 38: 77–114.
Prince, Alan, and Paul Smolensky. 1993. "Optimality Theory. Constraint Interaction in Generative Grammar." Technical Report TR-2, Center for Cognitive Science, Rutgers University, New Brunswick, NJ, and Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado, Boulder.
Rubach, Jerzy. 1984. *Cyclic and Lexical Phonology: The Structure of Polish*. Dordrecht: Foris.

Rubach, Jerzy. 2003a. "Duke-of-York Derivations in Polish." *Linguistic Inquiry* 34: 601–29.
Rubach, Jerzy. 2003b. "Polish Palatalization in Derivational Optimality Theory." *Lingua* 113: 197–237.
Rubach, Jerzy. 2007. "Feature Geometry from the Perspective of Polish, Russian, and Ukrainian." *Linguistic Inquiry* 38: 85–138.
Rydzewski, Paweł. 2014. "Phonological Consequences of the Backness Distinction in High Vowels with Reference to Selected Slavic Languages from the Perspective of Current American Phonological Theories." PhD diss., University of Warsaw.
Rydzewski, Paweł. 2016. "A Polish Argument for the Underlying Status of [ɨ]." *Studies in Polish Linguistics* 11: 111–31.
Rydzewski, Paweł. 2017. "Unwarranted Exceptionality: The Case of Polish *y*." *Lingua* 189–90: 75–95.
Treiman, Rebecca. 1983. "The Structure of Spoken Syllables: Evidence from Novel Word Games." *Cognition* 15: 49–74.
Treiman, Rebecca, and Catalina Danis. 1988. "Syllabification of Intervocalic Consonants." *Journal of Memory and Language* 27: 87–104.
Wierzchowska, Bożena. 1963. "Budowa akustycza a artykulacja dzwięków mowy." *Biuletyn Polskiego Towarzystwa Językoznawczego* 22: 3–23.
Wierzchowska, Bożena. 1971. *Wymowa polska*. Warszawa: Państwowe Zakłady Wydawnictw Szkolnych.

Appendix

ID	Word	Gloss	ID	Word	Gloss
1	pesto	"pesto"	13	szabla	"sabre"
2	nutki	"notes" (DIM)	14	wąsy	"moustache"
3	prądy	"currents"	15	węże	"snakes"
4	mleko	"milk"	16	diabły	"devils"
5	piece	"furnaces"	**17**	**bryki**	"cars" (colloq.)
6	**ryby**	"fish" (PL)	18	chaszcze	"thicket"
7	błoto	"mud"	19	słowo	"word"
8	ręce	"hands"	**20**	**cyfry**	"digits"
9	źródło	"source'	21	dziąsło	"gum"
10	krówki	"cows" (DIM)	22	róże	"roses"
11	**pyza**	"dumpling"	23	mewa	"seagull"
12	kredka	"crayon"	24	cacko	"gem"

ID	Word	Gloss	ID	Word	Gloss
25	ważka	"dragonfly"	53	noce	"nights"
26	foki	"seals"	54	tęcza	"rainbows"
27	**figi**	"figs"	**55**	**misie**	"teddy bears"
28	mięso	"meat"	56	butle	"bottles"
29	pączki	"doughnuts"	57	kreski	"lines"
30	tętno	"pulse"	58	dachy	"roofs"
31	ważki	"dragonflies"	59	grupy	"groups"
32	**ptysie**	"pastry puffs"	60	flądra	"flounders"
33	kable	"cables"	61	hasło	"password"
34	cążki	"nail clippers"	62	rzęsa	"eyelash"
35	łóżko	"bed"	63	gąszcze	"thickets"
36	zorza	"aurora borealis"	**64**	**wiśnia**	"cherry"
37	sito	"sieve"	65	kluchy	"dumplings"
38	chlewy	"pigsties"	66	tuba	"tube"
39	kozy	"goats"	67	znaki	"signs"
40	**piwo**	"beer"	68	żądło	"sting"
41	pręgi	"streaks"	69	ciąża	"pregnancy"
42	**łyżki**	"spoons"	70	lody	"ice cream"
43	**wiza**	"visa"	**71**	**życie**	"life"
44	**pysio**	"muzzle" (DIM)	72	rzeczy	"things"
45	gęsi	"geese"	73	łoże	"bed"
46	**mydło**	"soap"	74	kobra	"cobra"
47	**lisy**	"foxes"	75	jeże	"hedgehogs"
48	żądze	"craving" (PL)	75	piegi	"freckles"
49	kości	"bones"	**77**	**bitwy**	"battles"
50	**pikle**	"pickles"	78	nóżka	"leg" (DIM)
51	**kiwi**	"kiwi"	79	więzy	"bonds"
52	**wydra**	"otter"	80	zęby	"teeth"

Table 5. All items

Living on the Edge: Integration vs. Modularity in the Phonology of Czech Anglicisms

Tomáš Duběda

Institute of the Czech Language, Prague, Czech Republic

dubeda@ff.cuni.cz

Abstract: The article discusses different aspects of the phonology of Czech Anglicisms, categorizing the adaptation mechanisms into "integrative" and "modular." The integrative mechanisms include the Phonological Approximation Principle, the Spelling Pronunciation Principle, and different types of analogies. The modular mechanisms, which increase the autonomy of the subsystem of Anglicisms, include the Original Pronunciation Principle, the presence of marked phonemes or phonotactic structures, phonological variability resulting from the competition between adaptation principles, irregular mapping between phonology and spelling, as well as underlying links to English phonology. Quantitative evidence is provided for some of these tendencies, and several psycholinguistic hypotheses are formulated in connection with the adaptation model.

Keywords: Czech; English; Anglicisms; loanwords; loanword adaptation; phonology

1. Phonological Adaptation of Loanwords in Czech

Despite their peripheral status with respect to native words, loanwords form a numerous and constantly growing lexical subclass in Czech (Svobodová 2007, 6; Molęda 2011, 7). When migrating from one language to another, loanwords may be subject to phonological adaptation (Calabrese and Wetzels 2009, 4), which allows for their smooth integration in the target language system. In the case of Czech, an important part of the adaptation processes is phonological normalization, i.e., the elimination of foreign phonological elements. A recent study has shown that the "phonological invasiveness" of Czech with respect to Anglicisms is stronger than that of German, but weaker than that of French (Duběda 2016b). We shall refer to this first tendency of loanword adaptation as "integrative."

However, non-adapted pronunciations do occur as well, and even for adapted items, specific formal aspects, such as the word's phonotactics or the irregular mapping between pronunciation and spelling, make it possible for a language user to identify loanwords as such and to treat them within a specific subsystem of Czech phonology. This second tendency, increasing the autonomy of the subsystem of loanwords and leading to a greater or lesser "visibility" of this lexical stratum, will be termed "modular."

The present article applies this dualistic view of loanword adaptation to the subclass of Anglicisms, discussing its different aspects and providing examples and statistical data from recent empirical studies. We are concerned only with questions of phonological adaptation; semantic or morphological issues are not taken into consideration.

2. Integrative Tendencies in the Phonology of Czech Anglicisms

2.1 The Phonological Approximation Principle

Within the system of adaptation principles described in Duběda (2016a), which constitutes a useful framework for the study of loanwords and which has been applied with success to different lexical samples (Duběda et al. 2014; Duběda 2015a; Duběda, 2015b), the Phonological Approximation Principle stands out as the dominant force of phonological adaptation. Under this principle, non-native phonemes are replaced by their nearest native counterparts, unacceptable phonotactic structures are normalized, and stress is shifted to the first syllable where applicable. For example, the English word *soundtrack* [ˈsaʊndtɹæk] yields the form [ˈsa͡unttrɛk], where the phonemes /s, n, t, k/ remain basically unchanged, the phonemes /aʊ, ɹ, æ/ are replaced by their nearest equivalents available in the target phonological system, and /d/ loses its voicing due to regressive voice assimilation.

The Phonological Approximation Principle operates on a perceptual basis, within the limits of phonological contrasts available in the target language. It may lead to one-to-one phonemic projections (e.g., /ʌ/ > /a/ as in *punk* [ˈpaŋk]), phonemic mergers (e.g., /e, æ/ > /ɛ/; both *Ellen* and *Alan* may be pronounced as [ˈʔɛlɛn] in Czech), and, in some rare cases, to phonemic differentiation (e.g., /ə/ > /ɛ/ in non-rhotic contexts, as in *company* [ˈkampɛnɪ], and /ɻ/ in rhotic contexts, as in *hacker* [ˈhɛkɻ]). As an illustration of these processes, the mapping of the British English vowel system onto the Czech system is given in Figure 1.

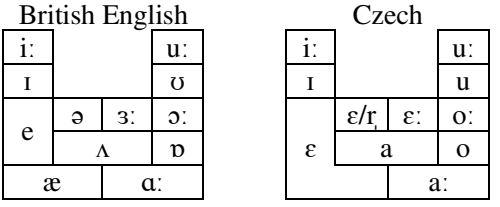


Table 1. Phonological Approximation of the English vowel system

This principle is considered to be the default mechanism in the adaptation of both loanwords and foreign proper names (Romportl 1978, 27; Hůrková 1995, 48). By studying a sample of 225 frequent Anglicisms used with non-adapted spelling we were able to demonstrate that the Phonological Approximation Principle is able to predict by itself the normative phonological form for 73% of the entries (Duběda et al. 2014).

The Phonological Approximation Principle is also responsible for most of the Anglicisms used with alternative or adapted spelling, e.g., *banjo*/*bendžo* [ˈbɛndʒo] "banjo," *víkend* [ˈviːkɛnt] "weekend." Furthemore, it interferes with the acquisition of English by Czech speakers, and may help explain many cases of negative phonological transfer. For example, the loss of contrastivity in the names *Ellen* and *Alan*, mentioned above, may occur not only when the words are used in Czech, in which case it is a consequence of due adaptation, but also in Czech speakers' English, where it is an undesirable feature that marks foreignness.

2.2 The Spelling Pronunciation Principle

Another option in the adaptation of loanwords is the Spelling Pronunciation Principle, whereby a word is adapted on the basis of its orthographic form rather than phonology. For instance, the words *bus* or *totem* are pronounced as if they were regular Czech words: [ˈbus], [ˈtotɛm], and not—as would have been predicted by the Phonological Approximation Principle—*[ˈbas], *[ˈto͡utɛm/ˈto͡utm̩]. In the aforementioned sample (Duběda et al. 2014), the Spelling Pronunciation Principle in its pure form is rare (3% of the items), while its co-occurrence with phonological approximation is quite frequent (24%): a word may either exist in two alternative variants (e.g., *holding*: phonological approximation [ˈɦo͡uldɪŋk] vs. spelling pronunciation [ˈɦoldɪŋk]), or be composed of two elements with different treatment (e.g., *antidumping* [ˈʔantɪdampɪŋk], where English phonology is deactivated in the prefix).

The influence of spelling is not surprising if we consider the importance of written communication in today's world, including new information channels such as electronic media (Molęda 2011, 20). Furthermore, the tendency towards spelling-based pronunciation is one of the ways of regularizing the relationship between phonology and orthography, the other being spelling adaptation (see 2.1).

2.3 Phonological Analogies

The peripheral nature of Anglicisms within the lexicon makes them phonologically less transparent than native words. This "phonological blurriness" is a breeding ground for various analogies, i.e., modifications of the standard phonological form influenced by other phonological forms which are more familiar to the language user. The source of analogy may be the target language, the donor language or a third language. The first category, also known as folk etymology, can be illustrated by the English verb *to browse* (*on the Internet*), which has led in Czech to the colloquial equivalent *brouzdat*

[ˈbro͡uzdat], identical to the native word *brouzdat* "to paddle (in shallow water)." As in other cases of folk etymology, this adaptation was triggered by similar phonological structure and semantic proximity. The influence of the donor language can be observed in forms in which Czech speakers wrongly apply their awareness of English phonology or that of the correspondence between spelling and pronunciation. For example, the attested pronunciation *[ˈɦo͡ustl̩] for *hostel* is most likely influenced by the English words *host* or *hotel*, which contain a diphthong. Other examples are *Robert* *[ˈro͡ubr̩t], *project* *[pro͡udʒɛkt], *corporate* (adj.) *[kɔrpɔrɛjt] and *Susan* *[sjuːzn̩]. Finally, the influence of a third language can be identified in some Anglicisms like *manažer* [ˈmanaʒɛr] "manager," probably adapted under the influence of French.

These alterations, while being indicative of phonological uncertainty, also give evidence of active phonological treatment. They represent an integrative force in that they "alienate" the word from its original phonological form; however, in the case of analogy with the donor language or a third language, they may be also categorized as modular tendencies because they impose phonological treatment based on other languages than Czech.

While the phonological exposure of Czech Anglicisms to third languages is very limited, the influence of English phonology on other loanwords, e.g., Gallicisms, is encountered more frequently. In other words, speakers sometimes treat loanwords as if they were Anglicisms, even if their origin is different. For example, the French poet's name *Charles Baudelaire* is occasionally pronounced as *[ˈʧaːrls ˈbodlɛːr] instead of [ˈʃarl ˈbodlɛːr], while, conversely, the pronunciation of, say, *Charles Darwin* as *[ˈʃarl ˈdaːrvɪn]—is highly unlikely. A number of examples from this category may also be found in Duběda (2015a): in this study, focused on the pronunciation of less known gastronomical terms of French origin in Czech, informants had recourse to English-influenced pronunciation in 8.2% of the items, e.g., *cordon bleu* *[ˈgoːrdon ˈbluː] instead of [ˈkordɔn ˈblɛː], *couvert* *[ˈkuvr̩t] instead of [ˈkuvɛːr] or *garni* *[ˈgaːrnɪ] instead of [ˈgarnɪ]. This "phonological spillover" is indicative of the prominent status of Anglicisms among loanwords. For less transparent loanwords, English may impose itself as "the default foreign language."

3. Modular Tendencies in the Phonology of Czech Anglicisms

3.1 The Original Pronunciation Principle

An obvious way of preserving the formal foreignness of a loanword is using it with its original pronunciation. This tendency, bordering on code mixing, can be observed in citations or foreign proper names, especially in intellectual contexts (Hůrková 1995, 64 *et infra*), as well as in informal communication involving topics such as pop culture or modern life style. Original pronunciation, however, is quite unusual and sounds unnatural in most situations (Hůrková 1995, 69), which makes Czech a language with high resistence to phonological import, except for foreign phonemes which are already well-integrated

(/f/, /g/, /ʤ/, /oː/, /a͡u/, /ɛ͡u/). Furthermore, it would be erroneous to believe that original pronunciation is an either-or category (Molęda 2011): in reality, its manifestations are often gradual and local. For example, if a Czech speaker opts for native-like or near-native pronunciation in the English given name *Jack*, he may use the native English sound [æ] or several other variants intermediate between [æ] and its Czech substitute [ɛ]. Where several phonemes may be pronounced in a non-adapted way, the speaker may choose a few of them or only one.

In a large-scale pronunciation survey described in Duběda (2016a), which included 138 Anglicisms and English proper names, instances of the Original Pronunciation Principle occurred at least once in 54 entries (i.e., 39%), but were mostly limited only to a small number of speakers. For example, the English given name *Chris* was pronounced by the overwhelming majority of the 300 speakers as [ˈkrɪs] or [ˈxrɪs] (Phonological Approximation and Spelling Pronunciation, respectively); only two speakers used phonetic elements imported from English: [ˈkʰrɪs], [ˈkʰɹɪs]. If we consider the total number of pronunciation variants obtained (138 items x 300 speakers = 41,400 variants), only 656 contain elements of original pronunciation (i.e., 1.6%). The analysis also confirms the claim made above: the influence of the Original Pronunciation Principle is mostly local, as in *Edward* [ˈʔɛdvɚt] (phonological approximation, except for the xenophoneme [ɚ], or in *William* [ˈwɪlɪjam] (spelling pronunciation, except for the xenophoneme [w]). The resulting effect corresponds to a discrete hint of foreign pronunciation rather than a true imitation.

In another study (Duběda 2015b), it has been shown that even in such a dynamic communication genre as TV advertising, original pronunciation plays only a marginal role. Despite potential gain resulting from foreign-sounding phonetic forms, English brand names (e.g., *Always Ultra*, *Head & Shoulders*, *Mr. Muscle*) are adapted to fit Czech phonology, and resist phonetic import.

3.2 Marked Phonemes and Phonotactic Structures

The Phonological Approximation Principle, which has been found to be the principal integrative force of Czech loanword phonology, yields results of two kinds: the adapted form may either contain only "genuine" Czech phonemes and phonotactic sequences (e.g., *body-check* [ˈbodɪʧɛk], *cup* [ˈkap]), or exhibit phonological properties which, though remaining within the limits of Czech phonology, reveal the foreign origin of the word. This latter case includes, on the one hand, well-integrated phonemes which occur only in loanwords or, marginally, in expressive words: /f/ (*film* [ˈfɪlm]), /g/ (*gag* [ˈgɛk]), /ʤ/ (*jazz* [ˈʤɛs]), /oː/ (*indoor* [ˈʔɪndoːr]), /a͡u/ (*joule* [ˈʤa͡ul]), /ɛ͡u/ (*terapeut* [ˈtɛrapɛ͡ut]). All of them are easily pronounceable, as they either correspond to variants of native phonemes (for example, [f] is a contextual variant of /v/ in native words, as in *včera* [ˈfʧɛra] "yesterday"), or to sequential combinations of native phonemes (/oː/ is a lengthened /o/, /ʤ/ is an articulatory blend of /d/ and /ʒ/, etc.). All of these marked phonemes with the exception of /ɛ͡u/ occur in

Anglicisms, and all of them also occur in words of other origins (Latinisms, Hellenisms, Gallicisms, etc.). On the other hand, the foreign origin of the lexeme may be indicated by unusual phonotactic structures, as in the words *error* [ˈʔɛror] "error (in computer slang)" or *Stephen* [ˈstiːvn̩], where the word-initial [ʔɛ] and the syllabic [vn̩] are felt as foreign-sounding, because they do not occur in native words.

3.3 Phonological Variability Resulting from the Competition of Adaptation Principles

Another indicator of a word's foreign origin can be sought in its potential variability: while native words exhibit very little phonological variation in Czech (leaving aside, of course, registers and regional accents), loanwords are more likely to have more than one standard pronunciation, e.g., *holding* [ˈɦo͡uldɪŋk/ˈɦoldɪŋk], *spam* [ˈspɛm/ˈspam] or *workshop* [ˈvɛrkʃop/ˈvorkʃop]. In the aforementioned pronunciation survey (Duběda 2016a), the average number of different pronunciation variants per item was 8.02 (4.01 excluding proper names, and 1.97 excluding variants with less than 5%). Only five of the 300 items were pronounced in the same way by all speakers. On the other hand, the greatest variability was recorded for the item *World* in the phrase *Miss World*: 54 different variants were identified, though only four of them had a frequency greater than 5%. The ten most frequent variants are: [ˈvort], [ˈvoːrt], [ˈvɝːlt], [ˈwɝːlt], [ˈvr̩l̩t], [ˈvoɹt], [ˈvərt], [ˈvoːɹt], [ˈvɝːt], [ˈvɛːɹlt].

Language users are exposed to this variability in speech perception, and may even switch between alternative pronunciations in their own production, as has been observed in the study mentioned above. This variability is not fortuitous, but can be explained by the competition of different adaptation principles, the most frequent case being the rivalry between the Phonological Approximation Principle and the Spelling Pronunciation Principle (cf. the three examples above).

Phonological variability of loanwords is socially structured, and may change in time (Štěpánová 2015). Furthermore, the pattern of phonological variability may be indicative of a word's origin. In Duběda (2016a), a set of variability paradigms was identified as being fully or mostly limited to Anglicisms, e.g., [a/ɛ] (*gang* [ˈgaŋk/ˈgɛŋk]), [Vr/V:r] (*software* [ˈsoftvɛr/ˈsoftvɛːr]) or [o/o͡u] (*Tony* [ˈtonɪ/ˈto͡unɪ]). It is probable that language users, being exposed to different pronunciation variants in their everyday communication, categorize words with similar variability patterns as belonging to the same class, thus making the lexical stratum of Anglicisms cognitively more coherent.

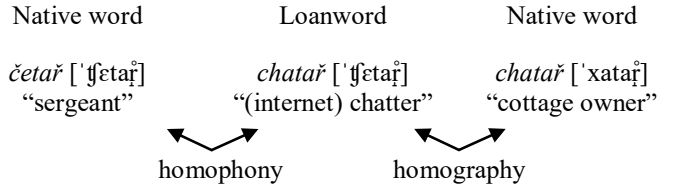
In the theoretical model supplied by Loanword Phonology (Calabrese and Wetzels 2009, 1ff.), phonological adaptation is described as an instantaneous process pertaining to a single speaker of the target language. This model may be useful as a theoretical construct, but from the sociolinguistic perspective, it is an oversimplification. A neological loan may appear in one or more foci, from where it spreads, provided that the speaking community or its part judges it worth adopting. During the process of spreading,

its phonological form is "negotiated" by language users, who may advance different adaptation preferences. In the case of Czech Anglicisms, this negotiation often leads to more than one generally accepted form.

3.4 Irregular Mapping between Pronunciation and Spelling

Yet another symptom of a word's foreign origin is the anomalous relationship between pronunciation and spelling. Native Czech words are characterized by a highly regular mapping between these two language supports: it is for this reason that native words are usually not provided with phonetic transcription in Czech dictionaries. On the other hand, orthographically non-adapted loanwords deviate from this regularity. An extreme example of this is the phonological form [ˈʧɛtaɽ̊], which corresponds to the native word *četař* "sergeant," but also to the Anglicism *chatař* "(internet) chatter." At the same time, the orthographic form *chatař* also corresponds to the native word [ˈxataɽ̊] "cottage owner":

(1)



The awareness of the anomalous relationship between spelling and pronunciation in the loanword *chatař* "internet chatter" also implies the awareness of its foreign origin, helping to keep it apart from its homophonous or homographic counterparts.

With respect to the native lexicon, loanwords form a peripheral lexical stratum, and are acquired later. Despite the lack of empirical evidence, it seems reasonable to argue that written forms play a greater role in their acquisition than it is the case for native words. Unlike the native lexicon, where the mapping between phonology and spelling is unproblematic, a successful acquisition of orthographically non-adapted loanwords requires double competence: spoken and written. Taking this idea further, we may hypothesize that the spelling form is latently present in the mental lexicon of the language user and may interfere if the phonological form is not fully activated.

3.5 Extant Link to the Source Language

An interesting aspect of loanword phonology, which is largely ignored in the literature, is the question of a possible underlying relationship with the source language in the speakers' mental lexicon. English is the most commonly taught and spoken foreign language in the Czech Republic (*Týdeník školství* 2010/17), which implies that many Czech speakers have at least a partial mastery of this language. Some lexical items thus exist in their mental lexicon in two forms: as an English word, and as a Czech word of English origin. Depending on their pronunciation skills, Czech speakers may keep both

phonological forms apart, or let the Czech phonological form, which they are more familiar with, influence the English phonological form. In this regard, the Czech accent in English would be, at least partly, explainable by the rules of phonological approximation applicable to loanwords (see Section 2.1). This hypothesis seems intuitively plausible, as both processes—phonological interference of the learner's mother tongue in second language acquisition and loanword adaptation—are based on the phonological categorization of phonetically similar units. Unlike for other, less known languages such as Latin or French, the widespread knowledge of English among Czechs probably helps maintain the phonological mapping defined by the Phonological Approximation Principle. One reason to believe this is that this mapping is surprisingly regular (cf. the aforementioned study in which it has been shown that the Phonological Approximation Principle is able to predict the adapted form of nearly three quarters of the entries analyzed).

The phonological subsystem of Czech Anglicisms, though deeply anchored in native phonology, thus shows, thanks to a regular phonological projection from English to Czech, features of what could be called a "shadow phonology" of English.

4. Conclusion

The different tendencies discussed above are summarized in Figure 1.

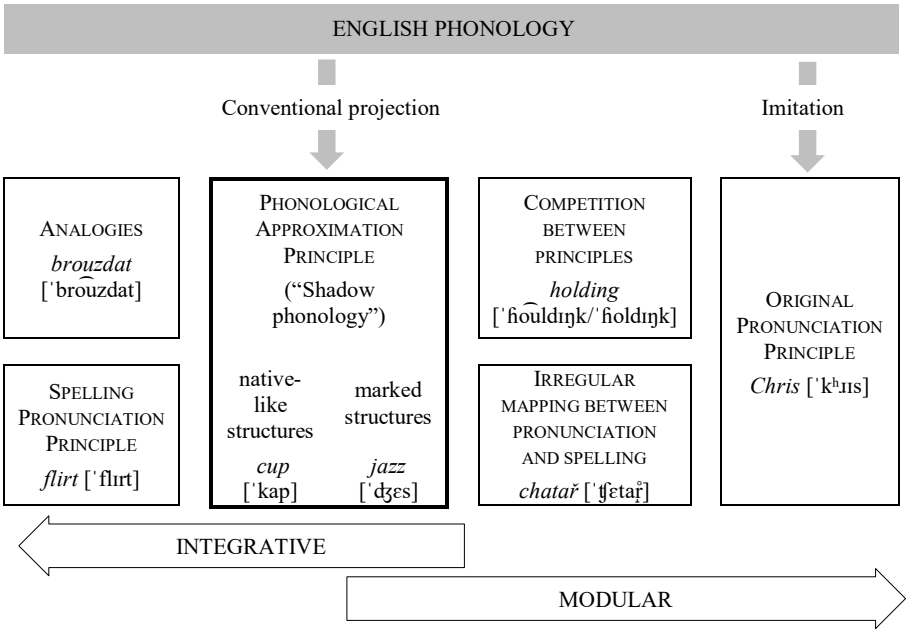


Figure 1. Summary of the integrative and modular tendencies in the phonological adaptation of loanwords

The diagram is organized along a horizontal axis, ranging from the most "integrative" to the most "modular" tendency. Each adaptation mechanism is illustrated by means of an example. The two major inputs of English phonology are marked with shaded vertical arrows. The Phonological Approximation Principle, highlighted by a thick frame, is the center of gravity of the whole system.

As the scheme suggests, Czech Anglicisms can be seen as a specific subsystem of Czech phonology, stretching between two poles—the integrative and the modular. At the integrative pole, the subsystem merges with native phonology, and at the modular pole, it projects beyond its limits. The center of the system is represented by the Phonological Approximation Principle, lying closer to the integrative pole, which establishes a conventional projection between original English phonology and the phonology of Czech.

Two different kinds of evidence can be sought to support this systemic view of Czech Anglicisms: First, quantitative analyses of several lexical samples were cited, which reveal the relative contribution of the different adaptation principles, as well as the degree of phonological variability in adapted loanwords. Second, several psycholinguistic hypotheses were formulated, whose investigation is beyond the scope of the present article. These include the question of formal recognizability of loanwords based on marked phonemes, phonotactic structures and paradigms of phonological variability, the role of spelling in the phonological treatment of loanwords, and the underlying relationship between the phonology of Anglicisms and English phonology for language users who are speakers of L2 English.

Funding Acknowledgement

This research was supported by the Czech Science Foundation grant Nr. 16-06012S.

Works Cited

Calabrese, Andrea, and W. Leo Wetzels, eds. 2009. *Loan Phonology*. Amsterdam: John Benjamins.

Duběda, Tomáš. 2015a. "L'adaptation phonologique des emprunts : le cas des gallicismes gastronomiques en tchèque." *Écho des études romanes* XI: 111–24.

Duběda, Tomáš. 2015b. "Jak se vyslovují názvy cizích produktů v televizních reklamách?" *Studie z aplikované lingvistiky* 2: 76–91.

Duběda, Tomáš. 2016a. "Empirické mapování výslovnostního úzu u cizích slov." *Slovo a slovesnost* 2: 123–42.

Duběda, Tomáš. 2016b. "L'invasivité phonologique dans le traitement des anglicismes: une étude quantitative de trois langues." *JEP-TALN-RECITAL* 2016: 401–9. Accessed September 9, 2016. https://jep-taln2016.limsi.fr/actes/Actes%20JTR-2016/V01-JEP.pdf.

Duběda, Tomáš, Martin Havlík, Lucie Jílková, and Veronika Štěpánová. 2014. "Loanwords and Foreign Proper Names in Czech: A Phonologist's View." In *Language Structure and Language Use. Proceedings of the Olomouc Linguistics Colloquium 2013*, edited by Joseph Emonds and Markéta Janebová, 313–21. Olomouc: Palacký University.

Hůrková, Jiřina. 1995. *Česká výslovnostní norma*. Prague: Scientia.

Molęda, Jacek. 2011. *A Comparative Study of Phonological Adaptations of Anglicisms in Czech and in Polish since the 1990s*. Racibórz: Państwowa Wyższa Szkoła Zawodowa v Raciborzu.

Romportl, Milan, ed. 1978. *Výslovnost spisovné češtiny. Výslovnost slov přejatých*. Prague: Academia.

Svobodová, Diana. 2007. *Internacionalizace současné české slovní zásoby*. Ostrava: Ostravská univerzita v Ostravě.

Štěpánová, Veronika. 2015. "Vokalická kvantita v přejatých slovech v češtině (výsledky výzkumu výslovnostního úzu)." *Naše řeč* 98: 169–89.

Týdeník školství 2010/17. Accessed September 9, 2016. http://www.tydenik-skolstvi.cz.

Word Study and the Lexicon: Corpus Approaches

So much as and *Even* in Downward Entailing Contexts: A Quantitative Study Based on Data from the British National Corpus

Volker Gast

Friedrich Schiller University Jena, Germany

volker.gast@uni-jena.de

Keywords: focus; polarity; scope; alternatives

Abstract: This contribution provides a comparison of the English focus operators *so much as* and *even* in downward entailing contexts on the basis of data from the British National Corpus. Five factors potentially determining the distribution of the two operators under analysis are taken into consideration: (i) the syntactic category of the co-constituent, (ii) the syntactic category of the focus, (iii) the type of downward entailing operator, (iv) the presence of focus alternatives in the clause, and (v) the source ordering the focus alternatives. It is shown that the two operators differ primarily in terms of the downward entailing operators they are licensed by. While *even* tends to occur more frequently in the scope of local negation than *so much as*, the latter operator is more commonly found in conditionals and *without*-PPs. A certain effect of the category of the co-constituent can also be observed. An explanation is offered for the affinity of *even* to local negation which derives tendencies in synchronic distributions from diachronic developments ("distributional inertia").[1]

1 An earlier version of this study was presented at Olinco 2016, at Palacký University Olomouc. I wish to thank the organizers for the invitation and the audience for valuable feedback and suggestions, especially (in alphabetical order) Markéta Janebová, Jaroslav Macháček, Michaela Martinková and Michaela Sedlářová. I am moreover indebted to two anonymous reviewers for valuable comments and suggestions. Any remaining inaccuracies are of course my own responsibility (and I apologize for the use of inappropriate deictics in combination with Czech place names).

1. Introduction

I use the term "scalar additive operator" for expressions such as Engl. *even*, Fr. *même*, Cz. *dokonce*, etc. (cf. König 1991, Giannakidou 2007, Gast and van der Auwera 2011, among others). These operators are additive, like Engl. *also*, *too*, etc., in the sense that they are appropriately used when there is a presupposition, or "focus supposition" (cf. Büring 2004), to the effect that the property attributed to the focus (or "added constituent") also holds of some other entity of a comparable category (cf. for instance König 1991, Reis and Rosengren 1997, Gast 2008). Unlike *also* and *too*, however, scalar additive operators are only used when the focus alternatives under discussion are ordered, thus forming a scale. Consider example (1a) and its "two-dimensional" representation in (1b) (I [arbitrarily] use the symbol † to indicate "unusual" utterances, under default assumptions about the world, and the number of †-symbols indicates degrees of [required] accommodation).

(1) Bill Nighy is very famous.
(a) *Even* the Queen congratulated him on his birthday.

(b) *Even* { the Queen / †his neighbour / ††his mother } congratulated him on his birthday.

(1a) asserts that Bill Nighy was congratulated by the Queen, and it requires the focus supposition that someone other than the Queen congratulated Bill as well. (1b) illustrates that the focus alternatives—the other potential congratulators, here *the Queen*, *Bill's neighbour* and *Bill's mother*—form a scale that is ordered in terms of what Gast and van der Auwera (2011) call "pragmatic strength"[2] (for similar, pragmatic analyses of *even*, cf. Fauconnier 1975, Anscombre and Ducrot 1983, Jacobs 1983, Kay 1990). The most common and prominent instance of pragmatic strength is probably unlikelihood, and *even* is actually "traditionally" analyzed as indicating that the attribution of the property in the background (in [1a], λx[*x* congratulated Bill]) to the focus (the Queen) is maximally unlikely (see for instance Karttunen and Karttunen 1977, Karttunen and Peters 1979, Rooth 1985).

As the ordering of focus alternatives in (1b) is compatible with our assumptions about the world, (1a) sounds natural. By contrast, (2) requires accommodation, e.g. in the sense that Bill is not on good terms with his mother.

(2) Bill is very happy.
††*Even* his mother congratulated him on his birthday.

2 "A proposition π is PRAGMATICALLY STRONGER (relative to a given quaestio *Q*) than a proposition ρ iff the RELEVANT CONTEXTUAL IMPLICATIONS of π (with respect to *Q*) entail the RELEVANT CONTEXTUAL IMPLICATIONS of ρ (with respect to *Q*)" (Gast and van der Auwera 2011, 9).

Under specific circumstances, scales of pragmatic strength are "reversed" (cf. Fauconnier 1975, König 1991, among others). This happens in the scope of "downward entailing" operators (Ladusaw 1979), e.g. negators.[3] (3a) therefore sounds natural, as under normal circumstances, it is Bill's mother who is most unlikely to *not* congratulate Bill. The ordering of alternatives in this case is shown in (3b).

(3) Bill is very unhappy.
  (a) Not *even* his mother congratulated him.
  (b) Not *even* { his mother / †his neighbour / ††the Queen } congratulated him.

The English operator *even* is compatible with upward as well as downward entailing contexts and can therefore be used in (1) as well as (3). Scalar additive operators may however be restricted to specific types of contexts. One such operator is the English multi-word expression *so much as*, which is only licensed in downward entailing contexts (cf. König 1982, Heim 1984, Gast and van der Auwera 2011). In such contexts it may be similar or even equivalent to *even*, cf. (4).

(4) If you *so much as* [~ *even*] parked on a yellow line they stuffed a mortgage application under your windscreen wipers. [BNC, SMA 4]

While there seems to be little difference in meaning between *even* and *so much as* in (4), there are (downward entailing) contexts where *so much as* is not commonly used in contemporary English, whereas *even* is fine. Two pertinent examples from my sample (cf. Section 2 and Note 7) are given in (5) and (6):

(5) Can the calculation of 165 deaths per one million rems be applied to all age groups—or *even* [?*so much as*] any? [BNC, EVEN 1065]

(6) That awful thing that so many groups get themselves involved in, when they're on a plane and they do a gig and they don't *even* [?*so much as*] know what city they're in, he'd manage to avoid. [BNC, EVEN 285]

3 An operator *Op* is downward entailing iff $(a \rightarrow b) \rightarrow (Op(b) \rightarrow Op(a))$. For example, *I saw a young man → I saw a man*; the negator *not* is downward entailing because *I did not see a young man → I did not see a man*; cf. also Section 2.2.1.

This contribution deals with the differences in the distribution of *even* and *so much as* in downward entailing contexts. It considers both (more or less) categorical differences and probabilistic factors, at a syntactic, semantic and pragmatic level. Specifically, the influence of the following variables is investigated:

- the syntactic category of the co-constituent of the operator;
- the syntactic category of the focus;
- the type of downward entailing operator;
- the relation ordering the set of focus alternatives;
- the presence or absence of focus alternatives in the clausal environment.

Following this brief introduction Section 2 contains a description of the data and the methodology, including the software used for the annotations. The (more) categorical differences between the two operators are discussed in Section 3. Section 4 presents a quantitative analysis taking into account probabilistic context conditions, in addition to the syntactic and semantic variables considered in Section 3. Section 5 briefly interprets the results against the background of the hypothesis that synchronic distributions mirror historical developments. Section 6 contains a summary and an outlook.

2. Data, Methodology and Software

2.1 Selection, Preprocessing and Syntactic Annotation of the Data

In a first step, I extracted all instances of *so much as* from the British National Corpus[4] using the BNCWeb-interface.[5] I filtered the 552 hits manually, identifying 261 of them as instances of the focus operator under study. Almost all of them—251—occurred in the register "Written books and periodicals." I therefore decided to focus on this register for my comparison of *so much as* and *even*.

In order to obtain a comparable sample of instances of *even*, I extracted a random sample of 5,000 examples from the BNC. I filtered out *even if*, *even though*, *even more* and combinations of *even* with an adjective in the comparative form (identified as such by the tag "AJC"), as they are not comparable to *so much as*. This left me with 3,881 examples of *even*. From the first 2,000 occurrences of this sample, I manually identified those that occurred in downward entailing contexts, resulting in a sample of 290 instances, 282 of which were from the register "Written books and periodicals." After a second round of manual inspection I identified eight false positives—instances of *even* that did not actually occur in a downward entailing context (cf. Note 1 and Section 2.2.1)—leaving me with 274 examples of *even*.

I used the whole sample for the more qualitative aspects of my analysis but extracted two random sub-samples of 100 examples of each operator for the quantitative analysis.

4 http://www.natcorp.ox.ac.uk/

5 http://www.bncweb.lancs.ac.uk/

These examples were tagged and parsed syntactically with the Stanford PCFG-parser (Klein and Manning 2003) and imported into GraphAnno (Gast et al. 2015), using the Python interface to GraphAnno, GraphPynt.[6] The parses were checked and manually corrected with GraphAnno. Example (7), thus processed, is represented in GraphAnno as shown in Figure 1.[7]

(7) [Its support began to crumble alarmingly even in its own working-class strongholds.] Far from being a challenger for power, it could not *even* hold on to its old citadels.

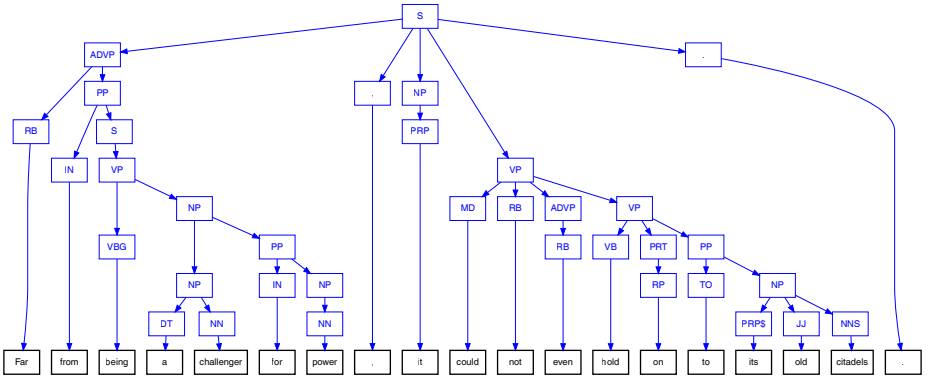


Figure 1. The syntactic structure of (7)

For a detailed semantic and pragmatic analysis of *even*, we have to annotate (minimally) the scope of the operator, its co-constituent and the focus (cf. also Gast, forthcoming). The scope of an operator is the (propositional) argument that the operator in question takes (cf. Gast and van der Auwera 2011). The co-constituent is the constituent that the operator combines with syntactically—the VP *hold on to its old citadels*, in (7). In (8) it is a NP and in (9) it is a verb (V).

(8) Relatively few voters read *even* $[_{NP}$ the $[_{A}$ best-selling$]_{F}$ papers]. [BNC, EVEN 139]

(9) It is not a problem that is $[_{V}$ solved$]_{F}$, or *even* $[_{V}$ touched$]_{F}$, by another 10s. [BNC, EVEN 166]

6 https://github.com/VolkerGast/GraphPynt

7 The data is available in csv-format on http://uni-jena.de/~mu65qev/data; EVEN and SMA stand for the sub-samples for *even* and *so much as*.

The focus, identified by a subscript F, is the element to which alternatives are considered.The exact extension of the focus is, to a certain degree, a matter of interpretation. In (7), there seems to be a paradigmatic contrast between *being a challenger for power* and *hold on to its old citadels*. I therefore assume that the whole VP is in focus (which is thus co-extensive with the co-constituent).[8]

The focus is invariably contained in the co-constituent of *even* / *so much as*, but it is not necessarily co-extensive with it. In (8), for instance, in one reading (suggested by the context of this example) it is only the adjective *best-selling* that is in focus, not the entire NP *the best-selling papers*.

The syntactic parameters characterizing sentences with focus operators can be represented in the annotations by classifying the relevant nodes accordingly. This was done using the semi-automatic annotation functionality of GraphAnno ("search-and-annotate," cf. Gast et al. 2016). The levels are differentiated by colours. With the nodes being classified in the way described above, sentence (7) is represented as shown in Figure 2 (in colour printing: grey: minimal sentence containing *even* [the highest S-node and any node dominated by it]; black: scope [the S-node at the first level of embedding and any node dominated by it]; orange: the operator itself [the RB-node immediately dominating *even*]; light blue: the downward entailing operator [the RB-node governing *not*], purple: the focus [the VP at the right margin of the sentence]).[9]

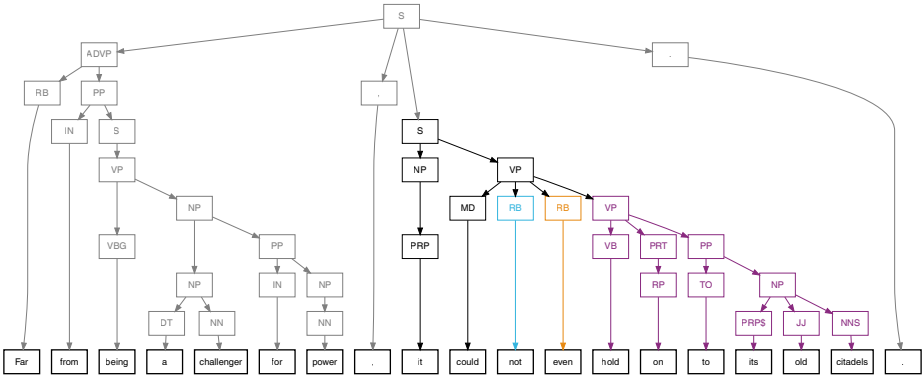


Figure 2. Example (2) with sentence constituents classified

8 An alternative interpretation would be to regard *the old citadels* as the focus, contrasting with other parts of the population where the party in question lost support. Given the "local" contrast within the sentence containing *even—being a challenger for power*—I assume the interpretation outlined in the main text.

9 See http://anglistika.upol.cz/olinco2016proceedings/ for the colour version.

2.2 Annotation of Further Variables

Given previous work on the topic (e.g. Hoeksema [2002, 2012] on Dutch; Gast [forthcoming] on German; Andorno and De Cesare [forthcoming] on French and Italian), I hypothesized that the two operators under study might be sensitive to the following properties of the context conditioning their distribution:

- the type of downward entailing operator;
- the ordering source of the alternatives;
- the presence or absence of alternatives in the clausal environment.

2.2.1 The Type of Downward Entailing Operator

The downward entailing operators occurring in my sample can be classified into the following types (they are sub-classified according to their entailment properties relative to the [extended] Zwarts hierarchy, cf. Zwarts 1995, Hoeksema 2012).[10] Roughly speaking, the hierarchy captures degrees of "strength" of negation and is primarily reflected in the distribution of negative polarity items (NPIs) such as *any*, *ever*, etc.[11] Questions and non-veridical superordinate predicates are regarded as a separate class because they exhibit a heterogeneous entailment behaviour).

- anti-morphic (anti-additive & anti-multiplicative)
  - negation with *not*
  - *without*-PPs
- anti-additive (not anti-multiplicative)
  - conditional operators (e.g. *if*)
  - (non-veridical) *before*-clauses
  - negation with *never* or a nominal *n*-determiner
- downward entailing (not anti-additive)
  - "weak negation" (e.g. *few* N)
- heterogeneous licensers
  - superordinate lexical triggers (e.g. *doubt*)
  - question operators (yes/no-questions, *wh*-questions)

Some pertinent examples are given in (10)–(15).

10 An operator *Op* is anti-additive iff $Op(\pi \vee \rho) \equiv Op(\pi) \wedge Op(\rho)$; an operator *Op* is anti-multiplicative iff $Op(\pi \wedge \rho) \equiv Op(\pi) \vee Op(\rho)$.

11 For example, "superstrong" NPIs such as *one bit* only occur in combination with anti-morphic operators, cf. *John wasn't one bit happy about these facts* vs. **No linguist was one bit happy about these facts*; cf. Krifka (1995, 217).

(10) negation with *not*
Danilov, The Voice observed, had probably **not** *even* begun to contemplate his murder when Dostoevsky was shaping Raskolnikov's. [BNC, EVEN 71]

(11) *without*-PP
There was another rustle of branches as the buffalo ran off **without** *so much as* another snort. [BNC, SMA 7]

(12) conditional
The new Lady Woodleigh looked as if she might take her riding-crop to him **if** he *so much as* uttered another word. [BNC, SMA 5]

(13) *before*-clause
It reminded me of all I disliked so much in the United States, of being called Ray **before** *even* shaking hands. [BNC, EVEN 94]

(14) superordinate lexical trigger (here, *unwise*)
Biggs is of the opinion that Mason would be unlikely to survive more than a couple of rounds against the world heavyweight champion and at this stage it would be **unwise** to *even* think of him as a genuine contender. [BNC, EVEN 10]

(15) negative determiner
No one *so much as* raised an eyebrow in their direction. [BNC, SMA 22]

2.2.2 The Ordering of the Focus Alternatives

The ordering of the focus alternatives distinguishes those cases where focus alternatives are ordered lexically from those where they are ordered on the basis of contextual knowledge only. (16) is an example of the former type. The focus *a million*, being a numeral, imposes an ordering on the set of alternatives. By contrast, the focus *inarticulate* in (17) does not in any obvious way constitute a scale with other focus alternatives—it forms a binary contrast with *articulate*.

(16) foci are lexically ordered
"If you'll pardon the correction, not *so much as* [a million]$_F$," said one of the lady lodgers. [BNC, SMA 38]

(17) not lexically ordered
She was not *even* [inarticulate]$_F$ in the sense that she could express her own feelings convincingly. [BNC, EVEN 87]

This variable is potentially informative because we can assume that *so much as*, as well as other expressions of its kind, is associated with foci that are lexically ordered, as it seems to exhibit an intrinsic association with (relatively) small quantities (*a million* in [16] contrasts with *over a million*; for an analysis of *so much as*-type operators as expressions of "small quantity" see for instance Vandeweghe [1981] on Dutch *ook maar*; cf. also Section 3.2).

2.2.3 Focus Alternatives in the Clausal Environment

Focus alternatives may be explicitly mentioned in the clausal environment or they may be implicit. In (18), there is an explicit contrast between *get a ride* and the focus alternative *win a race*. In (19), no such contrast can be retrieved.

(18) focus alternative present
Most stable-lads would have counted themselves lucky *even* to [get a ride]$_F$ let alone to [win a race]$_F$. [BNC, EVEN 211]

(19) focus alternative absent/implicit
[Picked him up at Imperial College. I gave them a three-hour lecture on the basic principles of stochastics, he said.]
Some composers today don't *even* understand [the simple calculus]$_F$, he said. [BNC, EVEN 16]

The presence or absence of focus alternatives has been shown to be an important factor determining the distribution of specific scalar operators in German (cf. Gast, forthcoming). Under the assumption that *so much as* contains more lexical information than *even*—"relatively small quantities" as opposed to "either relatively small or large quantities," depending on the type of context (upward entailing or downward entailing)—we can hypothesize that it is more prone to occurring without explicit focus alternatives in the immediate clausal environment than the latter operator.

2.2.4 Manual Annotations

The sample was annotated manually using GraphAnno by assigning the relevant properties to the nodes corresponding to the focus. The downward entailing operator had been identified at the preprocessing stage already (cf. Section 2.1). An example of a fully annotated sentence is shown in (3) (scal: lexical scalar ordering, "t" or "f"; conj: presence of explicit focus alternatives, "t" or "f"; remember that the structural annotations are represented by colours in GraphAnno, cf. Section 2.1).

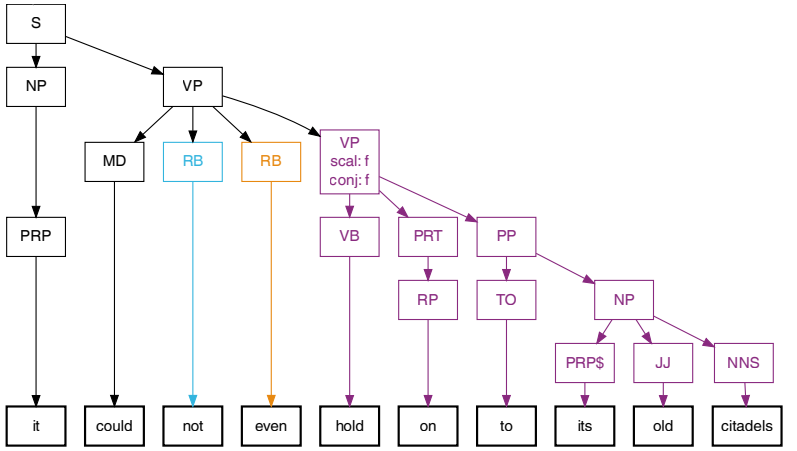


Figure 3. Fully annotated sentence

3. Some (More or Less) Categorical Distributional Differences between *So much as* and *Even*

3.1 Syntactic Differences

So much as is restricted much more severely in terms of its syntactic distribution than *even*. It occurs only with two types of co-constituents in my sample, with NPs (cf. [20]) and with VPs (cf. [21]):

(20) Musically this is very nearly correct, but not one recording includes *so much as* [$_{NP}$ a single word of Gilbert's spoken dialogue]. [BNC, SMA 104]

(21) For reasons best known to the fuel companies, the Gulf crisis never turned into an oil crisis, although petrol prices generally leap up and down quicker than a Tory backbencher during a Neil Kinnock speech if a dealer on the Amsterdam spot market *so much as* [$_{VP}$ sneezes over his computer screen]. [BNC, SMA 22]

The range of co-constituents that *even* combines with is much broader. Examples of co-constituents of category "adverb" and "PP" are given in (22) and (23). More precise quantitative data concerning the types of co-constituents will be provided in Section 4.

(22) Later I even appealed to the Member of Parliament for South Edinburgh, then the redoubtable Sir Will Y. I was afraid to leave Edinburgh, *even* momentarily, in case there was word from the War Office, but in September 1944 my mother persuaded me to go with her to Bedford for a short holiday. [BNC, EVEN 620]

(23) While Edward kept himself out of trouble with parents and school-masters without extending himself, he never, not *even* at St. Paul's, acquired the social ease of his schoolfellows. [BNC, EVEN 341]

Note that in some cases, *even*-sentences not allowing *so much as* for syntactic reasons can be paraphrased with the latter operator by placing it in the right position—cf. (24), where *so much as* is fine within the *to*-infinitive but not outside of it.

(24) (a) Amy won't do it and I can't find anyone [*even* [to [come in and keep it tidy]]]. [BNC EVEN 183]

(b) Amy won't do it and I can't find anyone [to [*so much as* [come in and keep it tidy]]]

Another property of *so much as* that restricts its distribution is the fact that it cannot occur at a distance from its focus. Consider (25a) and its counterpart with *so much as* in (25b).

(25) (a) It was an adventure *even* to find a stone, a clock movement, a tram ticket, a pretty leg, an insect, the corner of one 's own room; . . . [BNC, EVEN 491]

(b) It was an adventure to (†so much as) find (so much as) a stone, a clock movement, a tram ticket, a pretty leg, an insect, the corner of one's own room; . . .

If *so much as* precedes *find*, a reading is suggested in which *find* is part of the focus—i.e. the alternatives under consideration are different actions contrasting with *find a stone* etc. As the context does not suggest that reading, *so much as* would be expected to attach to the NP-conjunct [$_{NP}$ *a stone* . . .], if it was used to replace *even*.

3.2 Semantic Differences

As was mentioned in Section 3.1, *so much as* is restricted to two types of co-constituents, NPs and verbal projections (VPs and Vs). In either case, it tends to interact with foci that are associated with "littleness"—small quantities, such as *penny piece* (cf. [26]), or insignificant actions, such as *hint* in (27):

(26) Like other fellow scribblers whose squiggles seriously abuse the very title "shorthand notebook", I have nevertheless been generously given hours, sometimes even days, by sportsmen happy enough to rabbit on without *so much as* a penny piece being mentioned. [BNC, SMA 29]

(27) Anyone who *so much as* hints at a "third way" between communism and capitalism is considered naive; there is simply no time to try more experiments. [BNC, SMA 34]

Even is not sensitive to the lexical ordering of the focus alternatives in question. In context types that have undergone what we could call "double reversal," it is therefore commonly used, unlike *so much as*. Consider (28).

(28) And furthermore, F.L. Lucas (librarian of King's) would not *even* allow Eliot's work to be bought for his library. [BNC, EVEN 2533]

Eliot in (28) is not *per se* a low-ranking focus. What ranks low in terms of pragmatic strength is the VP *allow Eliot's work to be bought for his library* (a not unlikely state of affairs). Then, the scale is reversed by the negator *not*. This is an instance of "double scale reversal" if we assume that *Eliot's work* ranks high, *allow Eliot's work to be bought for his library* ranks low, and *not . . . allow Eliot's work to be bought for his library* ranks high again. If we inserted *so much as* as a co-constituent of *Eliot's work*, the resulting sentence would, if anything, imply that Eliot's work is insignificant. The effect of "double scale reversal" is even more clearly visible in (29):

(29) Indeed there are so many newspapers in contrast to the two television networks, that relatively few voters (as a percentage) read *even* the best-selling paper. [BNC, EVEN 139]

The NP *the best-selling paper* is not *per se* low-ranking; but the VP *read the best-selling papers* is likely and therefore ranks low on the scale of pragmatic strength. The downward entailing operator *few* in *few voters* inverts the scale of pragmatic strength a second time, rendering the proposition pragmatically strong or unlikely. And again, *even* is fine, because it is not sensitive to the lexical ranking of the alternatives, whereas *so much as* would be inappropriate here.

Another seemingly categorical restriction on *so much as* is that it does not combine with temporal foci such as *yesterday* (**so much as yesterday*). This restriction seems to be the reason why *so much as* cannot replace *even* in (30) or (31).[12]

(30) It is natural that our understanding of the solar system and our place in it be subject to periodic revision, continuing with a process of learning and discovery that began long before *even* the invention of the telescope. [BNC, EVEN 1287]

(31) It's not *even* 1989. [BNC, EVEN 1337]

12 While a detailed investigation of this restriction certainly deserves a study of this own, it seems reasonable to assume that the transparent meaning component of "quantity," reflected in the word *much*, requires some "vertical" scale for *so much as*, while temporal scales are normally conceived of as horizontal (in Western time metaphors).

It has moreover been observed that *so much as* (in comparison to other polarity-sensitive items such as *any* and *ever*) is bad in contexts of the type illustrated in (32b), expressing a (contingent) correlation between a generalizing relative clause (or some other clause denoting a condition or restriction) and episodic proposition in the main clause (Linebarger 1981, Heim 1984). It is fine in (32a), which suggests that the episodic context presents a problem.

(32) (a) Every restaurant that charges *so much as* a dime for iceberg lettuce ought to be closed down.

(b) *?Every restaurant that charges *so much as* a dime for iceberg lettuce happens to have four stars in the handbook. (Heim 1984, exx. 37/38)

While it is certainly true that (32b) sounds odd, my sample contains singular (though rare) examples of *so much as* in factive episodic sentences, cf. (33), so that I assume that the oddity of (32b) is not (exclusively) due to the episodic nature of the sentence.

(33) The only reason I *so much as* spoke to you last night, he said through his teeth, was because of my grandmother. [BNC, SMA 510]

Moreover, there does not seem to be a systematic difference between *even* and *so much as* in contexts of the type of (32). I have therefore not given this context parameter (generalizing vs. episodic sentences) any further consideration.

4. Differences between *So much as* and *Even*: Probabilistic Variables

Having discussed some more or less categorical factors distinguishing *even* from *so much as*, we can now turn to a quantitative analysis, also taking account of the "softer" distributional factors. As pointed out above, we will focus on five predictors in this section, i.e., variables that may influence the choice of *even* vs. *so much as*:

- The syntactic category of the co-constituent;
- the syntactic category of the focus;
- the type of downward entailing trigger;
- the ordering source of the scale;
- the presence or absence of focus alternatives in the sentence.

Note that the two syntactic variables—the category of the co-constituent and the category of the focus—are of interest despite the categorical restrictions that they exhibit. While *so much as* does not occur with specific types of co-constituents, it does occur with the

two most frequent types, i.e. NPs and VPs, so that a quantitative comparison with *even* makes sense.

In a first step, I used random forests (Breimann 2001) as implemented in the R-package party() (R Core Team 2015, Strobl et al. 2007) to determine the importance of each variable in a multivariate setting (for the sample of 2×100 examples, cf. Section 2). The results are shown in the form of a barplot in Figure 4.

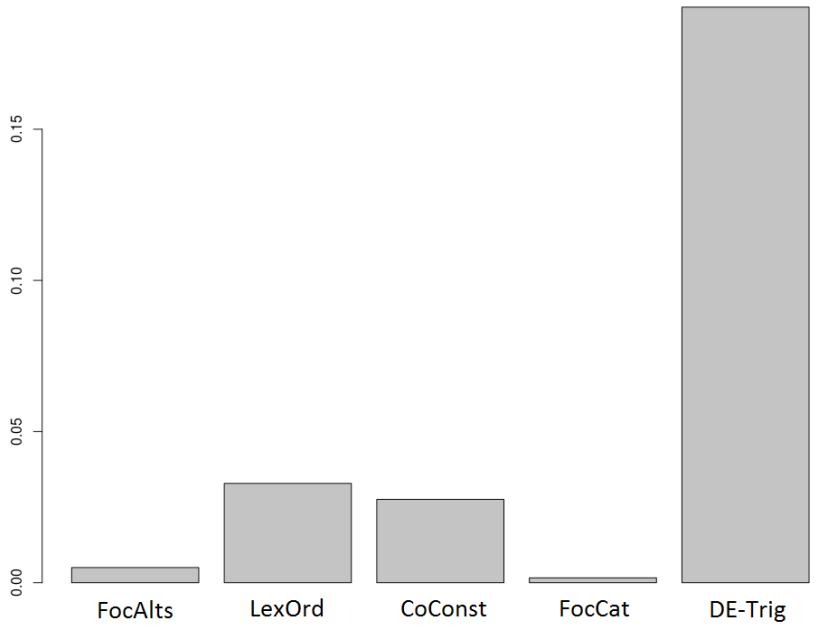


Figure 4. The importance of the five predictors according to a random forest analysis

The barplot in Figure 4 shows very clearly that the type of downward entailing operator ("DE-Trig") is the most important variable determining the choice of *even* vs. *so much as*. The ordering source ("LexOrd") and the category of the co-constituent ("CoConst") seem to have a certain impact, while "category of the focus" ("FocCat") and "presence vs. absence of focus alternatives" ("FocAlts") do not seem to have any influence on the choice of *even* vs. *so much as*. We will therefore focus on three variables in the following, proceeding from the weakest predictor (the category of the co-constituent) to the strongest predictor (the type of downward entailing operator).

Figure 5 shows the distribution of categories of co-constituents combining with *even* and *so much as*. As was pointed out in Section 3, *so much as* does not normally occur with co-constituents other than NPs or VPs. Within this major group, a clear tendency

can be observed for *so much as* to be associated with NPs, whereas *even* is comparatively more commonly used with VPs. The data is shown in the form of a barplot in Figure 5a and in the form of a Cohen-Friendly association plot in Figure 5b (cf. Cohen 1980, Friendly 1992).

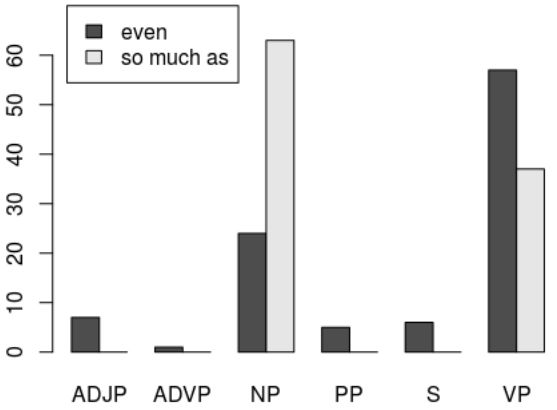


Figure 5a. The category of the co-constituent

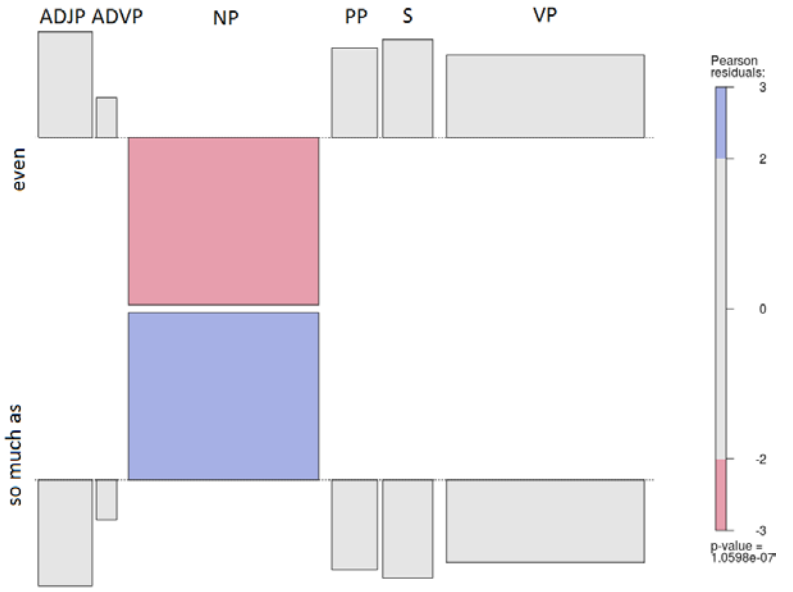


Figure 5b. The category of the co-constituent

The second most important variable is the source of the ordering relation. As Figure 6 shows, while being very rare overall, inherently (lexically) ordered foci (represented by the light grey area at the top of each bar) are significantly overrepresented in combination with *so much as*, in comparison to *even* (p=0.018, according to Fisher's Exact Test).

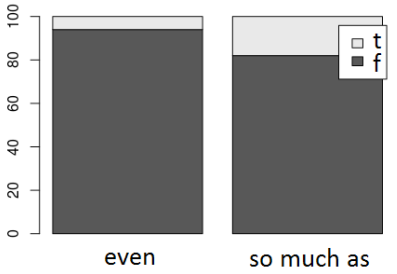


Figure 6. The ordering source (lexical vs. contextual) (p=0.018)

Finally, the frequencies of the various downward entailing operators in combination with either scalar operator are shown below in the form of a barplot (Figure 7) and a Cohen-Friendly association plot (Figure 8).

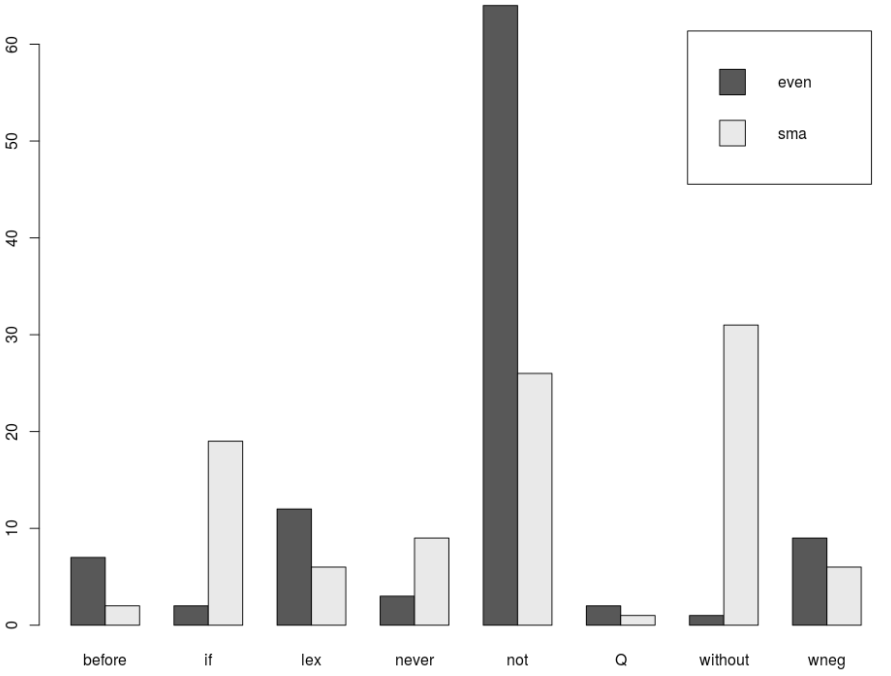


Figure 7. Frequencies of types of downward entailing operators in combination with *even* and *so much as*: a barplot

As we can see from the diagrams in Figures 7 and 8, *even* is significantly overrepresented in combination with *not*. *So much as* is significantly overrepresented in conditionals and *without*-PPs. Typical examples of instances of each operator are given in (34)–(36).

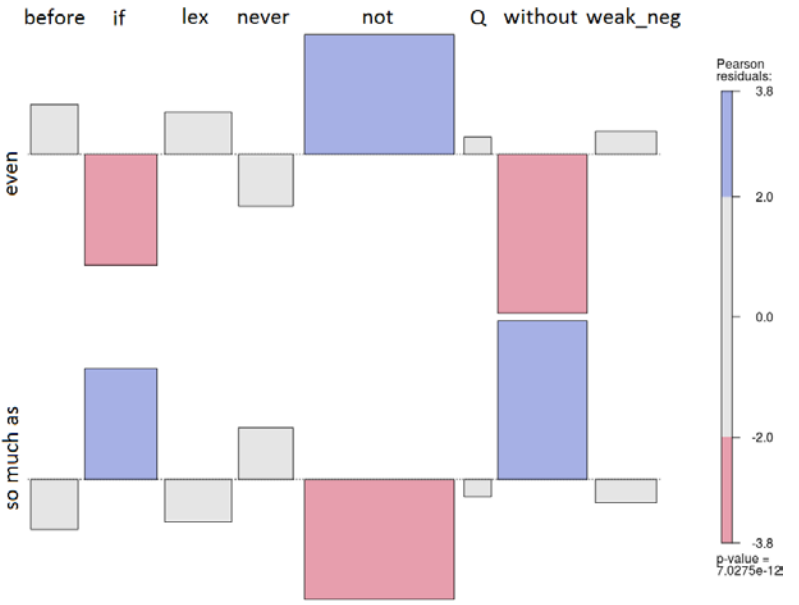


Figure 8. Frequencies of types of downward entailing operators in combination with *even* and *so much as*: a Cohen-Friendly association plot

(34) *even* under direct clause negation
But nowadays they call on new and brasher saints, whom St Margaret had not *even* met – Saints Epidura and Psychoprophilaxia. [BNC, EVEN 20]

(35) *so much as* in conditional
If you *so much as* parked on a yellow line they stuffed a mortgage application under your windscreen wipers. (=[4])

(36) *so much as* in *without*-PP
There was another rustle of branches as the buffalo ran off without *so much as* another snort. (=[11])

It is worth mentioning that *even* and *so much as* do not seem to be distributed along the dimensions of the (extended) Zwarts hierarchy (cf. Zwarts 1995, Hoeksema 2012). While *so much as* is particularly frequent within *without*-PPs, an anti-morphic context,

it is also frequently found in conditionals, which are not anti-morphic. The distribution of *even* does not follow a clear pattern along the Zwarts hierarchy either.

In order to get a more precise idea of the role played by the variables under study in combination, I fitted a logistic regression model. The goodness of fit is reasonable (Nagelkerke's pseudo-R2=0.52, C=0.86). The model shows that only two of the variables mentioned above are significant predictors for the choice of *even* or *so much as*, i.e. the type of downward entailing operator ($p < 0.001$), and the category of the co-constituent ($p = 0.001$). Figure 9 shows an association plot crossing these variable with the response variable "type of scalar additive operator" (*even* vs. *so much as*).

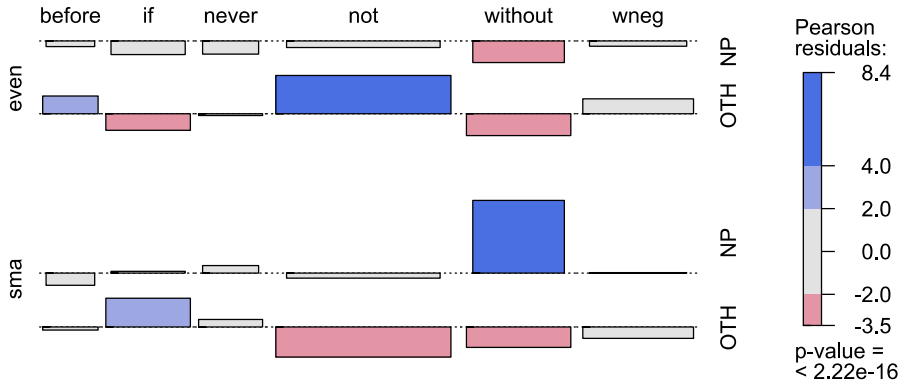


Figure 9. Logistic regression model with three variables

The association plot shows two particularly strong associations between feature combinations and operators (Pearson residuals > 4.0): *even* tends to co-occur with local negation (*not*) and non-nominal co-constituents ("other" [than NP]), while *so much as* is strongly attracted by *without*-PPs complemented by NPs. Somewhat weaker, but still significant, correlations are shown between *even* and *before* complemented by a non-nominal (i.e. clausal) constituent, and between *so much as* and *if*-clauses.

5. Interpreting the Data: Inertia in the Distribution of *Even*?

As I have argued elsewhere (e.g. Gast, forthcoming), tendencies concerning co-occurrence restrictions of the type identified in Section 4 often reflect historical developments, insofar as distributional properties of the source or "bridging context" (Heine 2002) are partially preserved in synchronic distributions. In research on grammaticalization, this phenomenon is known as "persistence" (Hopper 1991). In Gast (forthcoming), I have used to term "inertia" for it (note that the term is used in a slightly different way by Keenan 2002).

While I obviously cannot offer conclusive explanations for all of the tendencies pointed out in Section 4, one of the most striking asymmetries in the distribution of the two operators under analysis can reasonably be assumed to have its roots in historical developments. Remember that according to the analysis presented in Section 4, there are two types of downward entailing operators that are significantly associated with *even*: direct clause negation with *not* and *before*-sentences. The category of *before*-clauses can be regarded as belonging to the class of temporal foci, which do not normally go together with *so much as* (cf. Section 3.2). The first quantitative result—that *even* is overrepresented in combination with direct clause negation—is the one that lends itself to an explanation in terms of inertia (and this factor is also the only significant positive predictor according to a bivariate analysis, cf. Figure 8).

Gast and van der Auwera (2011) provide a brief summary of the historical development of *even* from a particularizer to a scalar operator. In downward entailing contexts, *even* came to be used at a relatively late stage—in the 17th century—often reinforcing *so much as*, which is attested some two hundred years earlier than *even* (in this particular context and function). A pertinent example from 1667 is given in (37).

(37) All which abuses, if those acute philosophers did not promote, yet they were never able to overcome; nay, *even* not *so much as* King Oberon and his invisible army. (Thomas Sprat, *The History of the Royal Society*, quoted from Gast and van der Auwera [2011, 43])

There is a type of context where *even* seems to have emerged independently of *so much as*, "emphatic negation." Before the 18th century, emphatic negation is often expressed in parentheticals following the main clause and introduced by *no not*. (38) is an example from the King James Bible (1611):

(38) Curse not the king, *no not* in thy thought. (Eccles. 10:20; quoted from Gast and van der Auwera [2011, 43])

Contexts of the type illustrated in (38) were probably important bridging contexts for the generalization of (scalar) *even* to downward entailing contexts, as *even* can be found accompanying *no not*. (39) is an example from the works of John Locke (1690):

(39) Expansion and duration have this further agreement, that, though they are both considered by us as having parts, yet their parts are not separable one from another, *no not even* in thought. (J. Locke, *An Essay concerning Human Understanding*, XV, 10; 1690; quoted from Gast and van der Auwera [2011, 43])

According to the type of development sketched above, *even* is a particularly clear example of a scalar operator that extended its domain from upward entailing contexts to direct

negation, and to other downward entailing contexts from there. Direct negation is thus the bridging context from upward to downward entailment, and can be regarded as the "archetypical context" within the latter class.

So much as has a different history. From its beginnings, it was not primarily associated with negation; it was simply a comparative expression stating the identity of two quantities. In its scalar function, the originally comparative meaning is "exploited" for what we could call a "reference-to-quantification transformation." For example, (40a) has a (wide scope) reading in which one of several apples will make the addressee sick. This reading is not available for (40b).

(40) (a) If you eat one apple, you'll get sick.

(b) If you eat *so much as* one apple, you'll get sick.

So much as one apple in (40b) is interpreted as "the quantity (of apples) corresponding to one apple." Note that a similar effect can be observed with *as much*/*many as* in upward entailing contexts, cf. (41):[13]

(41) John ate *as many as* eight apples!

It is an interesting question why *so much as*, unlike *as much*/*many as*, at some point specialized for downward entailing contexts. This asymmetry might be related to the fact that *as* (< OE *eallswā* "all so") contains an expression of precision, which would be incompatible with the establishment of a scale that comes with *so much as*. Answering this question more conclusively would of course require much more data, and more thorough semantic considerations.

For the time being, suffice it to say that the emergence of *so much as* was primarily semantically conditioned, and there was no preference for direct negation as opposed to other downward entailing contexts. In this respect, *so much as* seems to differ from *even*, and according to the "distributional inertia" hypothesis, this difference in the historical developments of the two operators is still reflected in their synchronic distributions.

6. Summary and Outlook

My comparative study of *even* and *so much as* has brought to light various categorical restrictions and significant correlations. *Even* and *so much as* show very different syntactic distributions, with *so much as* covering a subset of the context types where *even* is found. Moreover, they differ in terms of specific semantic properties of the focus,

13 I owe this observation to Ekkehard König.

e.g. insofar as *so much as* is not used in combination with temporal foci, and not with foci that "canonically" constitute the upper end of a scale.

Beyond these more or less categorical restrictions, *even* and *so much as* differ with respect to the types of downward entailing contexts that they typically occur in. Most importantly, *even* is significantly overrepresented in combination with direct negation. I have offered an explanation for this observation that derives tendencies in synchronic distributions from historical developments. Having emerged in upward entailing contexts, *even* —in its scalar function—extended its distribution to direct negation. It was only later that it came to be used in other types of downward entailing contexts. *So much as* does not seem to exhibit this kind of affinity to direct negation, from an either synchronic or diachronic point of view.

Needless to say, the hypothesis of "distributional inertia" for *even* should ideally be corroborated by studying historical corpus data—an endeavour that is hampered by the lack of richly annotated historical corpora, and the heterogeneity of the data, which renders (semi-)automatic annotation difficult, as even orthographies are not consistent across historical stages of English. Note that a comparison with other languages will also be instructive, for instance with Dutch *ook maar* and German *auch nur* (cf. Hoeksema 2002, 2012; Gast, forthcoming).

Works Cited

Andorno, Cecilia, and Anna-Maria De Cesare. Forthcoming. "Mapping Additivity through Translation. From French *aussi* to Italian *anche* and Back in the Europarl-Direct Corpus." In *Focus on Additivity. Adverbial modifiers in Romance, Germanic and Slavic languages.* Edited by Anna-Maria De Cesare and Cecilia Andorno. Amsterdam: John Benjamins.

Anscombre, Jean-Claude, and Oswald Ducrot. 1983. *L'argumentation dans la langue*. Brussels: Pierre Mardaga.

Breimann, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Büring, Daniel. 2004. "Focus Suppositions." *Theoretical Linguistics* 30: 65–76.

Cohen, Ayala. 1980. "On the Graphical Display of the Significant Components in a Two-Way Contingency Table." *Communications in Statistics Theory and Methods* A9: 1025–41.

Fauconnier, Gilles. 1975. "Pragmatic Scales and Logical Structure." *Linguistic Inquiry* 6: 353–75.

Friendly, Michael. 1992. "Graphical Methods for Categorical Data." *Proceedings of the SAS User's Group International Conference,* 17:1367–73. http://www.math.yorku.ca/SCS/sugi/sugi17-paper.html.

Gast, Volker. 2008. "The Distribution of *Also* and *Too*: A Preliminary Corpus Study." *Zeitschrift für Anglistik und Amerikanistik* 54 (2): 163–76.

Gast, Volker. Forthcoming. "The Scalar Operator *Even* and Its German Equivalents. Pragmatic and Syntactic Factors Determining the Use of *auch*, *selbst* and *sogar* in the Europarl Corpus." In *Focus on Additivity. Adverbial modifiers in Romance, Germanic and Slavic languages.* Edited by Anna-Maria De Cesare and Cecilia Andorno. Amsterdam: John Benjamins.

Gast, Volker, and Johan van der Auwera. 2011. "Scalar Additive Operators in the Languages of Europe." *Language* 87 (1): 2–54.

Gast, Volker, Lennart Bierkandt, and Christoph Rzymski. 2015. "Creating and Retrieving Tense and Aspect Annotations with GraphAnno, a Lightweight Tool for Multi-Level Annotation." In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Annotation*, edited by H. Bunt, 23–28. Tilburg: Tilburg Center for Cognition and Communication.

Gast, Volker, Lennart Bierkandt, Stephan Druskat, and Christoph Rzymski. 2016. "Enriching TimeBank: Towards a More Precise Annotation of Temporal Relations in a Text." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (*LREC* 2016), edited by Nicoletta Calzolari et al., 3844–50. Paris: European Language Resources Association.

Giannakidou, Anastasia. 2007. "The Landscape of *Even*." *Natural Language and Linguistic Theory* 25: 39–81.

Heim, Irene. 1984. "A Note on Negative Polarity and Downward-Entailingness." In *Proceedings of NELS* 14, edited by Charles Jones and Peter Sells, 98–107. Amherst, MA: GLSA.

Heine, Bernd. 2002. "On the Role of Context in Grammaticalization." In *New Reflections on Grammaticalization*, edited by Ilse Wischer and Gabriele Diewald, 83–101. Amsterdam: John Benjamins.

Hoeksema, Jack. 2002. "Polarity-Sensitive Scalar Particles in Early Modern and Present-Day Dutch." *Belgian Journal of Linguistics* 16: 53–64.

Hoeksema, Jack. 2012. "On the Natural History of Negative Polarity Items." *Linguistic Analysis* 38 (1/2): 3–33.

Hopper, Paul J. 1991. "On Some Principles of Grammaticization." In *Approaches to Grammaticalization*, vol. 1, edited by Elizabeth Closs Traugott and Bernd Heine, 17–36. Amsterdam: John Benjamins.

Jacobs, Joachim. 1983. *Fokus und Skalen: Zur Syntax und Semantik der Gradpartikeln im Deutschen*. Tübingen: Niemeyer.

Karttunen, Frances, and Lauri Karttunen. 1977. "*Even* Questions." In *Proceedings of the 7th meeting of the North Eastern Linguistic Society*, edited by Judy Anne Kegl, David Nash, and Annie Zaenen, 115–34. Cambridge, MA: North Eastern Linguistics Society.

Karttunen, Lauri, and Stanley Peters. 1979. "Conventional Implicature in Montague Grammar." In *Presuppositions*, vol. 11 of *Syntax and Semantics*, edited by Choon-Kyu Oh and David A. Dinneen. New York: Academic Press.

Kay, Paul. 1990. “Even.” *Linguistics and Philosophy* 13 (1): 59–111.
Keenan, Edward. 2002. “Explaining the Creation of Reflexive Pronouns in English.” In *Studies in the History of the English Language: A Millennial Perspective*, edited by Donka Minkova and Robert Stockwell, 325–54. Berlin: Mouton de Gruyter.
Klein, Dan, and Christopher Manning. 2003. “Accurate Unlexicalized Parsing.” In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423–30. Sapporo, Japan.
König, Ekkehard. 1982. “Scalar Particles in German and Their English Equivalents.” In *The Contrastive Grammar of English and German*, edited by Walter F. W. Lohnes and Edwin A. Hopkins, 76–101. Ann Arbor: Karoma Publishers.
König, Ekkehard. 1991. “Concessive Relations as the Dual of Causal Relations.” In *Semantic Universals and Universal Semantics*, edited by Dietmar Zaefferer, 190–209. Berlin: de Gruyter.
Krifka, Manfred. 1995. “The Semantics and Pragmatics of Polarity Items.” *Linguistic Analysis* 25: 209–57.
Ladusaw, William. 1979. “Polarity Sensitivity as Inherent Scope Relations.” PhD thesis, University of Texas.
Linebarger, Marcia. 1981. “Negative Polarity and Grammatical Representation.” *Linguistics and Philosophy* 10: 325–87.
R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Reis, Marga, and Inger Rosengren. 1997. “A Modular Approach to the Grammar of Additive Particles: The Case of German *auch*.” *Journal of Semantics* 14: 237–309.
Rooth, Mats. 1985. “Association with Focus.” PhD thesis, University of Massachusetts.
Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, Torsten Hothorn. 2007. “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.” *BMC Bioinformatics* 8 (25). https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-25.
Vandeweghe, Willy. 1981. “Ook maar x.” *Studie Germanica Gandensia* 21: 15–56.
Zwarts, Frans. 1995. “Nonveridical Contexts.” *Linguistic Analysis* 25: 286–312.

Lexical and Orthographic Distances between Bulgarian, Czech, Polish, and Russian, : A Comparative Analysis of the Most Frequent Nouns

Klára Jágrová, Irina Stenger, Roland Marti, and Tania Avgustinova

Saarland University, Saarbrücken, Germany

{kjagrova, avgustinova}@coli.uni-saarland.de,
{ira.stenger, rwmslav}@mx.uni-saarland.de

Abstract: This study analyses the share of cognates between four selected Slavic languages with special attention to their orthographic transparency. It is based on methods for determining lexical and orthographic distance between related languages. The underlying assumption is that the share of cognates and the transparency of orthography contribute to the mutual intelligibility of related languages. The material used is based on the respective national corpora. The distance measures serve as predictors for the performance of monolingual Slavic readers in their attempt to understand a related Slavic language. It aims at discovering the mechanisms by which intercomprehension in these closely related languages works. We observe lexical asymmetries for all language combinations and directions of reading.

Keywords: receptive multilingualism; Slavic languages; lexical distance; orthographic distance

1. Introduction and Motivation

This study is oriented on the methods for measuring linguistic distances between related languages applied by Heeringa et al. (2013) in their study on *Lexical and Orthographic Distances between Germanic, Romance and Slavic Languages and their Relationship to Geographic Distance*. The underlying assumption is that the intelligibility of related languages is, among other factors, influenced by the common share of cognates and

their orthographic transparency. This corresponds to the definition given by Heeringa et al. (2013, 102) that "a reader who is reading words spelled in a different but closely related language will understand the words relatively easily when cognates exist in his/her native language." Accordingly, readers will be more successful in identifying and understanding these cognates when they are spelled more similarly to their own language. Heeringa et al. (2013) measured lexical and orthographic distances for official EU languages of the Romance, Germanic, and Slavic language groups and their relationship to geographic distance. The method was grounded on a comparative analysis of translations of the 100 most frequent words of the British National Corpus (BNC)—the details are described in Section 3.

In the present study, similar linguistic methods are applied to different material. It aims at contributing further insights into receptive multilingualism among the selected Slavic languages (i) by including Russian into the comparison—not only because we focus on Bulgarian, Czech, Polish, and Russian, (hereafter referred to as BG, CS, PL, and RU, also in their adjective forms) in the framework of a larger research project, but also because it is relevant as the Slavic language with the biggest number of speakers and we expect more representative results from including another language with Cyrillic script in the comparison; and (ii) by using original sources of the languages under focus, viz. frequency lists based on the respective national corpora. Besides this, this study was conducted for reasons of verification and replicability of the findings of Heeringa et al. (2013) with regard to the linguistic distances. The distance measures should serve as predictors for the performance of monolingual Slavic readers in their attempt to understand a related Slavic language.

2. Objectives and Methods

We aim at finding predictors for the performance of Slavic readers in understanding a text in another unknown, but closely related language. Therefore, we determine the lexical and orthographic distance of four Slavic languages. We obtain the lexical distance of a language combination in a certain direction of reading by counting the amount of non-cognates that exist when translating the 100 most common nouns from one language to another.

We calculate the Levenshtein distance of the cognate pairs in a list to obtain scores of orthographic distance for each language combination and each decoding direction. CS and PL use Latin script, while BG and RU use Cyrillic script. We measure orthographic distance between the languages that use both of the different scripts. We obtain measures for both untransliterated and transliterated cognate pairs. This method is based on the methods applied in a related study by Heeringa et al. (2013).

3. Corpus-Based Material: The 100 Most Frequent Nouns of BG, CS, PL, and RU and Their Translations

3.1 Obtaining the Baseline Lists

The following section explains the procedure of obtaining the material used in order to measure the linguistic distances. In order to show the common points as well as the differences, we include a comparison of our method and that of Heeringa et al. While Heeringa et al. (2013) chose the 100 most frequent nouns from the British National Corpus and their translations into all other languages as the basis for comparison within the Romance, Slavic, and Germanic language families, we modified their method as follows:

Empirical basis	Heeringa et al. (2013)	Our approach
Source of data set	British National Corpus	frequency lists from the national corpora of BG, CS, PL, RU
Material	translations of the 100 most frequent English nouns	selection of the 100 most frequent nouns of BG, CS, PL, RU
Language family	Romance, Germanic, Slavic	Slavic
Slavic languages	Bulgarian, Croatian, Czech, Polish, Slovak, Slovene	BG, CS, PL, RU

Table 1. Explanation of method and material used

Instead of translating the 100 most frequent nouns from the British National Corpus, we systematically use original resources in order to establish a more representative picture of the individual languages. Among the 100 most frequent nouns from the British National Corpus, there are concepts that might occur relatively rarely in the Slavic languages: For instance, the translation equivalents for words like *policy*, *form*, *effect*, *police*, *use*, *line*, *study*, *report*, *rate*, *position*, etc., were in the list used by Heeringa et al. (2013), but are not part of our material (cf. The British National Corpus 2007). We expect that exploiting original language resources will have influence on the outcomes of such an analysis.

There are frequency lists based on the National Corpora readily available for all four languages, which of course is the most convenient way to obtain a list of the most frequent nouns. For CS and RU, the material is lemmatized, disambiguated,[1] annotated and sorted by POS (Křen 2010; Ljaševskaja and Šarov 2009; *Frequency Dictionaries of Bulgarian* 2011). Frequency lists for PL were published only recently by the LT group of the Politechnika Wrocławska. The list was generated on the basis of large corpora with an overall size of 1.8 billion tokens, including the IPI PAN corpus, Korpus Rzeczpospolitej,

1 According to the authors of the lists for both languages, the disambiguation is reliable for POS. In both cases, the authors point to the possibility that errors might still occur, especially with regard to homonymy.

Wikipedia (backup copy from early 2010) and a collection of large internet documents (Broda and Piasecki 2013).

For BG, there are frequency dictionaries sorted by subgenres available for download. However, the available files do not contain POS annotation along with the material. We unified the separate files from all subgenres into one big data file and counted the frequencies of all lemmata. We obtained a list of BG lemmata sorted by frequency and then manually extracted all nouns. This means that the BG list is not disambiguated by POS and e.g., nouns like *край* (*kraj*) or *син* (*sin*) could actually be representations of different lemmata:

(1) (a) *край* (*kraj*): nouns "end," "edge," "region," or preposition "at the end"

(b) *син* (*sin*): noun "son" or adjective "blue"

This should not happen for CS, PL and RU. However, as pointed out by the authors of the lists, cases of homonymy might occur in all four languages even though all words are nouns (see also [2] under Section 3.2).

We obtained 4 different baseline lists of the 100 most frequent nouns, each for one of the languages.

3.2 The Translation Process and Its Challenges

We translated our baseline lists from each language into all three other languages, obtaining 12 lists, each representing a language-reader combination in different directions. The translations were done by the members of our research group (the authors) themselves. We consulted other native speakers following the same principles as Heeringa et al. (2013): If a cognate translation of a word is possible in at least one context, then the cognate is chosen. The cognate translations can be "pairs of words which have the same meaning in both languages only in some contexts" (Heeringa et al. 2013, 103) as well. Cognates are consequently defined as both real cognates and partial cognates (ibid.). They have a common root and a common etymological origin. This principle represents an intercomprehension situation in which the reader would be able to identify the meaning of a word in a given context. If there is more than one possible cognate translation, we choose the cognate with the lower Levenshtein distance (hereafter referred to as LD; details will be explained in Section 3.3).

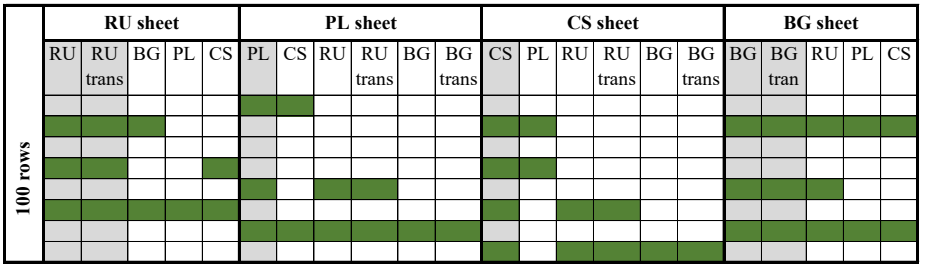


Figure 1. Visualization: translated cognate lists for each language combination and each direction of reading

The translated material consists of four sheets, one for each language. Each sheet contains the 100 most frequent nouns of the respective language (light grey column or columns in the case of BG and RU where there are two: Cyrillic original and Latin transliteration) and their translations into the other languages (white columns). Cognates are marked (green [dark grey] cells) and non-cognate translations remain unmarked (white cells).

Translating these words with cognates if possible turned out to be a non-trivial task, for meaning is a "messy" topic. In some cases we find cognates where the meanings overlap only in an extremely narrow context, e.g., in

(2) (a) PL *uwaga* ("caution," but also "consideration") and CS *úvaha* ("consideration")

(b) PL *ustawa* ("law," but also "statute") and CS *ustanovení* ("designation," but also "statute")

There are some clear cases of homonymy and polysemy in the lists. For instance, it is simple to determine which of the meanings of (3a–c) caused their placement among the top 100 of the frequency list.

(3) (a) BG *napa* (*para*[2] singular for "money" / "coin" or "steam"),

(b) PL *stan* ("state," but also "torso"),

(c) CS *stav* ("state" / "condition," but also "loom").

In other cases, it is arguable which of the possible meanings of a word is the most frequent one (the "main meaning"), e.g.,

2 For the transliteration employed, see Section 3.3.

(4) (a) RU *мир* (*mir* “peace” or “world”),

(b) CS *měsíc* (“month” or “moon”),

(c) PL *państwo* (“state” or “ladies and gentlemen”).

In other cases it is unclear whether we are dealing with polysemy or homonymy, e.g.,

(5) (a) RU *вид* (*vid*) could mean “kind,” “aspect,” “appearance,” “view,” etc.,

(b) RU *век* (*vek*) could mean “century” or “age,”

(c) CS *pohled* could mean “view,” “opinion,” “glance,” “glimpse,” “postcard,” etc.,

(d) PL *wzgląd* could mean “consideration,” “regard,” “respect,” “view,” etc.,

(e) BG *поглед* (*pogled*) could mean “view,” “glance,” “glimpse,” “look,” etc.,

(f) RU *взгляд* (*vzglâd*) could mean “look,” “glance,” “opinion,” “view,” etc.

The difficulties become apparent as soon as one is trying to find cognate translations, looking for the orthographically closest of them. Groups of cognates overlap in some nuances of their meanings, as shown in (5a–f). Here, context plays a crucial role in reading intercomprehension.

For the above-mentioned reasons, we decided to apply the principle that any meaning of a given word counts if there is a cognate translation in another language, even in cases where there are cognate translations only for the obviously non-frequent homonym and where there are no cognates in the “main” meaning of a word (3a–c). In a number of cases, ideal translations would be different from ours, as demonstrated in (2a) and (2b). The purpose of this study is to obtain measures of linguistic distance for the study of intercomprehension. For the same reason, we forego the distinction between “main” translations and rather rare translations. The main focus here lies merely on the understanding of linguistic code. The question is not how many different signifiers a concept has in the first place, but rather if readers are able to associate the signifier with the signified.

This principle holds also if the cognate translation chosen here is archaic or used in literary language. For instance, CS *oko* “eye” could be translated into Russian as *глаз* (*glaz*) or *око* (*oko* arch. “eye”). In this case, the cognate was chosen as a translation. This procedure aims at modeling an intercomprehension situation where the RU readers know the stimulus, e.g., from an archaic form of their language. Similar cases of cognate pairs were observed in all language combinations: BG *село* (*selo* “village”) translated

with PL *sioło* (arch. "village"), PL *osoba* "person" translated into BG *особа* (*osoba* lit. "person") or RU *комната* (*komnata* "room") translated into CS *komnata* (arch. "room").

As mentioned earlier, whenever there were multiple cognates, we chose the orthographically closest option (on the basis of LD), for instance PL *środek* has two possible translations into RU: *середина* (*seredina* "middle," "centre") or *средство* (*sredstvo* "means"). We chose *средство* (*sredstvo*) which has a LD of only 72% as opposed to *середина* (*seredina*) that has a LD of 81%.

We added English translations in an extra column, representing each of the meanings of the translations in the other languages. This implies that not all of the possible English translations are given in the list, but only those that are necessary to cover the meanings of the Slavic words. In the study of Heeringa et al. (2013), the English source words were translated into the individual languages by translators who were native speakers of these languages, with cognates to the English words if possible. However, the translators seem not to have followed these instructions strictly, as a number of words were not translated according to this principle, e.g., *system* was translated into CS as *soustava* instead of *systém*, *area* as *prostor* instead of *areál*, *service* as *služba* instead of *servis*, etc. (Golubović 2016, 210). In a second step, these baseline lists of each language were then again translated into all other languages within the same language family (Heeringa et al. 2013, 103), however, not necessarily taking into account the context of the initial English words. The same applies for the context of the words in our lists and their translations.

Country-specific nouns such as *sejm* (lower house of the Polish parliament), *koruna* (CS currency) or *лев* (*lev*, BG currency) were removed from our source lists. Also, obvious errors in the lists were corrected, e.g., *proca* "slingshot" in the PL list was replaced by *procent* "percent," because the abbreviation of *procent* was apparently mistaken for the genitive plural form of *proca* (*proc*) during processing.

Additionally, the words in Cyrillic were transliterated into Latin script according to ISO 9: 1986 for the measurement of orthographic distance between cognates in Cyrillic and Latin scripts. Orthographic distance is calculated both with and without transliteration, following the method applied by Heeringa et al. (2013). It is not possible to reverse the direction of transliteration since Latin script is only transcribed, but never transliterated into Cyrillic (cf. Wellisch 1978).

The translated lists and the word alignment matrices for the LD calculations (cf. Section 3.3) are made available by us under: http://www.coli.uni-saarland.de/~tania/incomslav.html/ (CC-NC-SA). An access code can be requested from the authors.

3.3 The Transliteration Process and Its Challenges

One of the challenges was how to calculate orthographic distance between Cyrillic and Latin script. As mentioned in Section 3.2, orthographic distance is calculated by means of the Levenshtein distance as in the study of Heeringa et al. (2013). Accordingly, orthographic distance of all cognate pairs was determined (i) with and (ii) without transliterations of the

Cyrillic script into Latin script. Heeringa et al. (2013) used the web application bg.translit.cc for Cyrillic to Latin transliteration. However, we found that the automatic transliteration is problematic in some cases: For instance, the BG word *община* (*obŝina* "community") is automatically transliterated into *obshtina*. This transliteration variant increases the Levenshtein distance by increasing the number of alignment slots by which the edit cost is divided:

# alignment slots	1	2	3	4	5	6	7	8	LD to CS
Original BG	о	б	щ			и	н	а	
Transliteration by web application	o	b	s	h	t	i	n	a	6/8 = 0.75
Aligned with CS	o	b				e	c		
Costs	0	0	1	1	1	1	1	1	

Table 2. Alignment of the automatic transliteration of BG *община* by the web application bg.translit.cc to its CS cognate *obec* "community"

# alignment slots	1	2	3	4	5	6	LD to CS
Original BG	о	б	щ	и	н	а	
Transliteration by ISO 9: 1986	o	b	ŝ	i	n	a	4/6 = 0.66
Aligned with CS	o	b		e	c		
Costs	0	0	1	1	1	1	

Table 3. Alignment of the ISO 9: 1986 transliteration of BG *община* to its CS cognate *obec* ("community")

Therefore, we decided to transliterate according to ISO 9: 1986, because each original sign corresponds to exactly one sign in the transliteration and hence, no additional alignment slots for calculating LD are necessary. Transliterating BG *община* by *obshtina* according to bg.translit.cc gives an LD of 75% (cf. Table 2), whereas our transliteration *obŝina* results in a lower LD of 66% (cf. Table 3). The Levenshtein distances were computed with a modified Levenshtein algorithm that aligns letters in slots according to weights for letter pairs, preferring an alignment of consonant letters to consonant letters and vowel letters to vowel letters. The computation consists of two steps: the automatic alignment of letters in slots and the calculation of the actual LD. The LD between two words is calculated based on the alignment.

In order to perform the alignment automatically, the algorithm is fed with letter weight matrices for each language combination and for the two additionally transliterated versions. Each matrix contains the complete alphabets of a language pair together with the costs assigned for every possible letter alignment. In order to guarantee that there is no alignment of vowel to consonant letters, all vowel-to-consonant combinations are given a weight of 4.5

(most expensive). Vowel-to-vowel and consonant-to-consonant combinations are given the weight of 1 and identical-to-identical letter combinations are given the weight of 0 (cheapest). Letters that differ only in their diacritical signs are given a weight of 0.5. The algorithm goes along each of the words, choosing the alignment combinations with the smallest final cost.

Once the algorithm performed the letter alignment in slots, the actual LD is calculated as demonstrated in Table 3. Insertions, deletions, and substitutions of letters cost 1 and differences in diacritical signs cost 0.5, identical letters cost 0. The costs are summed up and divided by the number of alignment slots in order to obtain a normalized LD value. No difference is made between the costs of the different diacritical signs that exist in CS and PL or in the transliterations: all combinations with either different diacritical signs or graphemes with or without diacritics always cost 0.5.

Heeringa et al. (2013) did not specify which letters of Cyrillic and Latin they considered identical in their calculations. For the untransliterated Cyrillic to Latin distance calculations we decided to make the following distinction: We assume that readers that are familiar only with Latin script will recognize Cyrillic *м*, *т*, *к*, *а*, *е*, and *о* as the equivalents of the Latin letters *m*, *t*, *k*, *a*, *e*, and *o*. From the other perspective, we assume that Cyrillic readers will recognize the Latin *a*, *e*, and *o* as correlates to their Cyrillic *а*, *е*, and *о*. We consider *t* a letter with different shape and account a cost of 1 in this direction of reading. Also *m* can be mistaken for the Cyrillic т[3] written in italics or in a different font. As for Latin *k*, it is not exactly identical to Cyrillic *к*. Therefore, we decided to treat it as equal to a difference in diacritical signs and assigned it a cost of 0.5 from the perspective of a Cyrillic reader. We assigned a cost of 0 for the other perspective, because we assume that *к* will be identified by readers used to the Latin script, as they know this shape of the letter from the Latin capital letter *K*. Other letters that are of identical shape, but have an entirely different phonetic representation, e.g., *р* and *p*, were treated as all other different letters and assigned a cost of 1. The following table demonstrates the calculation of the LD of BG towards PL with and without transliteration:

BG	PL	LD
страна	strona	0.6667
дейност	działalność	0.9090
ден	dzień	0.8
съд	sąd	1

BG	PL	LD
strana	strona	0.1667
dejnost	działalność	0.6818
den	dzień	0.5
sǎd	sąd	0.1667

Table 4. Orthographic distance of BG words for PL readers: example for comparison of LD without transliteration (left) and with transliteration (right)

3 This lowercase letter is not written in italics here in order to show its printed shape. When written in italics, Cyrillic т is usually rendered by *m*.

For instance, the difference between BG *съд* ("court") and PL *sąd* ("court") would be 100%. However, once BG *съд* is transliterated into *sâd*, the edit distance of the two words remains only 17%.

4. Results

Heeringa et al. conclude on the one hand that they "do not find asymmetric relationships on the lexical level" for the Slavic languages in their test set (2013, 117). On the other hand, however, there are asymmetric lexical distances in their result matrices in the appendix (2013, 134). In our case, we do find asymmetries on the lexical level for each of the language combinations. Examples of lexical asymmetries in the lists are demonstrated as follows:

CS	**PL**	**RU**	**RU translit**	**BG**	**BG translit**	**ENG**
dítě	dziecko	дитя	ditâ	дете	dete	child
oko	oko	око	oko	око	oko	eye
bod	punkt	точка	točka	точка	točka	point
PL	**CS**	**RU**	**RU translit**	**BG**	**BG translit**	**ENG**
sposób	způsob	способ	sposob	способ	sposob	way (manner)
pokój	pokoj	покой	pokoj	покой	pokoj	peace
RU	**RU translit**	**BG**	**CS**	**PL**		**ENG**
ребенок	rebenok	дете	dítě	dziecko		child
глаз	glaz	око	oko	oko		eye
дорога	doroga	път	dráha	droga		road
мир	mir	мир	mír	pokój		peace
BG	**BG translit**	**RU**	**PL**	**CS**		**ENG**
начин	način	способ	sposób	způsob		way (manner)
път	păt	путь	raz	krát		1) time, 2) way
точка	točka	точка	kropka	tečka		dot

Figure 2. Examples of lexical asymmetry in the lists: cognate translations (green [dark grey]) vs. non-cognates (white) of the original words from the lists (read: 1st column translated into the other columns)

The lexical asymmetry in the lists often emerges not only between two languages, but in some cases it may persist in all the other languages as well. For instance, all four languages share the cognates *dítě*, *dziecko*, *дитя* (*ditâ*), and *дете* (*dete*) "child." However, there is the word *ребенок* (*rebenok* "child") in the RU list which has no cognate translation in any of the other languages. The visualisation of the examples in Figure 2 represents the situations in which RU readers will for instance understand *dítě* because of the existence of the cognate *дитя* (*ditâ*), while CS readers will not understand *rebenok*.

Asymmetry in orthographic distance can occur even if the lexical distance is symmetric, as the same amount of cognate pairs for both directions of a language

combination does not imply the same set of cognate pairs. One of the reasons for this is that the two original lists are, of course, different, because the 100 most frequent nouns are a different set of words in every language.

4.1 Lexical Distance as the Percentage of Non-Cognates

We measure lexical distance by the share of non-cognates in language pairs. We assume that the higher the lexical distance score is, the more difficult it will be for readers to understand texts in an unknown language. At first glance, a partition by the sub-groups of the languages under focus becomes apparent by the distance score: the closest lexical relationships in the sample hold for BG and RU, South East Slavic and East Slavic, as well as for CS and PL, both West Slavic. For RU–BG, this is not surprising since RU has a substantial lexical layer of Church Slavonic which is South Slavonic, as is BG. On the other hand, BG has a substantial number of loan words from RU that were borrowed in the 19th century. However, we observe a lexical asymmetry between CS and PL that is larger than in the other pair and suggests that PL readers might find it harder to read and understand CS texts because of the higher share of non-cognates. The combination that is least intercomprehensive according to our results must be BG for a PL reader (33%). BG turns out to have higher distance scores for any reader when compared to the other languages read by other readers, meaning that BG is expected to cause the greatest lexical problems for other Slavic readers. The opposite holds for RU read by any other readers—the scores suggest a maximum distance of only 23%, meaning that RU is expected to cause less lexical problems than any of the other languages viewed here.

stimulus \ reader	BG	RU	CS	PL
BG		10	27	33
RU	11		20	23
CS	29	26		14
PL	27	20	10	

Figure 3. Lexical distance as the percentage of non-cognates

We found asymmetries in lexical distance for every language combination, depending on the direction of reading. The most remarkable asymmetries in lexical distance were observed for CS–RU 20% (CS decoder of RU stimulus) vs. RU–CS 26% (RU decoder of CS stimulus), as well as for PL–BG 27% (BG decoder of PL stimulus) vs. BG–PL 33% (PL decoder of BG stimulus), meaning that as far as vocabulary is concerned, CS readers are expected to find it easier to read RU and BG readers are more likely to succeed in reading and understanding PL than vice versa. Another minor asymmetry is

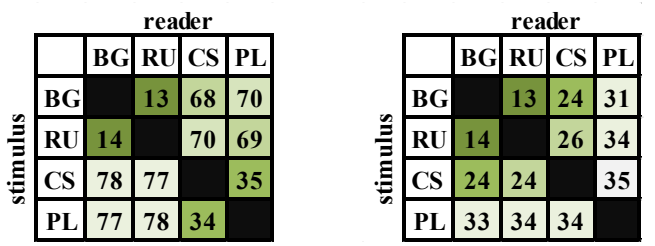
observed between RU and PL: 20% (RU decoder of PL stimulus) vs. 23% (PL decoder of RU stimulus). The scores surprisingly also imply that it must be harder for a PL reader (23% distance) than for a CZ reader (20% distance) to read RU, even though the geographical situation of these countries might lead us to assume the opposite.

4.2 Orthographic Distance of Cognates

We assume that the higher the orthographic distance, the more difficult it is to comprehend written cognates of the related language (cf. Gooskens 2007; Vanhove 2015). When comparing the languages sharing the same script, we find a remarkably lower orthographic distance in the pair with Cyrillic script (RU decoder of BG stimulus 13% vs. BG decoder of RU stimulus 14%) than in the pair with Latin script (PL decoder of CS stimulus 35% vs. CS decoder of PL stimulus 34%) meaning that the average LD would lead us to assume that CS and PL are less orthographically intelligible to each other than BG and RU are. This confirms our results from a previous study in which the orthographic distance between BG–RU and CS–PL was calculated on lists of Pan-Slavic vocabulary (CS–PL: 39% vs. BG–RU: 31%), internationalisms (CS–PL: 17% vs. BG–RU: 8%) and cognates from the Swadesh list (CS–PL: 42% vs. BG–RU: 33%) (Fischer et al. 2015; Stenger et al., forthcoming).

While the orthographic distance between BG and RU is only slightly higher than their lexical distance, there is a huge difference for CS and PL: despite the fact that these two languages are lexically relatively close when compared to other language combinations, their orthographic distance is the greatest of all combinations viewed here. This suggests that although readers can resort to a large share of cognate vocabulary, their mutual understanding is likely to be impaired by the different orthographies. It also leads us to conclude that orthographic opacity between these closely related languages might cause more problems in mutual intelligibility than is the case between other closely related languages.

In the transliterated version, we observe the highest orthographic distances always in combination with PL: not only is PL here the language that poses the greatest orthographic challenges to all other readers, but PL readers are also likely to face more difficulties caused by orthography when trying to read CS as well as BG and RU in transliteration. These comparably high orthographic distance values for PL might be due to the frequent consonant strings in PL that correspond to single letters in other languages, such as the Czech letters *č*, *ř*, or *š* with diacritics as opposed to their PL correspondences *cz*, *rz*, and *sz* which cause additional alignment slots and thus higher values in the LD calculation. Also for PL–RU and PL–BG, the greater orthographic distance might be caused by the additional alignment slots in digraph-to-monograph alignments such as *cz-č* or *sz-š* in the transliterated versions. Another factor is probably the fact that some of the letters that have diacritics in PL (albeit different ones, viz. *ć*, *ś*, *ź*) do not correspond to the seemingly similar CS letters *č*, *š*, *ž*.



stimulus \ reader	BG	RU	CS	PL
BG		13	68	70
RU	14		70	69
CS	78	77		35
PL	77	78	34	

stimulus \ reader	BG	RU	CS	PL
BG		13	24	31
RU	14		26	34
CS	24	24		35
PL	33	34	34	

Figure 4. Orthographic distance of cognate pairs without transliterations (left) and with transliterations (right)

Measuring the orthographic distance of non-transliterated Cyrillic to Latin script leaves us, as expected, with much higher LD than transliterated Cyrillic to Latin script. Consequently, a transliteration of Cyrillic reduces orthographic distance of cognate pairs dramatically. The scores representing readers of CS and PL trying to decode unknown Cyrillic code are somewhat lower than the scores of BG or RU readers decoding unknown CS or PL, which would suggest that CS or PL readers will perform better in reading Cyrillic even if they are not familiar with the script than vice versa. In spite of this it has to be taken into consideration that readers used to Cyrillic are more exposed to Latin than vice versa (due to IT and the learning of foreign languages using Latin script). Therefore these measured scores will most likely not have any predictive power.

5. Conclusions

We measured lexical and orthographic distance of the most frequent nouns of four Slavic languages: BG, CS, PL, and RU. The material that was used for this comparison was extracted from frequency lists based on the respective national corpora of the languages. We obtained scores for lexical distances by counting the number of non-cognates among the 100 most frequent nouns of a language and their translations into the other languages. We calculated the LD of these cognate pairs in order to determine the orthographic distance of these languages. The distance measures we obtained represent monolingual readers of the Slavic languages in their attempt to read an unknown, but related foreign language.

In general, our results reveal a partition into the subgroups of the languages: BG (South East Slavic) and RU (East Slavic) turn out to be the closest of these languages with regard to both lexis and orthography. In contrast to this, CS and PL display a large discrepancy between lexical closeness (only 10% distance for CS decoders of PL, resp. 14% distance for PL decoders of CS) and high orthographic distance (34% CS decoder of PL stimulus, resp. 35% PL decoder of CS stimulus).

We found lexical asymmetries in all combinations of languages, depending on the decoding direction. The greatest lexical asymmetries were found for CS–RU 20% (CS decoder of RU stimulus) vs. RU–CS 26% (RU decoder of CS stimulus), as well as for

PL–BG 27% (BG decoder of PL stimulus) vs. BG–PL 33% (PL decoder of BG stimulus). These scores suggest that CS readers are facing less difficulties when reading RU, while RU readers should find it harder to read and understand CS. Accordingly, BG readers are expected to have a slight lexical advantage when reading PL than vice versa.

We furthermore transliterated the BG and RU lists into Latin script and calculated the orthographic distance also of transliterated BG and transliterated RU to CS and PL. We obtained the greatest orthographic distances in all combinations with PL, confirming the distinct character of the PL orthography also shown in previous studies (cf. Heeringa et al. 2013).

6. Discussion and Future Work

Whether Slavic readers are indeed able to identify and understand the cognates will become apparent after obtaining results from translation experiments. We are currently preparing web-based experiments to investigate this.

In a next step, the distance between phonetic representations of these cognates could be determined. This has not been done by us so far, as this study is part of a research project on intercomprehension in reading. However, as most readers will try to pronounce the cognates with their inner speech (cf. Harley 2008), the phonetic representations, or at least what readers think might be the phonetic representations, are likely to be another factor worth investigating.

In the age of statistical language processing and machine translation, the problem of finding cognate translations could be approached with the help of parallel corpora. We also made use of large parallel resources and corpora-based online dictionaries (e.g., the Treq tool, www.glosbe.com). Relying solely on parallel corpora without any knowledge of the languages can lead to mistakes. For example, when looking for a cognate translation for CS *místo* "place," the Treq tool (based on InterCorp, release 9) suggests a list of possible translations, among others also PL *miasto* "city/town," which would be the orthographically closest translation proposed. However, the correct PL translation of *město* would be *miejsce* "town/city" and *miasto* is a false friend. The reason why tools such as Treq offer such translations is because there are several co-occurrences of *místo* "place" and *město* "city/town" within a sentence in the corpus and the alignment works statistically. Therefore, the translations were collected from various printed and online dictionaries as well as in extensive consultations with native speakers and were subsequently checked in parallel corpora.

One could argue that lexical and orthographic distance can be determined on the material of the traditional Swadesh lists (cf. Swadesh 1952). We measured the orthographic distance of the four languages on the Swadesh list as well as on two other cognate lists in a previous study (mentioned in Section 4.2, cf. Fischer et al. 2015). The main purpose of the previous study was to measure how often orthographic correlates apply in the two language pairs BG–RU and CS–PL. The results reflected a tendency that could be confirmed by the

present study: BG and RU are orthographically closer with less applicable orthographic correlates than CS and PL, although the overall values for orthographic distance in the previous study were higher (33% for BG–RU and 42% for CS and PL). This is most probably due to the fact that the Swadesh list does not contain internationalisms that have low orthographic distance. We expect a more representative synchronic sample of vocabularies from the frequency lists than from the Swadesh lists which do not contain internationalisms. In contrast to the present study, the values from the previous study are symmetric, because lexical distance was not measured and therefore there were no different word sets on which LD could have been measured in both directions of reading.

Funding Acknowledgement

This study was carried out in the context of a larger research project on mutual intelligibility among Slavic languages, concentrating mainly on BG, CS, PL, and RU. As a next step within this project, a large-scale web-based intelligibility test will be performed, including the material resulting from the present study. The INCOMSLAV project (Mutual Intelligibility and Surprisal in Slavic Intercomprehension) is part of the CRC 1102—Information Density and Linguistic Encoding at Saarland University, funded by the DFG.

Thanks

We wish to thank Varvara Obolonchykova for her support in the automatic calculation of orthographic distances and Aniko Kovać for corpus management. We owe special thanks to Juliana Stoyanova for consultations on the Bulgarian translations, and Magda Telus for the consultations on the Polish translations.

Works Cited

Broda, Bartosz, and Maciej Piasecki. 2013. "Parallel, Massive Processing in SuperMatrix—A General Tool for Distributional Semantic Analysis of Corpora." *International Journal of Data Mining, Modelling and Management* 5 (1): 1–19.

Fischer, Andrea, Klára Jágrová, Irina Stenger, Tania Avgustinova, Dietrich Klakow, and Roland Marti. 2015. "An Orthography Transformation Experiment with Czech-Polish and Bulgarian-Russian Parallel Word Sets. In *Natural Language Processing and Cognitive Science. Proceedings 2015*, edited by Bernadette Sharp, Wiesław Lubaszewski, and Rodolfo Delmonte, 115–26. Venezia: Libreria Editrice Cafoscarina.

Golubović, Jelena. 2016. "Mutual Intelligibility in the Slavic Language Area." PhD diss., University of Groningen.

Gooskens, Charlotte. 2007. "The Contribution of Linguistic Factors to the Intelligibility of Closely Related Languages." *Journal of Multilingual and Multicultural Development* 28 (6): 445–67.

Harley, Trevor. 2008. *The Psychology of Language: From Data to Theory*. 3rd ed. Hove: Psychology Press.

Heeringa, Wilbert, Jelena Golubović, Charlotte Gooskens, Anja Schüppert, Femke Swarte, and Stefanie Voigt. 2013. "Lexical and Orthographic Distances between Germanic, Romance and Slavic Languages and Their Relationship to Geographic Distance." In *Phonetics in Europe: Perception and Production*, edited by Charlotte Gooskens and Renee van Bezoijen, 99–137. Frankfurt a. M.: Peter Lang.

ISO 9: 1986. Documentation—Transliteration of Slavic Cyrillic characters into Latin characters. 1988. In *Documentation and Information.* (ISO standards handbook 1) 3rd edition. ISO, Genève.

Křen, Michal. 2010. "Srovnávací frekvenční seznamy." Czech National Corpus. Accessed September 11, 2016. http://ucnk.ff.cuni.cz/index.php.

Ljaševskaja, Oľga N., and Sergej A. Šarov. 2009. *Častotnyj slovar' sovremennogo russkogo jazyka (na materialah Nacional'nogo korpusa russkogo jazyka)*. Moscow: Azbukovnik.

Stenger, Irina, Klára Jágrová, Andrea Fischer, and Tania Avgustinova. Forthcoming. "'Reading Polish with Czech Eyes' or 'How Russian Can a Bulgarian Text Be?': Orthographic Differences as an Experimental Variable in Slavic Intercomprehension." In *Current Developments in Slavic Linguistics. Twenty Years After* [preliminary title], edited by Peter Kosta and Teodora Radeva-Bork. Bern: Peter Lang.

Vanhove, Jan. 2015. "The Early Learning of Interlingual Correspondence Rules in Receptive Multilingualism." *International Journal of Bilingualism* 20: 580–93.

Wellisch, Hans Hanan. 1978. *The Conversion of Script—Its Nature, History, and Utilization*. New York: Wiley.

Corpora

The British National Corpus. 2007. Version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk.

Czech National Corpus—Srovnávací frekvenční seznamy. (2010–) Institute of the Czech National Corpus. Accessed January 1, 2016. http://ucnk.ff.cuni.cz/srovnani10.php.

Czech National Corpus—InterCorp (version 9). Institute of the Czech National Corpus. Accessed February 3, 2017. http://treq.korpus.cz.

Frequency Dictionaries of Bulgarian. 2011. Department of Computational Linguistics, Bulgarian Academy of Sciences. Accessed April 5, 2016. http://dcl.bas.bg/en/tchestotni-retchnitsi-na-balgarskiya-ezik-2.

Lista frekwencyjna. 2016. Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej. Accessed September 8, 2016. http://www.nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/lista-frekwencyjna.

Ljaševskaja, Oľga N., and Sergej A. Šarov. 2009. *Novyj Častotnyj Slovar' Russkoj Leksiki* (*NČS*). Accessed April 5, 2016. http://dict.ruslang.ru/freq.php.

Emotions Translated: Enhancing a Subjectivity Lexicon Using a Parallel Valency Lexicon

Jana Šindlerová[a] and Aleš Tamchyna[b]

Charles University, Prague, Czech Republic

[a]sindlerova@ufal.mff.cuni.cz; [b]tamchyna@ufal.mff.cuni.cz

Abstract: This paper documents the behavior of verb valency complementations regarding the position of the target of evaluation within the valency frame. We classify the types of evaluative meaning expressed by the verbs and identify shared characteristic features considering the valency patterns of the verbs. In the analysis, we comment on three major issues of interest: the semantic classification of evaluative verbs and its relation to the propagation of sentiment value to the participants, the possible non-matching structural positions of the target of evaluation in the valency frame of a verb and its translation, i.e., the possible shift in evaluative focus and scope, and the possible loss of evaluative stance in the process of translation.

Keywords: sentiment; subjectivity lexicon; valency; parallel corpora

1. Introduction

In this paper, we present our efforts to enhance a Czech subjectivity lexicon with additional evaluative verb lemmas using a parallel valency lexicon as a relevant source of verb meanings. Also, we offer an analysis of the acquired verbs with respect to what happens to the evaluative state, the ordering of participants, the positions of the Source and Target of evaluation, and the evaluative strength of the verb in the process of sentence translation from one language to another (in this case, English to Czech).

Building subjectivity lexicons, or expanding them with additional meanings, can be done using various methods and resources. A popular resource for many languages is WordNet (Arora et al. 2012). Another option is employing unsupervised learning methods (Kanayama and Nasukawa 2006). Our approach is similar to the use of cross-lingual projections (Milhacea et al. 2007), but applied via a parallel lexicon and syntactically annotated corpus.

Subjectivity lexicons are valuable resources for identification of emotional, subjective and evaluative stances in the text. As verbs usually represent the core of the sentence, they often represent also the core of the so-called evaluative state,[1] in the case that the sentence expresses evaluative meaning. Their valency complementations, then, acquire the roles of the Source and Target of evaluation. By identifying evaluative verbs and their valency patterns in the text, we gain the ability to interpret the evaluative meaning of the sentence.

We believe that verbal valency may tell us much about the way evaluation is treated in a language, and moreover, a cross-lingual point of view may reveal some interesting facts about both the universal and language specific features of evaluative language as a linguistic construct.

2. Used Data and Theoretical Background

2.1 Czech Sublex 1.0

Czech Sublex 1.0 (Šindlerová et al. 2014; Veselovská 2013) is a Czech subjectivity lexicon, i.e., a list of subjectivity clues for sentiment analysis in Czech. It has been gained by automatic translation of a freely available English MPQA Subjectivity Lexicon (Wilson et al. 2005) using a Czech-English parallel corpus CzEng 1.0 (Bojar and Žabokrtský 2006). Additionally, some manual refinement of the lexicon followed in order to exclude controversial items. Finally, it contains 4,626 domain-independent evaluative items (1,672 positive and 2,954 negative) together with their part of speech tags, polarity orientation and source English lemmas. Of these, 1,549 are verbs.

2.2 Czengvallex 1.0

Czengvallex 1.0 (Urešová et al. 2016) is a parallel verb valency lexicon based on the Prague Czech English Dependency Treebank (PCEDT) (Hajič et al. 2012). It stores alignments between Czech and English valency frames and their arguments in about 22,000 English-Czech frame pairs. Aligned pairs of verb frames are grouped by the English verb frame, and for each English verb sense, their Czech counterparts are listed. For each such pair, all the aligned valency slots are listed and referred to by the functor assigned to the slot. So far, Czengvallex contains only the alignment of verb pairs, though an extension covering other parts of speech is planned.

2.3 Functional Generative Description Valency Theory

The Czengvallex has been built using the valency theory developed within the Functional Generative Description approach—the Functional Generative Description Valency

1 An evaluative state is a part of text where the speaker expresses evaluation towards any entity. An evaluative state consists of the Source, Target and Evaluative Expression.

Theory (FGDVT). The basics of the approach describe e.g., Lopatková and Panevová (2004). The FGDVT sees valency as a special relation between a governing word and its dependents, combining a syntactic and semantic approach for distinguishing valency participants. Verbs are considered to be the core of the sentence, governing both the morphological properties of their dependents and their semantic interpretation. The number and realization of the participants constituting the valency structure of the phrase is represented by valency frames. In the frame, each participant is represented by a functor, which is a label stating the value of a corresponding deep syntactic dependency relation, as well as expressing the function of the participant in the clause. Participant labels consist of two groups, the so called *inner participants* (Actor, Patient, Addressee, Origin and Effect) and *free modifications* (Cause, Location, Direction, etc.). Inner participants are considered as constituting the valency frame in any case, whether they are obligatory or optional. Free modifications belong to the valency elements only if they are obligatory. The first two positions in the valency frame, the Actor (ACT) and the Patient (PAT), are connected with no specific globally defined semantics. As a result, the FGDVT adopts the concept of shifting of "cognitive roles." According to this rule, the roles of Addressee (ADDR), Effect (EFF) and Origin (ORIG) are being shifted to the PAT position in case the verb has only two arguments, or any of the inner participant roles to the ACT position in case there is only one position in the frame.

3. Enhancing the Lexicon

In the process of lexicon enhancement, we utilized English lemmas from Czech Sublex, i.e., the original source English lemmas used in the task of Czech Sublex creation that correspond to the final lemmas included in Czech Sublex after manual cleanup. We used the lemmas as an input for the search of corresponding Czengvallex frame pairs. After sorting out translations already present in Czech Sublex, we gained 1,166 new verb translations corresponding to 578 unique lemmas. These 578 lemmas we subjected to manual cleanup, after which we ended up with 222 new true subjective lemmas to be included in Czech Sublex.

4. Analysis

4.1 Semantic Classes Reflecting the Type of Evaluative Meaning

We have analyzed the outcoming 222 verbs focusing on the question of which of the participants inherits the role of the Target of evaluation. Our first idea was to gather verbs into groups according to the functor label of the participant, to which the sentiment value is propagated by the verb. Nevertheless, due to the formal feature of cognitive role shifting (described briefly in Section 2.3) which in some cases re-labels the participants according to their syntactic closeness to the verb, this did not prove advantageous. Therefore, we decided to split the verbs into two categories onlythe first propagating the sentiment to the ACT position (making it the Target of sentiment), the second to a

non-Actor position (mainly PAT, ADDR, but also any other semantic modification, like, e.g., Cause (CAUS), appearing syntactically in the position of the second argument, labeled PAT in the FGDVT).

Within these two groups, we identified semantically close verb candidates that formed tight semantic classes considering the type of evaluative meaning carried by the verb. These classes did not match any previously created semantic classification, therefore, we used our own labels for their descriptions.

Verbs of success/failure[2] propagate sentiment to the ACT position (1a, b). In case they have two participants in the frame (1c, d), the PAT position is usually occupied by expressions of inherent "objective" sentiment value, also known as "Good/Bad news" (Veselovská et al. 2012, 300). This group of verbs includes, e.g., the lemmas *polepšit si* "improve one's position," *prospět* "benefit," *těžit* "profit," *užít/užívat si* "enjoy," *vychutnat si* "relish," *vydařit se* "turn out well," *zasloužit si* "deserve."

(1) (a) Večírek$_{\text{ACT-TARGET}}$ se vydařil$_{\text{EVAL}}$.
party REFL turned-out-well
"The party turned out well."

(b) Andrej Babiš$_{\text{ACT-TARGET}}$ pochybil$_{\text{EVAL}}$, když nepřiznal střet zájmů.
Andrej Babiš made-a-mistake when he-not-addmitted conflict interests
"Andrej Babiš made a mistake when he didn't acknowledge a conflict of interests."

(c) Williamsová$_{\text{ACT-TARGET}}$ si vychutnala$_{\text{EVAL}}$ vítězství$_{\text{PAT-GOODNEWS}}$ nad soupeřkou.
Williams relished victory over opponent
"Williams relished the victory over her opponent."

(d) Kvitová$_{\text{ACT-TARGET}}$ doplatila$_{\text{EVAL}}$ na svou nepřipravenost$_{\text{PAT-BADNEWS}}$.
Kvitová paid for her unreadiness
"Kvitová paid for her not being ready."

Verbs of Improvement/Deterioration, e.g., *zdokonalovat* "improve," *zkvalitnit* "enhance," *rozšířit* "extend," and *znásobit* "multiply," represent a class of verbs that balance on the verge between true evaluation and the category of Good/Bad News. It can be said that their evaluative strength is strongly dependent on the context. In evaluative contexts, they usually propagate sentiment value to the ACT position.

2 For some of the classes we use a joint label for both verbs expressing positive and verbs expressing negative polarity value since the polarity orientation of the verb may be simply turned over in a text by means of a negation prefix.

Verbs of Helping/*Harming* (e.g., *nahrát* "do sb. a good turn," *napomáhat* "assist," *deptat* "get sb. down," *znehodnotit* "destroy") and *Verbs of Praising*/*Disdain* (e.g., *cenit si* "appreciate," *přimlouvat se* "intercede," *bagatelizovat* "downplay," *poplivat* "denigrate," *ztrapnit* "embarrass") are (at least) ditransitive. Both classes share an interesting feature. They are subject to dual interpretation depending on the level of sentiment analysis desired. Either the Source of evaluation is sought within the sentence, then the ACT of the verb is considered the Source. Or, the Source of evaluation is sought outside the sentence, then it is the Author of the text which is considered the Source of evaluation and center of sentiment perspective. In the first case (lower level), the ACT as the Source of evaluation expresses his/her opinion verbally or in an action towards the non-Actor position (usually PAT) as the evaluated Target. In the second case (higher level), the Author of the text evaluates the ACT of the sentence for its involvement in a positive or negative action towards the other entity. The ACT then may be considered the Target of evaluation, while the other participant is perceived as carrying either no specific sentiment value, or a slight value of different orientation than the evaluative state expressed. Thus in (2a), the media is presented as expressing its negative evaluation of the president through the act of verbal attack, whereas in (2b), from the perspective of the Author/Reader, the media may be perceived negatively due to its involvement in a negative act (attack), whereas the president may be pitied as the victim of the act.

(2) (a) Média$_{\text{ACT-SOURCE}}$ opět napadají$_{\text{EVAL}}$ prezidenta$_{\text{PAT-TARGET}}$.
media again assault president
"The media once again assault the president."

(b) Média$_{\text{ACT-TARGET}}$ opět napadají$_{\text{EVAL}}$ prezidenta$_{\text{PAT}}$.
media again assault president
"The media once again assault the president."

Nevertheless, we must bear in mind that the use of Author perspective is strongly influenced by individual subjective attitudes of the Author (or Reader) and is therefore quite difficult to interpret. Therefore, it is usually avoided in the tasks for which Czech Sublex was designed originally.

Verbs of (*Dis*)*Liking* include verbs describing the feeling of liking either from the perspective of the experiencer, "liker," or from the perspective of the liked thing. The first group includes lemmas such as *zamilovat se* "fall in love" and *oblíbit si* "start liking." Here the verbs propagate sentiment value to the PAT (or, more generally, non-Actor) position, whereas the ACT position is occupied by the Source of evaluation. The second group includes lemmas such as *pobláznit* "craze," *uspokojovat* "satisfy," *zalíbit se* "appeal," and *odstrašit* "scare." In this instance, the ACT position is occupied by

the "liked" thing, thus being the Target of evaluation, whereas the non-Actor position is usually reserved for the Source.[3]

Verbs of Wanting express a desire (not) to own something or (not) to perform an action of some kind, including the implicit positive (or negative) attitude to the thing or action. This class includes lemmas such as *chtít* "want," *dožadovat se* "call for," *přát si* "wish for," *toužit* "long for," and *žádat* "plead."As for the Target of sentiment, this class behaves in an uncomplicated way, propagating the sentiment value to the non-Actor position (making it the Target of evaluation) and filling the ACT position with the Source of evaluation.

Verbs of Struggle, including the lemmas *potýkat se* "cope with," *přečkat* "endure," *protrpět* "suffer through," *strpět* "stand," and *vydržet* "bear," propagate negative polarity value to the PAT as the Target, while the ACT position is occupied by the Source.

Verbs of Judgment include two types of verbs. First, there are verbs with clear positive or negative value, such as *zazlívat* "hold st. against," *obhájit* "defend," and *vyčíst* "reproach."These verbs propagate sentiment to the non-Actor position (PAT or ADDR, usually) and the Source occupies the ACT position. Second, there are verbs in this group which express opinion, but without a clear positive or negative orientation, such as *posoudit* "assess," *přehodnotit* "revise," etc. In our approach, such verbs are marked as "Elusive Elements", i.e., elements which are evaluative, but it is not possible to decide their polarity value. These verbs appeared in the final collection of lemmas due to the fact that their English counterparts carried context polarity, i.e., their meaning was to be interpreted as evaluative or elusive, or even neutral, depending on the specific context in the sentence.

Sometimes, the polarity value was reduced in intensity, or even disappeared in the translation. In (3), the original English verb carries an evaluative meaning that might be considered as having a NONNEG polarity (the verb meaning implies downgrading of a strongly negative polarity). The Czech translation then only describes a certain nonspecific shift in polarity, not offering any specific information about the polarity orientation without a prior context.

(3) (a) $\text{He}_{\text{ACT-SOURCE}}$ $\text{softened}_{\text{EVAL}}$ $\text{the talk}_{\text{PAT-TARGET}}$ about a recession.

3 Unfortunately, the FGDVT treats valency frames of verbs of this group differently. For verbs expressing the "experiencer" of the liking feeling in direct case (accusative), the frame constitutes of an ACT tied to the subject position and the PAT tied to the object position. Nevertheless, for verbs like *zalíbit se* "appeal," which express the "experiencer" in an oblique case (dative), the syntactical subject position is labeled PAT, whereas the oblique object is considered ACT.

(b) Svá slova$_{\text{PAT}}$ o recesi přehodnotil$_{\text{ELUSIVE}}$.
his words about recession he-revised
"He changed his mind about his talk about a recession."

The last group to be mentioned here are the *Verbs of Communication*. Their meaning is the same as that of ordinary Communication Verbs, i.e., sharing a verbal message with another entity. Evaluative Communication Verbs though involve a semantic indication of the positive or negative attitude of the speaker. The group contains lemmas such as *brblat* "grumble," *čertit se* "talk in an angry manner," *lkát* "lament," *libovat si* "talk in a pleased manner," *ohradit se* "object," *pochvalovat si* "praise," *stěžovat si* "complain," *žalovat* "tell on sb.," etc.This class, unfortunately, does not behave homogenously enough considering the number and type of positions in the valency frame. The Target of the sentiment expressed by these verbs is in most cases the "message" (usually occupying the EFF or PAT position), as in (4a), for some verbs, it is the semantic addressee (ADDR or PAT position) (4b), in some cases, the Target of the sentiment may even be split into both the positions of the PAT and EFF (4c).

(4) (a) Stěžoval$_{\text{EVAL}}$ si úřadům$_{\text{ADDR}}$ na nedostatek informací$_{\text{PAT-TARGET}}$.
he-complained REFL authorities$_{\text{DAT}}$ about lack information
"He complained to the authorities about the lack of information."

(b) A já$_{\text{ACT-SOURCE}}$ s panem ministrem$_{\text{PAT-TARGET}}$ musím polemizovat$_{\text{EVAL}}$.
and I with Mr. minister must argue
"But I must disagree with the minister."

(c) Stěžoval$_{\text{EVAL}}$ si jim$_{\text{ADDR}}$ na syna$_{\text{PAT-TARGET}}$, že lže$_{\text{EFF-TARGET}}$.
he-complained REFL they$_{\text{DAT}}$ about son that he-lies
"He complained to them about his son's constant lying."

All the above mentioned classes apply also to the verbs in the original Sublex.

4.2 Target Functor Mismatch

In Czengvallex, it is often the case that frame elements do not align proportionally. There are two general types of disproportion in the data. In some cases, the aligned frame elements do not match in value; we call this type a "functor mismatch." In the case of the other type, we term it a "zero alignment," one or more frame elements do not have a counterpart in the parallel frame (Šindlerová et al. 2015).

Both functor mismatch and zero alignment are often caused by conversive translations, i.e., the translated verb depicts the situation from a different perspective than the original one. In the text, though, the perspectives match because one of the verbs

is used in an agent-backgrounding[4] diathesis (e.g., a passive, as in [5]), while the other verb involves the backgrounding of the semantic agent already in its unmarked form.

(5) (a) An airline buy-out bill$_{\text{PAT-TARGET}}$ was approved$_{\text{EVAL}}$ by the House$_{\text{ACT-SOURCE}}$.

(b) Zákon$_{\text{ACT-TARGET}}$ o skupování aerolinek prošel$_{\text{EVAL}}$ Sněmovnou$_{\text{DIR2-SOURCE}}$ …
law about buying-out airlines passed through-Parliament
"An airline buy-out bill was approved by the House."

Example (6) represents another case of zero alignment influencing not only the labeling of the Target, but also the evaluative strength of the translated verb. Both sentences include a semantic backgrounding of an Agent position. The English sentence involves a participial verb form implying a covert, coreferential agent. The Czech translation then changes the perspective of the clause from "the management" to "the employees," choosing an intransitive verb. The loss of the implicit Source of evaluation results in the lowering of the evaluative strength of the verb.

(6) (a) . . . by eliminating$_{\text{EVAL}}$ the typically long New York commutes$_{\text{PAT-TARGET}}$ between office and home, management will expect employees to work 40 hours a week in Dallas, rather than a 35-hour work week in New York . . .

(b) Díky tomu, že odpadne$_{\text{EVAL}}$ typicky dlouhé
thanks it that will-not-take-place typically long
newyorské dojíždění$_{\text{ACT-TARGET}}$ mezi kanceláří a domovem,
New-York$_{\text{ADJ}}$ commute between office and home
bude management od zaměstnanců v Dallasu
AUX.FUT management from employees in Dallas
očekávat 40hodinový pracovní týden …
expect 40-hour-long working week

"By eliminating the typically long New York commutes between office and home, management will expect employees to work 40 hours a week in Dallas…"

4 Some participants receive a semantic priority in the situation perspective. As such, they tend to be overtly expressed, receive prominent syntactical positions (e.g., subject, object) and prominent morphological forms (e.g., direct case). Others are linguistically constructed as being "in the background of" the situation, they are perceived as not necessary for the interpretation, too general, etc. They tend to remain unexpressed in the sentence, or they receive oblique morphological forms and syntactic positions outside the valency frame, etc.

Thus, it may happen that in the process of translation, the loss of the Source of evaluation from the evaluative meaning may lead to the use of a verb that implies a Source interpreted as the Author of the text, or even to a complete loss of evaluative strength of the verb. Nevertheless, it is highly improbable that a zero alignment in the data would affect the Target at any time.

Conversive translations do not constitute a notable portion of PCEDT verb translations, but this is probably caused by the fact that the English translations of the Czech sentences in the treebank were made with a special regard to the treebank purpose, and the maximal possible syntactic similarity to the original sentence was explicitly declared in the instructions. In commonly produced translations, we expect more substantial portion of conversive translations to appear.

Another type of mismatch is represented by the Abstract Cause-Subject Alternation, where a single lemma may function in dual perspective configuration. One (the causative) having the semantic agent in the syntactic subject, semantic patient in the object and an oblique cause, which is also affected by the sentiment value as a "secondary target." The other involves the abstract cause shifted into the subject position and a strongly backgrounded agent, see e.g., the case of the verb *pobuřovat* "to offend" in (7).

(7) (a) A poll of South Koreans showed overwhelming opposition to efforts to curb dog-meat consumption just because it$_{\text{ACT-TARGET}}$ offends$_{\text{EVAL}}$ foreigners$_{\text{PAT-SOURCE}}$.

(b) Z průzkumu veřejného mínění mezi Jihokorejci
from poll public opinion among South-Koreans

vyplynulo, že většina obyvatel je proti
followed that majority residents is against

snahám ukončit konzumaci psího masa jen proto, že
efforts to-end consumption dog$_{\text{ADJ}}$ meat only because that

to$_{\text{ACT-TARGET}}$ cizince$_{\text{PAT-SOURCE}}$ pobuřuje$_{\text{EVAL}}$.
it foreigners offends

"A poll of South Koreans showed overwhelming opposition to efforts to curb dog-meat consumption just because it offends foreigners."

(c) Z průzkumu veřejného mínění mezi Jihokorejci
from poll public opinion among South-Koreans

vyplynulo, že většina obyvatel je proti
followed that majority residents is against

snahám ukončit konzumaci psího masa jen proto, že
efforts to-end consumption dog$_{ADJ}$ meat only because that

tím$_{MEANS\text{-}TARGET}$ cizince$_{PAT\text{-}SOURCE}$ pobuřují$_{EVAL}$ [oni]$_{ACT\text{-}TARGET}$.
it$_{INS}$ foreigners they-offend [they]

"A poll of South Koreans showed overwhelming opposition to efforts to curb dog-meat consumption just because it offends foreigners."[5]

4.3 Lost Evaluation

Even though the analysis of truly evaluative verbs brought some significant findings, it may be even more interesting to look at the lemmas that were excluded during the manual revision. Since most of them originated as corpus translations of evaluative verbs, it was initially uncertain as to what happened to the originally subjective content during the translation. We were able to identify four major reasons explaining why the translated verbs did not come out as evaluative.

Sometimes, the verb lost its subjectivity during the translation, while the subjectivity was transferred to another participant or a verb modifier in the text. This happened especially when the translation included a light verb construction, or another instance of a semantically general verb in combination with an evaluative nominal element (phrasemes etc.), i.e., a single English verb was translated by a combination of a Czech verb and another lexical item to which the evaluation was transferred, see (8).

(8) (a) Americans$_{ACT\text{-}SOURCE}$ didn't dislike$_{EVAL}$ metrics$_{PAT\text{-}TARGET}$; they simply ignored them.

(b) Ne že by Američané$_{ACT\text{-}SOURCE}$ neměli metrický
not that AUX.COND Americans not-had metric

systém$_{PAT\text{-}TARGET}$ rádi$_{DPHR\text{-}EVAL}$, oni jej prostě ignorovali.
system like$_{ADJ}$ they it simply ignored

"Americans didn't dislike metrics, they simply ignored them."

A different case is represented by originally evaluative verbs that were used in metaphorical, non-evaluative contexts, or specific jargon, in the parallel treebank, and therefore,

5 The translation in (7c) is a possible variant not appearing directly in the PCEDT.

they got into the pairing with a non-evaluative verb in the Czengvallex and the non-evaluative lemma was then harvested into the candidate list, see the case of the verb "to enjoy" translated by *zaznamenat* "to notice" in (9).

(9) (a) Hawker Siddeley said its core electrical products division enjoyed strong growth, with a 20% rise in operating profit during the period.

(b) Společnost Hawker Siddeley oznámila, že její divize základních
company Hawker Siddeley announced that her division basic

elektrických produktů zaznamenala silný nárůst s 20%
electrical products noticed strong growth with 20%

vzrůstem provozního zisku během tohoto období.
growth operating profit during this period

"Hawker Siddeley said its core electrical products division enjoyed strong growth, with a 20% rise in operating profit during the period."

With verbs that did not possess an inherent, prior polarity, but only a "functional" context polarity, it was simply the case that the translation of the non-evaluative meaning of the lemma was collected, see (10a, b) in contrast to (10c, d).

(10) (a) Market Airlines tried to restrict the program substantially by limiting$_{\text{NONEVAL}}$ the offer$_{\text{PAT}}$ to certain days$_{\text{EFF}}$ of the week.

(b) Aerolinie se pokoušely tento program značně
airlines REFL tried this program substantially

omezit$_{\text{NONEVAL}}$ tím, že nabídku$_{\text{PAT}}$ vymezily na
restrict it$_{\text{INS}}$ that offer they-limited to

některé dny$_{\text{TOWH}}$ v týdnu.
some days in week

"Market Airlines tried to restrict the program substantially by limiting the offer to certain days of the week."

(c) Advocates hope that such standards will improve treatment while limiting$_{\text{EVAL}}$ unnecessary tests$_{\text{PAT-TARGET}}$ and medical procedures$_{\text{PAT-TARGET}}$.

(d) Zastánci doufají, že tyto normy zlepší léčbu,
advocates hope that these norms will-improve treatment

když omezí$_{EVAL}$ zbytečné testy$_{PAT}$ a lékařské procedury$_{PAT}$.
when they-limit unnecessary tests and medical procedures

"Advocates hope that such standards will improve treatment while limiting unnecessary tests and medical procedures."

And last, but not least, there was a number of English lemmas in the source material to Czech Sublex creation that were more "subjectivity clue verbs" lacking directly evaluative features. This applied especially to plain verbs of communication, such as *prohlašovat* "claim" or *uvádět* "state This is connected to the fact that the Czech Sublex is aimed at a substantially narrower concept of evaluation than the original MPQA lexicon.

5. Conclusions and Future Work

The analysis suggests that the relation between the valency frame patterns of evaluative verbs and the positions of the Source and Target of evaluation is complex. Generally, the verbs of similar evaluative meaning (verbs within our evaluative "semantic classes") propagate sentiment to the same participants of the frame. Nevertheless, we have seen that there are some complicated cases.

For some of the identified verb classes of evaluative meaning, we observed that the sentiment value might attach to more than one position in the frame. There are essentially three major cases:

- One position in the verb frame is occupied by the Target, the other by an inherently Evaluative Expression, a Bad News/Good News item (Verbs of Success/Failure).
- The two affected positions receive a dual interpretation, depending on whether we choose the Author/Reader perspective, or the "Source in the text" perspective (Abstract Cause-Subject Alternation).
- The Target of the sentiment is split evenly between two positions in the frame (PAT and EFF of some Communication Verbs).

Also, the analysis partially answered the question regarding what happens to sentiments in translation. We have seen that in translating evaluative states, we come across numerous evaluation-changing phenomena, starting from the change of situation perspective, propagation of sentiment value to different participants, shifts from prior polarity verbs to context polarity verbs, lowering evaluative strength, and even the complete loss of sentiment value.

Considering the description of evaluative state using valency, and the above mentioned shifts in translation, it eventually appears that the FGDVT framework might not be the most suitable one for relating the positions of the evaluation Target and Source

to syntactic participants of a kind because of its formal feature of shifting of cognitive roles, i.e., the formal labeling of obligatory participants on the basis of their position in the sentence, not their original, semantic value. Also, the syntactic actant labeling is strongly dependent on the morphosyntactic form of the expression, thus, e.g., a dative experiencer is likely to be labeled ACT, whereas an accusative experiencer is prohibited from being labeled ACT. Therefore, our semantic classes are not homogeneous with respect to the syntactic labeling of the Source and Target though they involve similar semantic participants.

In the future, we would like to enrich and deepen our analysis using the annotated parallel treebank data, focusing on the syntactic mismatches in evaluative constructions in a greater detail.

Funding Acknowledgement

This work was supported by the Czech Science Foundation (GAČR), project number GA15-06894S, and by the SVV project number 260 333. This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth, and Sports of the Czech Republic (project LM2015071).

Works Cited

Arora, Piyush, Akshat Bakliwal, and Vasudeva Varma. 2012. "Hindi Subjective Lexicon Generation Using WordNet Graph Traversal." *International Journal of Computational Linguistics and Applications* 3 (1): 25–39.

Bojar, Ondřej and Zdeněk Žabokrtský. 2006. "CzEng: Czech-English Parallel Corpus Release Version 0.5." *Prague Bulletin of Mathematical Linguistics* 86: 59–62.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. "Announcing Prague Czech-English Dependency Treebank 2.0." *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*: 3153–60. Accessed June 1, 2016. http://www.lrec-conf.org/proceedings/lrec2012/index.html.

Kanayama, Hiroshi and Tetsuya Nasukawa. 2006. "Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis." In *Proceedings of EMNLP 2006*, edited by Dan Jurafsky and Eric Gaussier, 355–63. Stroudsburg, PA: The Association for Computational Linguistics.

Lopatková, Markéta, and Jarmila Panevová. 2004. "Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank." In *Insight into Slovak and Czech Corpus Linguistic*, edited by Mária Šimková, 83–92. Bratislava: Veda, Publishing House of the Slovak Academy of Sciences.

Milhacea, Rada, Carmen Banea, and Janyce Wiebe. 2007. "Learning Multilingual Subjective Language via Cross-Lingual Projections." In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, edited by Antal van den Bosch and Annie Zaenen, 976–83. Stroudsburg, PA: Association for Computational Linguistics.

Šindlerová, Jana, Eva Fučíková, and Zdeňka Urešová. 2015. "Zero Alignment of Verb Arguments in a Parallel Treebank." In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, edited by Eva Hajičová and Joakim Nivre, 330–39. Uppsala: Uppsala University.

Šindlerová, Jana, Kateřina Veselovská, and Jan Hajič, Jr. 2014. "Tracing Sentiments: Syntactic and Semantic Features in a Subjectivity Lexicon." In *Proceedings of the XVI EURALEX International Congress: The User in Focus*, edited by Andrea Abel, Chiara Vettori and Natascia Ralli, 405–13. Bolzano: Institute for Specialised Communication and Multilingualism.

Urešová, Zdeňka., Eva Fučíková, and Jana Šindlerová. 2016. "CzEngVallex: A Bilingual Czech-English Valency Lexicon." *The Prague Bulletin of Mathematical Linguistics*, 105 (1): 17–50.

Veselovská, Kateřina. 2013. "Czech Subjectivity Lexicon: A Lexical Resource for Czech Polarity Classification." In *Proceedings of the Seventh International Conference Slovko*, edited by Katarína Gajdošová and Adriána Žáková, 279–84. Lüdenscheid: RAM Verlag.

Veselovská, Kateřina, Jan Hajič, Jr., and Jana Šindlerová. 2012. "Creating Annotated Resources for Polarity Classification in Czech." In *Empirical Methods in Natural Language Processing. Proceedings of the Conference on Natural Language Processing (KONVENS) 2012*, edited by Jeremy Jancsary, 296–304. Wien: ÖGAI.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis." In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP) 2005*, edited by Joyce Chai, 347–54. Madison, WI: Omnipress Inc.

English Translation Counterparts of the Czech Particles *copak, jestlipak, kdepak*

Denisa Šebestová[a] and Markéta Malá[b]

Charles University, Prague, Czech Republic

[a]sebestovadenisa@gmail.com; [b]marketa.mala@ff.cuni.cz

Abstract: The paper examines English translation counterparts of Czech sentences containing the particles *copak, jestlipak* and *kdepak*, which represent elements of the "third syntactical plan," i.e., they "place the content of the sentence in relation to the individual and his special ability to perceive, judge and assess" (Poldauf 1964, 242). This paper a) identifies and describes English means which perform the same communicative functions as Czech sentences containing the *-pak* particles; b) specifies the functions of the Czech particles. The paper employs contrastive analysis, which allows a comparison of meanings that stem from the same notions and serve the same communicative functions but are conveyed by different means in the respective languages. The affix *-pak* is shown to be a polyfunctional indicator of communicative function (Grepl and Karlík 1998): the *-pak* particles have content/speaker-related functions as well as communication/addressee-oriented functions (Kranich and Gast 2015).

Keywords: third syntactical plan; translation counterparts; Czech particles

1. Introduction

The present paper examines English translation counterparts of Czech sentences containing the particles *copak, jestlipak* and *kdepak*. These particles share the expressive and intensifying postfix *-pak* (Dokulil et al. 1986; Komárek et al. 1986). A postfix is defined as a type of affix which follows an inflectional suffix (Karlík, Nekula and Rusínová 2000, 109).

The examined particles have been described as elements of the "third syntactical plan," i.e., "components which place the content of the sentence in relation to the individual and his special ability to perceive, judge and assess" (Poldauf 1964, 242). The third syntactical plan is fully developed in Czech, but to a much lesser degree in English. This paper has a twofold aim: a) to identify and describe English means which perform

the same communicative functions as Czech sentences containing the *-pak* particles, b) to further specify the functions of the Czech particles.

With regard to the findings of Dušková et al. (2012) and Poldauf (1964), English may be expected to prefer syntactic means (specific grammatical structures) of expressing the speaker's stance and evaluation where Czech employs lexical means, cf. particles.

The particle *jestlipak* marks the polar question it introduces as deliberative, i.e., the speaker is considering whether or not the content of the question is true (Dušková et al. 2012, 313). A similar function was shown to be performed by English sentences introduced by *I wonder* (Poldauf 1964, 253; Dušková et al. 2012, 313), which are therefore likely to occur as a frequent translation counterpart of the *jestlipak* questions.

The English translation counterparts of Czech sentences introduced by the particle *copak*/*cožpak* are likely to include rhetorical polar questions, i.e., clauses which are formally identical with questions but do not prompt the addressee to provide a reply. Rhetorical questions are emotionally expressive; their illocutionary force is an emphatic assertion of the reversed polarity (Dušková et al. 2012, 316). According to Dušková et al., rhetorical polar questions are similar in their function to indicative clauses with a question tag of the opposite polarity (ibid.). Poldauf (1964, 254) also refers to question tags as elements of the English third syntactical plan. They may, therefore, be expected as another type of translation counterpart of *copak*.

While both *copak* and *jestlipak* introduce interrogative sentences, *kdepak* occurs in sentences of the declarative type. *Kdepak* is classified as an epistemic modal particle (Komárek et al. 1986, 233), expressing the speaker's certainty that the content of the clause is not true.[1]

Words with the postfix *-pak* are generally described as expressive (Komárek et al. 1986, 393). Postfixes which evolved from enclitic particles, such as *-pak*, are frequently used in spoken language (Balhar et al. 2011, 570). *Jestlipak* is characterized as colloquial (Filipec and Kroupová 2005, 121; Trávníček 1951, 657), an element of "common Czech" (Havránek et al. 1960, 786), i.e., the variety of the Czech language which is most frequently used in spontaneous everyday spoken discourse (Karlík, Nekula and Pleskalová 2002, 81).

2. Material and Method

The material was drawn from the fiction and drama core of the parallel translation corpus *InterCorp*, version 9 (2016). Our search was limited to texts whose source language is Czech[2] and their English translations. The numbers of instances of individual particles are given below in Table 1. We have included variant forms of the particles which were

1 Štícha et al. (2013, 534) also mention the possible contrastive function of *kdepak.*

2 The size of the subcorpus defined for the purposes of the present research (31 Czech original fiction and drama texts) was 2,915,456 tokens.

present in the corpus, namely the dialectal or colloquial forms *cák*, *depak*, *depák*, as well as the variant *cožpak*. The instances of *copak* also included the particle with the attached morpheme *-s,* which is a contracted form of the auxiliary *být* (*be*) in the second person singular (1).

Copak may occur as a particle, interrogative pronoun (2), or an interjection (3).[3] In the present paper we are only interested in particles. Therefore, irrelevant instances of *copak* were not included in our analysis. We have also excluded examples of the type presented in (4), which we interpret as elliptical constructions with the pronoun *copak* in Czech.

(1) (a) Copaks do svých padesáti let neviděl ženskou jen tak?
(b) Haven't you ever seen a naked woman before in all your fifty years?

(2) (a) Copak jste jí udělal?
(b) What have you done to her?

(3) (a) Copak, snad se nebojíte?
(b) You're not scared, are you?

(4) (a) A copak Angela Davis?
(b) And what about Angela Davis?

particle	number of instances	total number
jestlipak	34	34
copak(s)	251	270
cožpak	12	
cák	7	
kdepak	77	86
depak/*depák*	9	
total		390

Table 1. Instances of the *-pak* particles analysed in this study

3 We classify instances of *copak* of this type, separated from the clause by a comma, as interjections, in accordance with Havránek et al. (1960, 222).

We employ the methodology of contrastive analysis, which allows for a comparison of meanings that stem from the same notions and serve the same communicative functions but are conveyed by different means in the respective languages, since "linguistic structure is language-specific while the cognitive and functional-communicative substance which constrains it is potentially universal" (Boye 2012, 7; cf. Haspelmath 2010; Malá 2013; Šebestová and Malá 2016).

3. Analysis

3.1 *Jestlipak*

Based on Poldauf's (1964) observations, we expected English sentences introduced by *I wonder* to occur as the major translation counterparts of *jestlipak*. This hypothesis was confirmed as 47.1 per cent of counterparts indeed contained this construction. One of them was sentence-final (5), others were in the prototypical introductory position. The final position indicates a lower degree of integration of *I wonder* into the sentence (cf. Poldauf's (1964, 253) description of *I wonder* as an "introductory/epenthetic marker"), which may point towards its discourse particle status (cf. Aijmer 2013).

(5) (a) Jestlipak to ještě dovedu, bejt mlsná.
(b) Do I still have a sweet tooth? I wonder.

We have identified two distinct uses of *jestlipak* in our material, which correspond to two uses of its most frequent English counterpart *I wonder*. These two different uses of the English verb *to wonder* can be defined as follows:

> A. *wonder about something* "to think about something and try to decide what is true, what will happen, what you should do, etc." (*Oxford Advanced Learner's Dictionary* 2005, 1693)
>
> B. *wonder [wh-]* "used as a polite way of asking a question or asking somebody to do something" (ibid.)

Firstly, *I wonder* may mark an utterance as deliberative, i.e., the speaker poses a question to himself (Štícha et al. 2013, 763). This use (6) corresponds to definition A.

(6) (a) Jestlipak vůbec ví, že je vlastně král?
(b) I wonder if he knows he's a King?

Definition B can be illustrated by (7), where *jestlipak*/*I wonder* is used for establishing or maintaining contact (7). The contact function of *jestlipak* is made explicit in the English counterpart (*tell me*) in (8).

(7) (a) Jestlipak ti dají také zakouřit . . .
(b) I wonder if they will give you a smoke . . .

(8) (a) Ale jestlipak poznáš, co to je za keř?
(b) But tell me this, do you recognise this bush?

The English counterparts of *jestlipak* show that, at the same time, the particle may function as a means of indicating tentativeness or politeness: in (9b) it is the past tense that serves as a politeness/tentativeness marker (cf. Dušková et al. 2012, 223). The same function can be performed by modal verbs (*could* in [10], *would*) and epistemic adverbials in the English translations (*by any chance* in [11], *perhaps*).

(9) (a) … jestlipak víte, že jedu za tři dny na Slovácko…
(b) … did you know I was going to Moravia...

(10) (a) Jestlipak znáte časopis Svět zvířat?
(b) Could it be that you know the magazine The Animal World?

(11) (a) Jestlipak znáte ještě vzoreček pro výpočet plochy kruhové výseče?
(b) Do you recall, by any chance, the formula for calculating the area of a sector?

On the other hand, similarly to other discourse markers (Aijmer and Altenberg 2002), the function of *jestlipak* may be lost in the translation: 26.5 per cent of *jestlipak* sentences have zero counterparts (viz. unmarked positive polar questions).

3.2 *Copak*

The English counterparts of *copak* can be divided into three main groups.

- Rhetorical questions

Our material contained rhetorical questions of both polarities[4] (12), (13). The communicative function of the rhetorical question is an assertion of the opposite polarity (Dušková et al. 2012, 316). Typically, the subject has generic reference (*all women* in [12], *we* in [13]), and the verb is in the atemporal present simple tense.

Some of the positive rhetorical questions contained epistemic modal verbs (*can*) or the epistemic content disjunct *really* (13) (Quirk et al. 1985, 621). These support

4 Most of the rhetorical questions were polar. The *wh*-questions (always positive) were represented marginally, e.g., *Copak poznám složenou básničku od napsaný?—How can I tell a composed poem from a written one?* They are often introduced by *how could*, *how can.*

the appeal function of the rhetorical question: the speaker appeals to the addressee to reaffirm the speaker's view.

(12) (a) Copak netrpí všechny ženy měsíčním krvácením?
(b) Don't all women suffer from monthly bleeding?

(13) (a) Copak je nutné se starat – dnes, kdy se konečně může říkat všechno – komu nahraje pravda?
(b Do we really have to worry – today, when at last everything can be said – about those whose hands the truth plays into?

- Negative polar questions

Negative polar questions express a change in the speaker's assumption concerning the validity of the statement (14)—the speaker expected the statement to be true but the new context leads him to reassess the situation (Dušková et al. 2012, 314). At the same time, the speaker appeals to the addressee to confirm the speaker's inference. *Copak* here functions as a pragmatic presupposition trigger (Hirschová 2013).

Where the likely re-interpretation is contrary to what the speaker considered appropriate or advisable, the appeal is combined with an overtone of reproach (signalled by the exclamation mark in [15]).

(14) (a) Copak nechápete?
(b) Don't you understand?

(15) (a) Copak nevidíš, že je nemocný!
(b) Can't you see that he is sick!

- Declarative clauses

The English declarative clauses used as counterparts of *copak* questions have the opposite polarity with respect to the Czech question. The majority of these declarative clauses were negative or comprised lexical negators (*hardly* in [16]).

Declarative clause counterparts directly correspond to the communicative function of the Czech rhetorical question introduced by *copak.* The given statement is expressed explicitly (17). Often, these sentences contained an emphatic, sometimes emotionally expressive element, such as *for Heaven's sake* conveying irritation (18). These counterparts highlight the emphatic and emotionally expressive character of *copak*.[5]

5 This emotional expressivity may contain the speaker's negative evaluation of the addressee's supposed attitude (15).

Alternatively, declarative clause counterparts of *copak* questions may function in a way similar to negative polar questions, signalling a change in the speaker's assumption demanded by the context (as indicated by *I thought* in [19]).

(16) (a) „Životní štěstí," řekl jsem posléze bezradně, „– copak to jde vyučovat?"
(b) "Happiness—" I eventually said nonplussed, "—that's hardly something you can teach."

(17) (a) Copak to potřebuju?
(b) I don't need that kind of trouble.

(18) (a) Copak jsem pořád malé dítě?
(b) For Heaven's sake, I'm not a child any more!

(19) (a) Copak ty nejsi posrpnovej, Franku?
(b) I thought you were post-invasion yourself, Frank.

The declarative clause may be followed by a question tag of the reversed polarity, whose function is maintaining contact with the addressee, as well as appealing to the addressee to confirm or refute the given statement (20).[6] This appeal can also be expressed explicitly in sentences introduced by (*do*) *you mean . . . ?* (21).

(20) (a) Copak jsem se tvářil andělsky?
(b) I didn't make an angel face, did I?

(21) (a) „Copak Lucii nemiluješ?" zeptal se Harýk.
(b) "You mean you don't love Lucie?" said Haryk.

- Inferentials

Other significant counterparts of *copak* questions included inferential constructions of the types *is it that . . .?* / *could it be that . . .?* (Delahunty 1995). These constructions reflect the use of *copak* as a marker of epistemic modal meaning whereby the statement is labelled as the speaker's inference (22).

(22) (a) Ty vopice jedna, copak myslíš, že se budu jen s tebou bavit?
(b) You singular monkey, is it that you think that I'd be prattling with you?

6 The tag may be a general extender, such as *or what, or something* following a question (3 cases in our material): *Copak jste němý?—Are you dumb, or what?*

- Introducing two contrasted elements

Sometimes *copak* is used to introduce a pair of elements which are contrasted with each other. This construction may present several slightly different types of contrast (cf. Čermák et al. 2009, 139). The translation counterparts of such constructions were varied, only *never mind* occurred repeatedly (twice).[7]

Most frequently in our material, the former element (often previously mentioned by the addressee) is presented as unimportant (23), or of less importance (24) than the latter. The latter element may be omitted, as in (23).

(23) (a) Copak on!
(b) What did he matter?

(24) (a) Copak trapné, ale přišli bychom o Dvořákův violoncellový koncert!
(b) Never mind the embarrassment, think of the Dvořák's cello concerto we'd be missing!

The idiom *copak X, ale Y* sometimes corresponds to the English expression *That's all very well* (*but* . . .) (Čermák et al. 2009, 139)—however, this particular translation counterpart was not present in our material.

Copak in the contrasting constructions is evaluative—implying either negative (24) or positive evaluation (25), (26). As shown by the translation equivalent, in (25) the speaker implies that the publishing business will survive, but there is another more serious problem.

(25) (a) Copak nakladatelství, to vydrží.
(b) I'm not worried about her publishing business—that will hang together.

(26) (a) Teta ho rozmazluje a pořád o něm říká: „Copak náš Milouš!" Tvrdí o něm, že je neobvykle nadaný.
(b) My aunt spoils him and invariably comes out with the remark: "He really is something, our Bertie!" She claims that he has an unusual talent.

In practice, the above-mentioned functions of *copak* are combined. Let us illustrate a possible combination by (27), which shows the following functions of *copak*:

- Epistemic modality (certainty): The communicative function is an assertion of the opposite polarity. This assertion is even reinforced by the following sentence: *Naopak*/*On the contrary* . . .

7 *Kdepak* seems to have a similar function in examples such as *Kdepak Luis, ten by se nesek.—Luis! Are you kidding? He wouldn't mess up*, implying a contrast between Luis and someone else.

- Inference: The speaker is reacting to his partner's previous statement, based on which the speaker makes an inference paraphrasable as "I assume that you believe that everything that is not a mad chase after a final resolution is a bore."

- Appeal: The speaker is trying to persuade the addressee of the opposite.

(27) (a) „Když tě slyším," řekl nesměle profesor Avenarius, „bojím se, aby tvůj román nebyl nuda."
—"Copak všechno, co není bláznivý běh za konečným rozuzlením, je nuda? . . . Naopak . . ."
(b) "When I hear you," Professor Avenarius said uneasily, "I just hope that your novel won't turn out to be a bore."
—"Do you think that everything that is not a mad chase after a final resolution is a bore? . . . On the contrary . . ."

Notably, there is no one-to-one straightforward correspondence between a particular type of translation counterpart and a particular discourse function (cf. Petrová 2016).

3.3 *Kdepak*

Kdepak differs from the other particles not only in that it does not introduce questions but also in that it is frequently (62.8 per cent of clauses) used as a clause equivalent (cf. Komárek et al. 1986, 234), separated from the following clause by a comma (28) or constituting an independent sentence. It can also be appended at the end of a negative declarative clause (12.8 per cent of *kdepak*-clauses), further intensifying the preceding negation (29). Where *kdepak* itself is not separated from the rest of the sentence by punctuation (24.4%), it is used to front a clause element (30).

(28) (a) Kdepak, to bylo fakt ohromný.
(b) Not at all, that was quite fantastic.

(29) (a) Tady na východě se žádný záznamy nevedly jako u nás, kdepak.
(b) Here in the East they didn't keep records like we did, nowhere near it.

(30) (a) Ale kdepak já, běžel jsem, jako když mně hlavu zapálí, na Berounsko a víckrát jsem se na Kladencku neukázal.
(b) But as for me, forget it. I ran to the Beroun region as if they had set my head on fire and I never showed myself in Kladensko again.

In our material, emphatic initial signals of negative attitude were employed as translation counterparts of *kdepak* most frequently, corresponding to the prototypical initial

position of *kdepak*. They include a) introductory emphatic negative expressions (*oh no*, *not at all*, *of course not* . . . [28]); b) introductory negated clause element, syntactically not integrated in the clause (*not me*, [31])

(31) (a) O věcech Boga jsem se ani nezmínil, kdepak, já byl rád, že tam můžu ležet.
(b) I didn't even mention that stuff about Bog, not me, I was glad to be there.

The translation counterparts of *kdepak* also included idiomatic expressions *some hope*, *not a hope*, *what a hope*, which express "little confidence that expectations will be fulfilled" (*Collins English Dictionary*)—they carry epistemic modal meaning (32).

(32) (a) Kdepak, teď už bych nic neufoukal.
(b) Not a hope. Couldn't blow now.

All the initial signals point to the emphatic and emotionally expressive character of *kdepak*.

Often, the use of *kdepak* involves a pragmatic presupposition, i.e., the speaker merely supposes the addressee to hold a particular opinion. The counterparts show that *kdepak* introduces a statement which reacts to an immediately preceding utterance and (emphatically) denies either its explicit content (33), or a message inferred by the speaker on the basis of the content (in [34], the inferred statement might be "perhaps you have not forgotten your 'Our Father'"). *Kdepak* also emphasizes a previous negative statement, reinforcing it and even introducing the modal meaning of impossibility (cf. the emphatic negation in the counterparts *Oh, no*; *No*. *Not the* . . .; *certainly not*). These two aspects are both illustrated by the use of *kdepak* in (35).

(33) (a) „Kolik že vám je roků? Šedesát?"
—"Ale kdepak, pane profesore, před dvěma měsíci mi bylo pětaosmdesát."
(b) "How old did you say you were? Sixty?"
—"Oh, no, professor. I was seventy-five two months ago."

(34) (a) Jestlipak jste, vy syčáci, ještě nezapomněli otčenáš? Tak to zkusíme—Nu, já věděl, že to nepůjde. Kdepak otčenáš, takhle dvě porce masa a fazulový salát, napráskat se, lehnout si na kavalec, dloubat se v nose a nemyslit na pánaboha, nemám pravdu?
(b) Could it be, you bums, that you have forgotten your 'Our Father'? No? Then, let's try it. (Silence) Well, I knew you couldn't do it for me. No, not the 'Our Father'. Maybe two portions of meat and a bean salad . . . Stuff yourselves . . . Lay down on your bunks, pick your noses and never think of the Lord God! Am I not right?

(35) (a) A pak toho pána, co žertuje o plzeňském pivu, že musí být přímo od pípy, jinak to není ono . . . kdepak v láhvích v Mourek Innu.
(b) And then the gentleman who makes jokes about Pilsener beer, that it has to come straight from the tap or it's just not the same thing—certainly not from bottles in the Benes Inn.

Note that (34) also contains the particle *jestlipak*, which marks the question as dubitative (expressing doubt, Štícha et al. 2013, 763) and deliberative (the speaker is asking himself and the addressee/s at the same time, Zouharová 2008). The emotionally expressive tone of the utterance points to the particle's expressivity. Moreover, stylistically the particle is in line with other expressive expressions in the utterance, namely *syčáci*, *napráskat se.*

While the Czech *-pak* particles merely indicate negative epistemic modality (Komárek et al. 1986) and are employed in sentences whose communicative function is objection, reproach, disagreement, or expressing surprise (Grepl and Karlík 1998), in English the negative epistemic modal meaning tends to be expressed explicitly (e.g., by negative declarative clauses or introductory negative expressions, such as *not at all*).

4. Conclusions

4.1 Characteristics of the Postfix *-pak* and the *-pak* Particles

The postfix *-pak* has been shown to act as a polyfunctional indicator of communicative functions. Generally, its functions may be subsumed under the categories of expressing epistemic modality, appeal and contact. These functions are usually combined. Different shades of meaning are carried by different particles as the postfix *-pak* interacts with the different lexical bases.

- Epistemic modality

Jestlipak marks dubitative, deliberative meaning. *Copak* can signal the speaker's inference, or certainty of the opposite polarity. *Kdepak* marks the epistemic modal meaning of impossibility.

- Appeal

A *jestlipak*-question is neutral in the sense that it implies no expectations on the speaker's part, which is related to the capacity of *jestlipak* to make a question more tentative or polite. In contrast, *copak* questions present the speaker's expectation which is to be confirmed or refuted by the addressee.

Kdepak clauses, being declarative, have no function of appeal. The function of appeal seems to be an exclusive attribute of the interrogative *-pak* particles.

- Contact (+ politeness/tentativeness)

Jestlipak can be used to establish/maintain contact. This is again linked to its polite/tentative character.

The contact function of *copak* is most prominent in question tags and in non-rhetorical negative questions. It seems to be linked to the function of appeal (establishing contact and prompting the addressee to reply).

Although the function of establishing contact is by definition closely tied to questions, *kdepak* may be considered a means of establishing/maintaining contact as well, if understood widely as a "clausal particle"[8] (Komárek et al. 1986, 234). Štícha et al. take a similar view of *kdepak* as a "response interjection"[9] (2013, 534), resembling in words such as *yes* and *no* (in fact, our material shows that *kdepak* often functions as an emphatic equivalent of *no*).

In all these three areas, the emphatic and emotionally expressive character of the postfix *-pak* asserts itself.

Within our material the *-pak* particles occurred either in fiction dialogues or interior monologues. These contexts might suggest that the postfix is used more frequently in spoken discourse. However, the *-pak* particles are not very frequent in corpora of spoken Czech (*ORAL*). This may be caused by the small size of the available spoken corpora. Another possible interpretation is that the particles are more frequent in written texts, in which they may fulfil specific functions of simulating speech, perhaps compensating for the absence of prosody. This hypothesis will be subject to further research.

4.2 English vs. Czech in Terms of Their Third Syntactical Plans

Both English and Czech employ lexical as well as grammatical indicators of the speaker's stance. In both languages, these indicators tend to be sentence-initial: The Czech *-pak* particles are most frequently clause-initial. The English *I wonder* is prototypically (though not universally, as noted by Poldauf 1964) introductory. The same applies to the negators, idioms and other means which corresponded to *kdepak* in our material. As for the translations of *copak,* the vast majority is represented by questions. Whether polar or introduced by a *wh*-interrogative pronoun, questions are marked by a fixed word order in which the initial position is taken by a specific element, whereby the structure is immediately identified by the addressee as a question. Sentence-initial auxiliaries or *wh*-words could therefore be viewed in terms of their function as a specific type of opening markers.

The difference between the two languages lies in the frequency of each type of indicators. In English, the grammatical ones are more frequent where Czech uses lexical

8 "Větotvorné částice" in Czech (Komárek et al. 1986). Transl. DŠ.

9 "Odpověďová citoslovce" in Czech (Štícha et al. 2013, 534). Transl. DŠ.

means (particles). English often employs grammatical-lexical signals (specific syntactic structures)—e.g., negative polar questions, question tags. However, the English repertoire contains lexical means as well, mainly certain fixed constructions. These were represented in our material especially among the counterparts of *jestlipak* (*I wonder*) and *kdepak* (idioms, such as *not a hope*).

Works Cited

Aijmer, Karin. 2013. *Understanding Pragmatic Markers. A Variational Pragmatic Approach*. Edinburgh: Edinburgh University Press.

Aijmer, Karin, and Bengt Altenberg. 2002. "Zero Translations and Cross-linguistic Equivalence: Evidence from the English-Swedish Parallel Corpus." In *From the COLT's Mouth . . . and Others. Language Corpora Studies in Honour of Anne-Brita Stenström*, edited by Breivik, Leiv Egil, and Angela Hasselgren, 19–41. Amsterdam: GA: Rodopi.

Balhar, Jan, Pavel Jančák, et al. 2011. *Český jazykový atlas 5*. Prague: Academia.

Boye, Kasper. 2012. *Epistemic Meaning. A Crosslinguistic and Functional-Cognitive Study*. Berlin: De Gruyter Mouton.

Collins English Dictionary—Complete and Unabridged, *12th Edition.* 2014. S.v. "not a hope." Accessed September 10, 2016. http://www.thefreedictionary.com/not+a+hope.

Čermák, František, Jan Holub, Renata Blatná, and Marie Kopřivová. 2009. *Slovník české frazeologie a idiomatiky. 4: Výrazy větné.* Prague: LEDA.

Delahunty, Gerald P. 1995. "The Inferential Construction." *Pragmatics* 5 (3): 341–64. Amsterdam: John Benjamins Pub. Co. Accessed August 30, 2016. https://benjamins.com/#catalog/journals/prag.5.3.03del/fulltext.

Dokulil, Miloš, Karel Horálek, Jiřina Hůrková, Miloslava Knappová, and Jan Petr. 1986. *Mluvnice češtiny. 1, Fonetika, fonologie, morfonologie a morfemika, tvoření slov*. Prague: Academia.

Dušková, Libuše, Dagmar Knittlová, Jaroslav Peprník, Zdenka Strnadová, and Jarmila Tárnyiková. 2012. *Mluvnice současné angličtiny na pozadí češtiny*. Prague: Academia.

Filipec, Josef, and Libuše Kroupová. 2005. *Slovník spisovné češtiny pro školu a veřejnost*. Prague: Academia.

Grepl, Miroslav, and Petr Karlík. 1998. *Skladba češtiny*. Olomouc: Votobia.

Haspelmath, Martin. 2010. "Comparative Concepts and Descriptive Categories in Cross-linguistic Studies." *Language* 86 (3): 663–87.

Havránek, Bohuslav, Jaromír Bělič, Miloš Helcl, Alois Jedlička, et al. 1960. *Slovník spisovného jazyka českého I; A – M*. Prague: ČSAV.

Hirschová, Milada. 2013. *Pragmatika v češtině*. Prague: Karolinum.

Karlík, Petr, Marek Nekula, and Jana Pleskalová. 2002. *Encyklopedický slovník češtiny.* Prague: Nakladatelství Lidové noviny.

Karlík, Petr, Marek Nekula, and Zdenka Rusínová. 2000. *Příruční mluvnice češtiny.* Prague: Nakladatelství Lidové noviny.

Komárek, Miroslav, Jan Kořenský, Jan Petr, and Jarmila Veselková. 1986. *Mluvnice češtiny* 2. *Tvarosloví.* Prague: Academia.

Kranich, Svenja, and Volker Gast. 2015. "Explicitness of Epistemic Modal Marking: Recent Changes in British and American English." *Proceedings of Modality in English* IV. Accessed July 19, 2016. http://www.personal.uni-jena.de/~mu65qev/papdf/kranich_gast.pdf.

Malá, Markéta. 2013. "Translation Counterparts as Markers of Meaning." *Languages in Contrast* 13 (2): 170–92.

Petrová, Zuzana. 2016. *Czech "copak" and Its English Translation Equivalents in Parallel Texts.* MA thesis, Charles University in Prague.

Poldauf, Ivan. 1964. "The Third Syntactical Plan." *Travaux Linguistiques de Prague 1, L'école de Prague Aujourd'hui*: 241–55.

Quirk, Randolph, Stephen Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* London: Longman.

Šebestová, Denisa, and Markéta Malá. 2016. "Anglické překladové protějšky českých vět s částicemi *copak* a *jestlipak*." *Časopis pro moderní filologii* 98 (2): 228–40.

Štícha, František et al. 2013. *Akademická gramatika spisovné češtiny*. Prague: Academia.

Trávníček, František. 1951. *Mluvnice spisovné češtiny. Část II., Skladba.* Prague: Slovanské nakladatelství.

Wehmeier, Sally, ed. 2005. *Oxford Advanced Learner's Dictionary.* Oxford: Oxford University Press.

Zouharová, Marie. 2008. *Slovník frekventovaných syntaktických termínů*. Accessed July 24, 2016. http://clanky.rvp.cz/clanek/c/Z/1940/slovnik-frekventovanych-syntaktickych-terminu.html.

Corpora

Czech National Corpus—InterCorp (version 9). Institute of the Czech National Corpus. http://www.korpus.cz.

Czech National Corpus—ORAL2008. Institute of the Czech National Corpus. http://www.korpus.cz.

Parallel Corpus in Translation Studies: Analysis of Shifts in the Segmentation of Sentences in the Czech-English-French Part of the InterCorp Parallel Corpus

Olga Nádvorníková

Charles University, Prague, Czech Republic

olga.nadvornikova@ff.cuni.cz

Abstract: This study explores reasons for and consequences of shifts in the segmentation of sentences, i.e., the joining and splitting of sentences, in translations into English, Czech and French. On the basis of data from the core of the InterCorp parallel corpus, which contains mainly narrative texts, we explore two different explanations of these shifts: on the one hand, the hypothesis of information density, suggesting structural differences between languages regarding the preferred ways of information packaging; and, on the other hand, the theory of translation universals, assuming the influence of inherent features of the language of translation, such as simplification, explicitation and normalization.

Keywords: sentence splitting; translation universals; parallel corpus; information density

1. Introduction

In the process of translation, shifts in the segmentation of sentences are caused by two opposite operations: either the translator splits a single sentence of the source text into two or more sentences in the target text (1), or he/she joins two sentences of the source text together into one complex (or compound) sentence (2).

(1) (a) Il y avait toujours eu, sur la planète du petit prince, des fleurs très **simples, ornées** d'un seul rang de **pétales, et qui** ne tenaient point de **place, et qui** ne dérangeaient personne. (A. de Saint-Exupéry, *Le Petit Prince*, 1946/1999)

(b) On the little prince's planet the flowers had always been very **simple. They** had only one ring of **petals; they** took up no room at **all; they** were a trouble to nobody. (transl. K. Woods, 1943)

(c) Na jeho planetě rostly prosté květiny, ozdobené jedinou řadou okvětních **plátků. Nezabíraly** místo a nikoho nerušily. (transl. Z. Stavinohová, 1989)

(2) (a) Le consul n'acheva pas sa **phrase. En** ce moment, on frappait à la porte de son cabinet, et le garçon de bureau introduisit deux étrangers, dont l'un était précisément ce domestique qui s'était entretenu avec le détective. (J. Verne, *Le tour du monde en quatre-vingt jours*, 1873)

(b) The consul did not finish his **sentence, *for*** as he spoke a knock was heard at the door, and two strangers entered, one of whom was the servant whom Fix had met on the quay. (transl. not indicated)

(c) Konsul **nedořekl, *protože*** v tom okamžiku někdo zaklepal na dveře pracovny a kancelářský zřízenec uvedl dovnitř dva cizince, z nichž jeden byl právě onen sluha, který předtím hovořil s detektivem. (transl. J. Pospíšil, 1971)

In (1), both the Czech and the English translator split the source sentence and shift information from subordinate structures (relative clauses, in English also from the participle *ornée*) to coordinated structures, i.e., to a hierarchically higher level. In (2), on the contrary, both translators opt for joining the first short sentence with the second one; moreover, they both express overtly the implicit causal relationship between them (*for*/*protože*). Both operations (splitting and joining) may also occur at the same time: the translator may compensate for the splitting of the source sentence at one point by joining it at another one:

(3) (a) For a split second he hesitated, his hand on the window latch, wondering whether to slam it **shut. But** then the bizarre creature soared over one of the street lamps of Privet **Drive, and** Harry, realizing what it was, leapt aside. (J.K. Rowling, *Harry Potter and the Prisoner of Azkaban*, 1999)

(b) Pendant une fraction de seconde, il hésita, la main sur la poignée de la fenêtre, en se demandant s'il ne ferait pas mieux de la **refermer mais** au même moment, la créature passa au-dessus d'un réverbère de Privet **Drive. Harry** vit alors de quoi il s'agissait et fit aussitôt un pas de côté. (transl. J.-F. Ménard, 2000)

(c) Na zlomek vteřiny zaváhal s rukou na okenní klice a přemýšlel, jestli by raději neměl okno **přibouchnout, pak** ale onen bizarní tvor přelétl nad jednou z pouličních **svítilen. Harry** si uvědomil, co to je, a uskočil stranou. (transl. P. Medek, 2001)

In (3), we again observe a similarity in the translation strategies, although the source language is now English: both translators append the first sentence to the second one, using the conjunction *but* (*mais*/*ale*), and split the second sentence, after having shifted the subordinate non-finite clause (*realizing*) to an independent sentence (*vit*/*si uvědomil*).[1]

In parallel corpora, where the source text is aligned with the target text, the split/joined sentences result in so-called non-1:1 segments, i.e., aligned pairs of segments consisting of more than one segment on either side: 1:2 in (1), 2:1 in (2). In our research, we analyze also alignment pairs of equal number of segments, such as 2:2 in (3), because they are also the result of shifts in the segmentation of sentences.[2] In translation studies, various explanations for these shifts may be found; in what follows, we shall discuss two of them: information density, and translation universals.

1.1 Information Density

According to Fabricius-Hansen (1996 and 1999) or Solfjeld (1996), the splitting of sentences may be caused by structural differences between the source and the target languages; more specifically, by a difference in their (relative) information density. Fabricius-Hansen (1999, 203 and 1996, 558) argues that high information density languages (such as German) encode the discourse information in complex, hierarchical sentences, whereas low information density languages (e.g., Norwegian) prefer a more incremental, paratactic style.[3] According to this theory, sentence splitting is more likely in translations from a high information density language to a low information density language—and vice versa.[4] Solfjeld (1996,

1 The other *-ing* form (*wondering*) is maintained in the French translation at the non-finite level (gerund *en se demandant*), but rendered as a finite verb in Czech (*a přemýšlel—and he was wondering*); for an analysis of these shifts, see 1.1.

2 It is important to point out that the definition of an s-unit (sentence-unit) is not the same as the linguistic concept of a sentence. Alignment tools may also put the sentence boundary after a colon or a semi-colon in addition to the end-of-sentence punctuation such as period or exclamation mark (for more about automatic alignment, see, e.g., Rosen [2005]).

3 Fabricius-Hansen analyzes the information density not only at the sentence level, but also at the clausal level, e.g., also the cases of a non-clausal (phrasal) constituent, such as a gerund, turned into a subordinate clause (see [3]). Such a change does indeed reduce the information density, but it does not affect the sentence boundaries. Therefore, in what follows, we will not focus on this type of change.

4 We are aware of the fact that the level of the information density depends on the choice of the linguistic phenomenon on which it is based: in the number of words or syllables, French would score far higher than Czech; in the number of clauses, relevant for this paper, the result is the opposite.

567) defines this difference between languages in terms of "sententiality," i.e., the number of (finite) clauses per sentence.

English is seen by Fabricius-Hansen as less incremental than German, but not to the extent of Norwegian, especially due to the wider use of participial clauses and a more refined use of punctuation, including colons and semicolons (Fabricius-Hansen 1999, 204). Comparing French and English, Cosme (2006) shows that shifts from coordinate to subordinate constructions are more frequent in translations from English to French than in the opposite direction (18% and 5% respectively) and that "English inter-clausal *and* is used in a wider range of contexts than French *et*" (Cosme 2006, 94).[5] This result, suggesting a higher degree of incrementality in English in comparison with French, is corroborated by traditional contrastive literature comparing these two languages (Vinay and Darbelnet 1995 or Guillemin-Flescher 1981).

A detailed analysis of the degree of information density of Czech has not been carried out yet, but traditional contrastive analyses point out, e.g., the tendency of Czech to render non-finite verbal forms (gerunds or participial adjuncts) by (coordinate or subordinate) finite verbal forms (see example [3]—*wondering*/*en se demandant* vs. *a přemýšlel—and he was thinking*). These statements have been recently confirmed by several contrastive corpus researches, e.g., Čermák and Nádvorníková et al. (2015) show that adverbial *-ing* forms (*gérondif*, *gerundio*) in French, Italian, Portuguese and Spanish are rendered in Czech in approx. 50% of the occurrences by a coordinate finite verb form.[6] A similar result has been observed in translations from English: 60% of adverbial participial constructions have as equivalent in Czech a coordinate finite clause (Malá and Šaldová 2015).[7]

The tendencies in (relative) informational density observed in French, English and Czech are still only approximate, but we consider them sufficient to suggest the hypothesis that in comparison to English and French, Czech is a language of the lowest degree of information density. Therefore, sentence splitting will occur significantly more often in translations from English/French *into* Czech while sentence joining in the opposite direction.

5 A similar difference was observed when comparing English to another Romance language: Musacchio (2005, 93–94) notes the preference for juxtaposition and parataxis in English and for long complex sentences in Italian (she specifies, however, that the syntactic preferences of English influence the target Italian texts).

6 Analysis of Czech counterparts of French *gérondif* is available also in Nádvorníková (2010) (in English).

7 Martinková and Janebová (2017, 74) notice a loss of a sentence boundary in the English translation of a reported complex containing the Czech evidential particle *prý* (sentence joining).

1.2 Translation Universals

Bisiada (2016) argues that splitting of sentences is not due to the differences in the structural conventions of the source and target languages, i.e., the ratio of their respective information density, but that it is also the effect of a global translation strategy, i.e., independent of the source language. On a corpus of business and management articles translated from English into German, Bisiada shows that splitting of sentences occurs also in translations *into* German, i.e., a high information density language according to Fabricius-Hansen. Bisiada attributes this fact to a general tendency of translated language to explicitation.[8]

As defined by Baker (1996, 176–77, apud Olohan 2004, 91), *explicitation* is a translation universal based on "the tendency to spell things out in translation, including, in its simplest form, the practice of adding background information."[9] At the level of syntax, this tendency may involve the introduction of connectives explicitly specifying the relationship between clauses/sentences (see example [2]) or the transposition of non-finite forms to finite ones, which entails the specification of tense, modality and subject in the target text (Fabricius-Hansen 1999, 179). Some of the explicitation shifts are obviously due to the structural differences between the source and target languages, e.g., the transposition of the French gerund into a finite verb in Czech, already mentioned above. Similarly, the shifts from and-coordination to subordinate constructions in French, observed by Cosme (2006, see above), necessarily involve the explicitation of the inter-clausal semantic relationship. However, as shown, e.g., by Blum-Kulka (1986), explicitation may be observed in both translation directions (from French to English as well as from English to French).

The exaggeration of the features of the target language may, however, be a manifestation of another translation universal, called *normalization*: "the tendency to conform to patterns and practices that are typical of the target language" (Baker 1996, 176–77, apud Olohan 2004, 91). As shown by Vanderauwera (1985), Malmkjær (1997) and May (1997), changes in punctuation, often also involving shifts in segmentation of sentences, may make the target text more readable and closer to the usage of the target language, but at the same time affect the specific style of the source text.[10]

8 Bisiada points out that sentence splitting is influenced also by editorial guidelines (Bisiada 2016, 354).

9 As pointed out by Zanettin (2013, 25), some scholars consider the concept of *translation universals* controversial (see also Kruger [2002, 99] or Malmkjær and Windle [2012, 6]); however, we find the three specific features (normalization, simplification and explicitation) useful for the purposes of this work.

10 Fabricius-Hansen (1996, 561) notes that the acceptability of sentence splitting may be different in narrative and in argumentative texts. This factor of text genre is also taken into account by Cosme (2006, 94), who concludes that the inter-clausal *and* is more frequent in English than in French only in fiction: in journalese no such difference has been observed.

Improvements of the "readability" of the target text involve the third translation universal which may shed some light on the consequences of sentence segmentation shifts in translation, namely *simplification*—"the idea that translators subconsciously simplify the language or message or both" (Baker 1996, 176–77, apud Olohan 2004, 91). In sentence organization, this may mean that long, complex sentences are split into shorter, simpler ones.

Based on these observations, our second hypothesis is that the segmentation shifts will be motivated not only by structural differences between languages, but also by the inherent features of translated language (normalization, explicitation and simplification). Hence, the splitting or joining of sentences may also occur independently of the direction of the translation and the degree of information density of the source and target languages (e.g., Czech and French).

2. Analysis of the Czech-French-English Parallel Data

Each of the above hypotheses needs a different type of linguistic material to be tested on: the theory of information density requires extensive quantitative data, whereas for the analysis in terms of translation universals a refined qualitative approach is more suitable. These two approaches do not exclude each other. We will first analyze a large number of non-1:1 segments (2.1) and then—manually—a sample of the parallel segments. On this sample we can observe not only the types of shifts in segmentation, but also the reasons for and more importantly, the consequences of these shifts.

Both parts of the analysis are based on the same corpus: the core of the Czech-French-English part of the parallel corpus InterCorp (www.korpus.cz/intercorp, https://kontext.korpus.cz). Unlike its other subsections (called "collections": Acquis communautaire, Europarl, Subtitles, etc.), its "core" section contains mainly fiction. Moreover, it has two advantages, important for the analysis of the sentence segmentation shifts:

(i) a higher quality of alignment, due to the proofreading step in the pre-processing of core texts, supported by the InterText parallel text editor (Vondřička 2014);
(ii) a possibility of reliably identifying the direction of translation.

As we were interested only in non-1:1 segments, we did not use KonText, the standard corpus interface KonText of InterCorp, but rather a list of such segments, provided by the Institute of the Czech National Corpus.[11] These data are rich, but limited by two aspects of InterCorp:

(i) Czech is the pivot language of the parallel corpus, which means that all the texts are aligned first to the Czech version, and *through* this alignment, to other languages. Consequently, the shifts between English and French cannot be observed directly and Figure 1 shows only the combinations of non-1:1 segments including Czech.

11 I owe thanks to Pavel Procházka and Alexandr Rosen for their help and cooperation.

(ii) The intersection between Czech, French and English in the corpus InterCorp is so far quite limited, especially in the FR-cs-(en) part, i.e., translations from French including not only Czech, but also English:

EN-cs-(fr)	CS-en-(fr)	FR-cs-(en)
Carroll *Alice in Wonderland*	Hašek *Good Soldier Švejk*	Foucault *The Order of Things*
Kipling *The Jungle Book—Mowgli*	Havel *Disturbing the Peace*	Verne *Around the World in Eighty Days*
Kipling *The Jungle Book—other*	Havel *Largo Desolato*	Saint-Exupéry *Little Prince*
Rowling *H.P. and the Philosopher's Stone*	Hůlová *All This Belongs To Me*	
Rowling *H.P. and the Prisoner of Azkaban*	Jirotka *Saturnin*	
Tolkien *Lord of the Rings 1*	Klíma *Love and Garbage*	
Wells *Time Machine*	Kundera *Immortality*	
Wells *War of the Worlds*	Kundera *Unbearable Lightness of Being*	
	Kundera *Joke*	
	Otčenášek *Juliet and Darkness*	
	Topol *The Devil's Workshop*	
	Viewegh *Bringing Up Girls in Bohemia*	

Table 1. The intersection between the English, Czech and French parts of the core of the parallel corpus InterCorp

Table 1 shows that the Czech-French-English subcorpus is unbalanced not only in the representation of the different source languages (EN, CS, FR), but also in its textual types (there are only two non-fiction texts—*The Order of Things* by M. Foucault and *Disturbing the Peace* by V. Havel—and a single play—*Largo Desolato* by V. Havel)) or in the publication year of the texts and translations (mainly in the EN-subcorpus, where the texts published in the 19th century prevail). Last but not least, we have to mention the potential influence of the authors' (and the translators') idiolects, cf. the three texts by Milan Kundera in the CS-subcorpus. This is why we specify not only the author and

the title of the source text, but (if possible) also the translator and the year of publication for each example mentioned in this paper.

These limitations mainly affect the manual analysis of non-1:1 segments, involving all three languages at once (see 2.2); on the other hand the quantitative analysis is based only on the pairs of languages, which allows us to rely on a much larger amount of data.

2.1 Information Density

As mentioned in 1.2, the theory of information density is based on the assumption that structural differences between languages influence the way they encode information in sentences. We indicated that Czech has a lower information density than English and French. On these bases, we can expect more splitting of sentences in translations into Czech and more joining in the opposite direction of translation (into French/English).

Language pair	splitting	joining	splitting-joining	non1:1 total	total align-ments	number of texts
en-cs	44,302	26,550	1,855	72,707	754,181	178
% of non1:1	60.9%	36.5%	2.6%			
% of alignments	5.9%	3.5%	0.2%	9.6%		
cs-en	12,029	16,834	1,460	30,323	147,958	25
% of non1:1	39.7%	55.5%	4.8%			
% of alignments	81%	11.4%	1.0%	20.5%		
fr-cs	9,436	6,436	544	16,416	223,729	60
% of non1:1	57.5%	39.2%	3.3%			
% of alignments	4.2%	2.9%	0.2%	7.3%		
cs-fr	5,224	9,598	433	15,255	91,880	20
% of non1:1	34.2%	62.9%	2.8%			
% of alignments	5.7%	10.4%	0.5%	16.6%		

Table 2. Number of non-1:1 segments in translations from/into Czech in the core of InterCorp

Table 2 shows the numbers of non-1:1 segments in the four translation subcorpora (translations from English into Czech: EN-CS, from Czech into English: CS-EN, etc.). If the number of units marked as <s> (sentence) in the source language segment is lower than in the target language segment (e.g., 2:1 in [1]), then splitting occurs; the opposite case is the occurrence of joining (see [2]); if the number of sentences is equal (e.g., 2:2, see [3]), it is the case of "splitting-joining," i.e., usually a compensation.

Column 5 of Table 2 shows that for the translations from English into Czech, 72,707 non-1:1 segments were analyzed, which represents 10% of all the alignments of this English-Czech subcorpus (column 4). We can also observe that in this English-Czech language pair, splitting occurred much more frequently than joining: 60.9% of splitting against 36.5% of joining in all non-1:1 segments.

If we consider the other language pairs, we can conclude that the information density hypothesis is more or less confirmed: splitting is always more frequent in translations *into* Czech and joining occurs more often in the opposite direction.[12]

However, these results have to be specified in several other aspects:

First, the comparison with Czech does not confirm a higher information density in French than in English: if this were the case, the proportions of shifts in the French-Czech translation pair would be much more important than those in the English-Czech translation pair. We hope that in the future, it will be possible to extract non-1:1 segments between French and English directly and to analyze the shifts; at the moment, we cannot present a more robust conclusion.

Secondly, other factors than just information density, such as standard practices of translation, clearly influence the results. For this reason, the proportion of non-1:1 segments (joining-splitting together) is more significant in translations *from* Czech than *into* Czech. This may suggest that Czech translators are more respectful to the style (or segmentation) of the source text than English and French translators. However, these results have to be verified on a larger corpus, because the corpora of translations *from* Czech are much smaller than the corpora of translations *into* Czech. Such a corpus is thus more sensitive to the specifics of a single text.[13]

The third point is the most intriguing: even in the language pairs where we would expect more splitting (in translations into Czech), there are a considerable number of the cases of joining: they represent more than 30% of all changes. This means that to explain this tendency, we have to look not only at these rough data, but at the specific shifts in detail.

2.2 Translation Universals

It is obvious that shifts in the segmentation of sentences and changes in punctuation between three languages must be multifarious and complex. Moreover, as pointed out also by Fabricius-Hansen (1996, 561), shifts in translation should not be analyzed only for separate

12 According to the chi-squared test, all the differences are significant at $p<.001$; the effect size calculation shows that the proportions in the analyzed subcorpora differ by about 21 pp (with limit values 20.6% and 21.9%).

13 These differences between Czech on the one hand, and English and French on the other, may also be caused by another factor, discussed in detail by Vanderauwera (1985): the status of "minority" literature.

sentences, but rather in the context of the whole text. Given the limited space for the presentation of this research, we will focus on the shifts concerning the tendencies against the splitting/joining processes presented above.[14] Additionally, we will try to illustrate them by examples representing not just a single case in the translation, but apparently a global translation strategy of the translator.

As already mentioned, the manual analysis of the triplets of non-1:1 segments was undertaken on a much smaller corpus than the one used for the analysis of the binary non-1:1 segments (see Table 2), as the intersections of the three languages in the corpus are also limited (mainly in the translations from French into Czech and English, see Table 1 and Table 3). Furthermore, given the impossibility of direct alignment between French and English (Czech being the pivot language), the third segment of the triplet is sometimes missing.[15] Despite these limitations, from each combination of the aligned texts, we extracted a sample of 500 complete non-1:1 segments to be used in the analysis of the shifts in segmentation and punctuation.[16]

Combination of languages	Number of texts	Total number of non-1:1 segments	Sample non-1:1
EN-cs-(fr)	8	3,446	500
CS-en-(fr)	12	11,471	500
FR-cs-(en)	3	960	500
CS-fr-(en)	12	10,426	500
Total	23	15,877	2,000

Table 3. Czech-French-English subcorpora in InterCorp (the source text is indicated by capital letters) with samples of manually analyzed segments

In all the analyzed directions of translation, the most frequent change from the original involved—not surprisingly—the most frequent punctuation mark: the comma. Comma is the source of change in about 25–30% of all the shifts. Most frequently, comma changes into an end-of-sentence period or semi-colon (in all the translation directions, see, e.g., [1] and [3]), which means increased segmentation. Ranking second among the most frequent sources of changes, we either find the end-of-sentence period, the second most frequent punctuation mark (if the source language is Czech), or the semi-colon (if the source language is French or English).

14 For a thorough analysis of all the types of shifts in segmentation and punctuation between Czech and French, see Nádvorníková and Šotolová (forthcoming).

15 These combinations of segments were excluded from the analysis.

16 The non-1:1 segments in CS-fr-en and CS-en-fr overlap only partially; therefore, we retained them both for the analysis.

Some changes in segmentation are to a large extent only typographical, e.g., a very frequent change at the border of narrative levels—between an introductory clause (*proposition incise*) and the second part of the direct speech:[17]

(4) (a) "No," I **cried,** "**that's** impossible! . . ." (Wells, *The War of the Worlds*, 1898)
(b) – Non, m'écriai-**je, c'est** impossible. (H. D. Davray, 1917)
(c) „Ne! Tohle ne!" zaprotestoval **jsem.** „**To** není možné! . . ." (V. Svoboda)

Other changes in punctuation in the analyzed samples are motivated by structural differences between languages (see the *-ing* forms in [3]) and thus confirm the hypothesis of information density. Stylistic conventions are close to these structural differences, e.g., the tendency to avoid repetition or the use of semi-colon. The former tendency may be illustrated by (1), where the Czech translator had to avoid the repetition of the relative pronoun *qui*/*který*, which is considered stylistically clumsy in Czech. In the latter case, the difference is based on frequency: in English and in French, the semi-colon is much more frequent than in Czech (see [1b] and [6b]).[18]

Nevertheless, many changes cannot be explained by any of these systemic or stylistic differences. This is especially the case of **splitting** in translations from Czech into French/English, which seems to be a global translation strategy mainly in texts where the narrator wants to emulate spoken language, using long sentences based on paratactic style:

(5) (a) Ale táta přece nemohl vědět, že babička umře, to jsem chápala už tehdy, a tak jsem se táty zastávala, a že nic špatnýho neudělal, si myslim i teď. (P. Hůlová, *Paměť mojí babičce*, 2002)

17 Likewise, only minor modification may be observed between *"No! no!" cried Frodo*, and *„Ne! Ne!" vykřikl Frodo.* (Tolkien, *The Fellowship of the Ring*), although in Czech the automatic segmenter identified two segments (according to the rule of an end-of-sentence punctuation mark followed by a capital letter). Likewise, the sequence after the colon in (6b) is treated as a separate s-unit, as it starts with a capital letter. This type of changes represents about 5% of all the non-1:1 segments analyzed.

18 In French narrative texts, the semi-colon is twice as frequent as in Czech; in non-fiction, it is nearly four times as frequent (data based on the FRANTEXT corpus for French and the SYN corpus for Czech). Nevertheless, data from Jerome, the Czech comparable translation corpus (Chlumská 2013) suggest that Czech translations may be influenced by the source texts, as the relative frequency of the semi-colon in translations is seven times higher than in the non-translated texts, both in fictional and non-fictional texts (see Nádvorníková and Šotolová, forthcoming; also Šotolová 2013).

(b) There was no way for Papa to know that Grandma was going to die, **though. I** realized that even **then. So** I told him he didn't do anything wrong, and I still think that today. (transl. A. Zucker, 2009)

(c) Mais papa ne pouvait savoir que grand-mère allait **mourir. Je** le comprenais déjà à l'époque, je prenais sa défense, et je continue de penser qu'il n'a rien fait de mal. (transl. H. Rihova-Allendes; A. Maréchal, 2005)

The translated sentences in French and in English are more logically structured, more "readable" and closer to the typical sentence of the target languages, but the specific style of the original, stressed in Czech also by morphological traits of colloquial language (*špatnýho*, *myslim*), is lost. We can see the same global translation strategy in the French and English translations of Jáchym Topol, another contemporary Czech author. May (1997) observed a similar tendency to a more logical, structured text in Russian and French translations of Faulkner's and Woolf's novels, explaining it as the effect of *normalization*. We agree with this explanation of splitting of (long) sentences; we want to show, however, that the effect of the same translation universal may also shed some light on the *joining* of sentences.

The normalization-based **joining** of sentences appears easily and very frequently when the final stop in the source text is followed by the conjunction *Et*/*And*/*A* or *Mais*/*But*/*Ale* (see [3b, c]). Bisiada (2016, 374) argues that "splitting sentences at the point of the conjunction may be the least intrusive way of introducing full stops," but that "[t]his may introduce false emphasis on the logically subordinated propositions." In the case of *joining* at this point of the sentence, the effect is inverse: the stress put by the author on the separate sentence is "diluted" in the resulting complex sentence.[19] The same effect may be observed in the case of joining sentences without the support of the conjunction in the source text; see (2) and the following example:

(6) (a) Voici mon **secret. Il** est très simple: on ne voit bien qu'avec le cœur. (A. de Saint-Exupéry, *Le Petit Prince*, 1946/1999)

(b) "And now here is my **secret, a** very simple secret: It is only with the heart that one can see rightly; . . . " (transl. K. Woods, 1943)

19 In a non-fictional text (FR—source text): Comme au XVI^{e} siècle, ressemblance et signe s'appellent **fatalement. Mais** sur un mode nouveau. (M. Foucault, *The Order of Things*, 1990). CS—Stejně jako v 16. století na sebe podobnost a znak osudově **odkazují, ale** ovšem jiným způsobem. (transl. J. Rubáš, 2007). EN—As in the sixteenth century, resemblance and sign respond inevitably to one **another, but** in a new way. (transl. A. M. Sheridan Smith, 2001)

(c) „Tady je to mé **tajemství, úplně** prostinké: správně vidíme jen srdcem; . . . “ (transl. Z. Stavinohová, 1989)

Intentionally short, segmented sentences in the original Saint-Exupéry's text are rendered in *both* translations (Czech and English) by a more hierarchical structure.[20] Hence, the resulting effect of this joining of sentences is the same as in the case of splitting observed above: a "normalized" text.

3. Conclusions

This article has argued that changes in the segmentation of sentences in translation may be triggered by two (sometimes conflicting) forces: on the one hand, structural differences between the source and target languages, causing differences in their respective information density; and, on the other hand, tendencies inherent to the process of translation, so-called translation universals, independent of the source language.

As for the information density hypothesis, we assumed a lower degree of information density in Czech, in comparison with French and English. A large set of non-1:1 segments extracted from the Czech-English-French part of the InterCorp parallel corpus (mainly fictional texts) confirmed this general difference (significantly more splitting in translations from English/French into Czech and more joining in the opposite direction). We have to specify, however, that other factors may influence these results, especially the composition of the analyzed subcorpora, the standard translation practice or the minority status of the translated literature. In addition, these results, based especially on narrative texts, should be verified on non-fiction. Future research should also concentrate on a more thorough analysis of the direct differences between French and English and on comparison with other textual types.

We completed this large quantitative research by a manual analysis of Czech-English-French non-1:1 segments. We showed that the splitting/joining tendencies which are contrary to the hypothesis of information density may be explained by the influence of translation universals, especially normalization: intentionally short or long sentences are joined or split to a "normal" length. This change, sometimes mixed with simplification or explicitation, makes the target text smoother, more logical and "readable," but if it becomes a global translation strategy, it wipes away the specific style of the source text, so important in fiction.

20 Joining of sentences is often done also by using the conjunction *and*/*et*/*a*, resulting in a coordinate structure; this type of shift is frequent (in both translations), e.g., in the J.K. Rowling text: EN—Fortunately, Dumbledore arrived moments **later. The** babble died away. (J.K. Rowling—*Harry Potter and the Philosopher's Stone*, 1997). CS—Okamžik nato dorazil naštěstí **Brumbál a šuškání** v síni ustalo. (transl. V. Medek, 2000). FR—Heureusement, Dumbledore arriva à son **tour et la** rumeur des conversations s'évanouit. (transl. J.-F. Ménard, 2005).

Hence, the difficult mission of the translator is to constantly search for balance between the structural and stylistic conventions of the target language, and the specificities of the style of the source text.

Funding Acknowledgement

This study was supported by the Charles University project Progres 4, Language in the shiftings of time, space and culture.

Works Cited

Baker, Mona. 1996. "Corpus-Based Translation Studies: The Challenges That Lie Ahead." In *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*, edited by H. Somers, 175–86. Amsterdam: John Benjamins.

Bisiada, Mario. 2016. "'Lösen Sie Schachtelsätze möglichst auf': The Impact of Editorial Guidelines on Sentence Splitting in German Business Article Translations." *Applied Linguistics* 37 (3): 354–76.

Blum-Kulka, Shoshana. 1986. "Shifts of Cohesion and Coherence in Translation." In *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, edited by J. House and S. Blum-Kulka, 17–35. Tübingen: Narr.

Čermák, František, and Alexandr Rosen. 2012. "The Case of InterCorp, a Multilingual Parallel Corpus." *International Journal of Corpus Linguistics* 13 (3): 411–27.

Čermák, Petr, and Olga Nádvorníková et al. 2015. *Románské jazyky ve světle paralelních korpusů*. Praha: Karolinum.

Cosme, Christelle. 2006. "Clause Combining across Languages. A Corpus-Based Study of English-French Translation Shifts." *Languages in Contrast* 6 (1): 71–108.

Fabricius-Hansen, Cathrine. 1996. "Informational Density: A Problem for Translation and Translation Theory." *Linguistics* 34: 521–65.

Fabricius-Hansen, Cathrine. 1999. "Information Packaging and Translation: Aspects of Translational Sentence Splitting (German–English/Norwegian)." In *Sprachspezifische Aspekte der Informationsverteilung*, edited by M. Doherty, 175–214. Berlin: Akademie Verlag.

Guillemin-Flescher, Jacqueline. 1981. *Syntaxe comparée du français et de l'anglais: problèmes de traduction* (Réimp.). Paris: OPHRYS.

Kruger, A. 2002. "Corpus-Based Translation Research: Its Development and Implications for General, Literary and Bible Translation." *Acta Theologica Supplementum* 2: 70–106.

Malá, Markéta, and Pavlína Šaldová. 2015. "English Non-Finite Participial Clauses as Seen through Their Czech Counterparts." *Nordic Journal of English Studies* 14: 232–57.

Malmkjaer, Kirsten. 1997. "Punctuation in Hans Christian Andersen's Stories and in Their Translations into English." In *Nonverbal Communication and Translation. New Perspectives and Challenges in Literature, Interpretation and the Media*, edited by F. Poyatos, 151–63. Amsterdam: John Benjamins.

Malmkjær, Kirsten, and Kevin Windle, eds. 2012. *The Oxford Handbook of Translation Studies.* Oxford: Oxford University Press.

Martinková, Michaela, and Markéta Janebová. 2017. "What English Translation Equivalents Can Reveal about the Czech 'Modal' Particle *prý*: A Cross-Register Study." *In Contrastive Analysis of Discourse-Pragmatic Aspects of Linguistic Genres. Yearbook of Corpus Linguistics and Pragmatics* 5, edited by Karin Aijmer and Diana Lewis, 63–90. Cham: Springer.

May, Rachel. 1997. "Sensible Elocution: How Translation Works in & upon Punctuation." *The Translator* 3 (1): 1–20.

Musacchio, Maria Teresa. 2005. "The Influence of English on Italian: The Case of Translations of Economics Articles." In *In and Out of English: For Better, for Worse?*, edited by Gunilla M. Anderman and Margaret Rogers, 71–96. Clevendon, GB: Multilingual Matters.

Nádvorníková, Olga. 2010. "The French Gérondif and Its Czech Equivalents." In *InterCorp: Exploring a Multilingual Corpus*, edited by František Čermák, Patrick Corness, and Aleš Klégr, 83–96. Prague: Nakladatelství Lidové noviny/Ústav Českého národního korpusu.

Nádvorníková, Olga, and Jovanka Šotolová. Forthcoming. "Změny v segmentaci na věty v překladových textech: analýza dat z francouzsko-českého paralelního korpusu." [Changes of Segmentation in Phrases in Translation]. In *Jazykové paralely*, edited by Anna Čermáková, Lucie Chlumská, and Markéta Malá. Prague: ÚČNK/NLN.

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies.* London: Routledge.

Rosen, Alexandr. 2005. "In Search of the Best Method for Sentence Alignment in Parallel Texts." In *Computer Treatment of Slavic and East European Languages: Third International Seminar, Bratislava 10–12 November 2005*, edited by R. Garabík, 174–85. Bratislava: VEDA.

Solfjeld, Kåre. 1996. "Sententiality and Translation Strategies German-Norwegian." *Linguistics* 34: 567–90.

Šotolová, Jovanka. 2013. "Sur le point-virgule et autres détails éphémeres." *Études Romanes de Brno* 34 (1): 28–40.

Vanderauwera, Ria. 1985. *Dutch Novels Translated into English: The Transformation of a "Minority" Literature.* Amsterdam: Rodopi.

Vinay, Jean-Paul, and Jean Darbelnet. 1995. *Comparative Stylistics of French and English: A Methodology for Translation.* Amsterdam: Benjamins.

Vondřička, Pavel. 2014. "Aligning Parallel Texts with InterText." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 1875–79. European Language Resources Association (ELRA).

Zanettin, Federico. 2013. "Corpus Methods for Descriptive Translation Studies." In *Procedia—Social and Behavioral Sciences* 95: 20–32.

Corpora

Chlumská, Lucie. 2013. *Korpus Jerome – srovnatelný translatologický korpus překladové a nepřekladové češtiny.* Prague: Institute of the Czech National Corpus, Charles University. www.korpus.cz.

Nádvorníková, Olga, and Martin Vavřín. 2015. *Korpus InterCorp—French, version 8*. Prague: Institute of the Czech National Corpus, Charles University. http://www.korpus.cz.

Pragmatics of "Saying" Routines in Police Interviews

Magdalena Szczyrbak

Jagiellonian University, Kraków, Poland

magdalena.szczyrbak@uj.edu.pl

Abstract: The verb *say*, it can be argued, plays a role in formulating legal narratives and, consequently, in constructing evidence in legal-lay communication. In light of the above, the current study examines the patterns of use involving *say* in police interviews carried out in a homicide investigation. The aim of the analysis is threefold: (1) to determine the frequencies of selected *say* forms; (2) to explore the speaker-form correlation, i.e. to establish how legal professionals and laypersons deploy *say* in interaction and (3) to compare selected "saying" routines in police interviews and in trial data.

Keywords: communication verbs; institutional talk; police interview; stance

1. Introduction

The subject of power asymmetry has figured, in one way or another, in numerous scholarly investigations. The role of various linguistic resources in controlling discourse, including questioning strategies, has been studied in detail, too. Pragmatic uses of *say*, however, have attracted considerably less attention. To fill this gap, this paper revisits the role of *say* and seeks to relate its use to the institutional practices of police interviewers. In doing so, it first explains the communicative context of the police interview and then discusses some of the interactional routines pursued by questioners, stressing the role that *say* plays in constructing authority and shaping the epistemic asymmetry between institutional and non-institutional speakers.

2. Police Questioning Routines

Like other institutional practices, police questioning too complies with the externally imposed rules of procedure and norms of interaction. Typically, police interviews follow

the four-stage format including: opening, recount, questioning and closure. What is more, in this hybrid genre, two discoursal planes are clearly distinguishable, that is "primary reality," referring to the circumstances of the interview itself, and "secondary reality," referring to the portrayal of investigated actions and events, which must necessarily fit relevant legal theories and concepts (Gibbons 2005, 142–50). The point also needs to be raised that apart from the real information gathering, police interviews seek "to obtain confirmation of a particular version of events that the questioner has in mind" (Gibbons 2005, 95). Thus, non-coercive elicitation of information ("the pursuit of truth") competes with coercive interrogation aimed at obtaining a confession ("the pursuit of proof") (Baldwin [1993] quoted in Gibbons [2005, 96]). This, in turn, affects the design of the questioning agenda and the choice of the discursive strategies employed to achieve the desired effect.

Apart from the less or more coercive question types,[1] quotation and reformulation seem to take center stage among the favored interview tactics (see, e.g., Gaines 2016; Szczyrbak, forthcoming). In fact, it would be difficult not to agree with Johnson (2014, 546), who observes that in legal interaction quotation is a marked source introducing "a context-shift that is controlled" or, more broadly, that "selectivity involving quotation is a potent linguistic resource" (Johnson 2014, 526). Reformulation does significant pragmatic work as well, since it enables the questioner to frame the respondent's replies and to seek confirmation of the preferred version of events. Naturally, both quotation and reformulation are linked to the use of the most common reporting verb, i.e. *say*, which is intentionally used by questioners and which, in the courtroom context, allows "the lawyer to ventriloquize and animate the voice of the other side making it present" (Johnson 2014, 545), at the same time "contrasting what was said with what is said now and here to construct truth and lies and to construct evidence within a defense case" (Johnson 2014, 645). In police interviews, similarly, *say* is employed by questioners to build authority through quoting and then challenging the respondent's earlier words, which agrees with the claim that police officers have not only physical, but also linguistic power over interviewees, as they are in a position "to direct a witness's story, to choose what aspects to focus on in a summary of their story, and to ask questions of suspects in a coercive way" (Eades 2010, 180). These practices, involving selected patterns with *say*, will be examined in Section 3.

3. "Saying" Routines in Police Interviews

Building on earlier research into pragmatic uses of *say* in legal communication (Taylor 2009; Johnson 2014; Szczyrbak 2016, forthcoming), the current analysis investigates the interactional use of *say* in police interviews. In particular, it seeks to relate speaker- and hearer-oriented forms of *say* to two participant roles: that of questioner and that of respondent. In doing so, the study focuses on patterns related to "saying what is said"

1 For a typology of the least and most coercive leading questions, see, e.g., Berk-Seligson (1999).

(Johnson 2014) or "talking about the talk" (Taylor 2009, 220), which inevitably involves evaluation of the accounts provided by the respondents as well as the positioning of the speakers themselves.

3.1 Data and Method

The material used for the analysis comprises 30 randomly chosen police interviews[2] (totaling 271,544 words) from a homicide investigation which was launched after the shooting of an unarmed African American by the police officer Darren Wilson in a suburb of St. Louis, Missouri, in 2014. As is usually the case, the content of the interviews was intended for a non-participatory audience (here: the grand jury), and not just the participants present in the interview room (i.e. investigating officers, lay respondents and legal representatives).[3] Somewhat unusually, however, in order to ensure transparency, anonymized interview transcripts were made publicly available,[4] thanks to which they could easily be accessed for research purposes (St. Louis Post-Dispatch 2014).

In the study, the CADS approach (Partington et al. 2013) was adopted, with the emphasis being placed on the qualitative interpretation of the data, rather than the quantitative findings alone. And even though the corpus itself can rightly be described as small by today's standards, it is believed that it contains a sufficient number of occurrences to suggest certain trends in the use of *say*. As will become evident in the ensuing sections, the CADS perspective is combined with the assumptions underlying interactional linguistics. Along the same lines, the use of *say* is linked to the notion of stance conceptualized as a sequentially organized public act involving the positioning of subjects and the evaluation of objects (du Bois 2007).

Regarding the corpus search itself, it started, using WordSmith Tools (Scott 2012, version 6), with a general query for the node word *say*, out of which only *I*- and *you*-oriented patterns were distinguished. These, in turn, included strings with present, past and modalized forms of *say*.[5] The six most frequent items in the corpus (with 40 occurrences established as a cut-off point) were then selected for a manual reading focused on the distribution of individual *say* patterns as well as the pragmatic functions associated with them. To facilitate comparison with other datasets, the raw scores were normalized to show frequencies per million words.

2 For the purpose of this study, the term "police interviews" embraces interviews conducted by the police and the FBI.

3 Cf. Haworth's (2009, 115–21) description of the audience design in police interviews, including addressees, auditors, overhearers and eavesdroppers.

4 This happened despite the fact that the federal investigation cleared Wilson of civil rights violations and, consequently, the grand jury decided against indictment.

5 The query was part of a larger research project into the use of *say* in different legal genres (cf. Szczyrbak, forthcoming).

3.2 General Description of Patterns with *Say*

As the corpus search revealed, the speakers used a variety of *I*- and *you*-oriented forms of *say*, some of which were more prominent than others. It was noted, for instance, that hearer-oriented forms (e.g. *and you say*, *when you say*, *you're saying*) significantly outnumbered speaker-oriented ones (e.g. *and I say*, *when I say*, *what I'm saying is*). At the same time, speaker-oriented *like I said* alone was the most frequent form analyzed (Table 1). By contrast, items such as *I have to say*, *I must say* or *I should say* were attested only by single occurrences.

POLICE INTERVIEWS		
***Say* patterns**	**Raw count**	**Normed score (per million words)**
like I said	84	309.34
and you say	55	202.55
and you said	54	198.86
when you say	50	184.13
you're saying	45	165.72
you know what I'm saying	40	147.31

Table 1. Most frequent *say* forms in police interview data

Several distinct patterns were observed regarding distribution as well (Table 2). Predictably, it was revealed that *you*-oriented forms were clearly favored by the questioners, except for *you know what I'm saying* which was used repeatedly by the respondents and which, despite the presence of the pronoun *I*, can be regarded as a hearer-oriented form, too. In addition, the only frequent *I*-oriented form of *say*, i.e. *like I said*, was used almost exclusively by the respondents.

***Say* patterns**	**POLICE INTERVIEW PARTICIPANTS**		
	Questioners	**Respondents**	**Total**
	Raw count	**Raw count**	**Raw count**
like I said	4	**80**	84
and you say	**55**	0	55
and you said	**54**	0	54
when you say	**48**	2	50
you're saying	**45**	0	45
you know what I'm saying	0	**40**	40

Table 2. Distribution of *say* per participant

3.3 Pragmatics of "Saying" Routines

As reported in an earlier study focusing on trial data, *say* can be linked to the stancetaking functions of "shifting standpoints," "challenging standpoints," "reality reconstruction" and "standpoint continuity" (Szczyrbak 2016). The current research, in the same fashion, investigates interview data, with the aim of revealing the most salient pragmatic functions of *say* in police questioning, which, it can be argued, "probes" rather than "cross-examines" and "suggests" rather than "demands" (cf. Shuy 1998, 12–13). Consequently, as will be shown, in the interview format, the role of *say* in reality reconstruction and challenging standpoints appears to be more prominent than its usefulness for shifting standpoints or marking standpoint continuity.

Moving on to the specifics, the item *like I said* was the only prominent speaker-oriented form of *say* which resurfaced from the data.[6] Found chiefly in the interviewees' turns, *like I said*, it can be speculated, served not only to organize the narrative, but also to emphasize the speaker's perspective.[7] In (1a), for instance, the respondent inserts *like I said* between the repeated words "he was ducked in between" which may be interpreted either as a filler device or an emphasis marker, depending on the actual stress falling on these words (which, however, is not annotated in the transcript data). Similarly in (1b), when saying "cause like I said", the speaker seems to underline the accuracy of his earlier words, insisting that his account is based on what he himself saw.[8]

(1) (a) A: He was right behind the officer's—
B: His friend was?
A: —yes he was ducked in between, **like I said**, he was ducked in between the car that's behind the police vehicle and the police vehicle. He was ducked right behind it.
B: Now how can you see that from this angle? How do you know that he was behind there, cause right now this, how— (UI)
B: —how can you see where his friend was?
A: Like, like I say, he was standin' behind 'em.
B: Um hum.

6 Note that in the adversarial context, where self-promotion and contestation of opposing viewpoints play a greater role, *I*-oriented *say* forms are more frequent than *you*-oriented items (Szczyrbak 2016). It was also noted that in the (British) trial data analyzed, the forms *as I say* and *as I said* are preferred, rather than *like I say* or *like I said* found in the (American) police interview data analyzed in the current study.

7 When used by the interviewers, conversely, *like I said* seemed to mark amenity and solidarity, rather than uncertainty or insistence.

8 These instances can be contrasted with the "self-promotional" uses of *as I said* / *like I said* attested in academic genres (Gawlik 2010).

(b) A: Okay. Is there anything else that you can think of that you think would help us with this case? Any other information that, that you can think of that you saw that day, that you heard that day, that you smelled that day, anything else that you can think of that would help us?
B: No sir. Uh, no, '**cause like I said**, sat there and watched them with a clear vision. Like, everything I'm givin' you is . . .
A: Um hum.
B: My—you know, actually saw.

Additionally, a closer examination of the concordance lines revealed the following patterns with *like I said*: *and like I said* (ten tokens), *because*/*'cause like I said* (eight tokens), *but like I said* (seven tokens) and *like like I said* (five tokens), which in themselves might be worth investigating in a wider discoursal perspective. Also worthy of note is the cluster *like I said I* (17 tokens) suggesting, again, the speaker's perspective or even insistence (given the repetition of *I*).

The hearer's perspective, in turn, was visible in all the remaining *say* patterns, including *and you say* and *and you said*, which occurred with almost identical frequencies. As noted by Johnson (2014, 536), *and*- and *so*-prefaced questions serve "to summarise and resume the narrative with the police officer telling the story in the witness's words and formulating it as a question with assumed confirmation." Such was also the case with the examples with *say* attested by the data. As an illustration of this, consider (2), where *and you say* serves to elicit confirmation, even though, in this case, it prefaces an assertion and not a real question.

(2) A: Was it a big car or a small car?
B: Four door car.
A: Four door, ok. And so he reaches his hand out with a, the police officer reaches his hand out the window with his gun in it, **and you say** that the smaller boy runs.
B: He ran behind.
A: He ran behind. Did he run behind the car or to the passenger side of the car?
B: He ran this a way.
A: Like you have to say, was that to the back of the car?
B: Yeah.

In addition to the confirmation-seeking function, the evaluative role of *and you say*/*said* patterns should not be overlooked, either. It is with such forms that interviewers tend to mark their institutional dominance, framing respondents' narratives and, ultimately, "labelling" them as reliable or deceitful. Similar observations regarding the evaluative function of *say* and *tell* in courtroom interaction can be found in Taylor (2009, 218–220). As this study reveals, the past form *you said*, which is more frequent in hostile examination than in friendly examination, points to the witness's own testimony in a bid to undermine its

reliability. Likewise, *told us* is employed to contradict the witness's narrative and, unlike the present form *tell us*, it is never found in friendly examination. In the current study, on the other hand, *and you say* and *and you said* are equally frequent; however, while *and you say you* occurs only once, *and you said you* is attested by 12 tokens (six of which co-occur with the past form of the verb *saw*). In this case, it cannot be unambiguously decided whether this pattern is linked to positive or negative evaluation (or else, whether it threatens the respondent's face or not), as intonation marking is absent in the data.

Notwithstanding the above, certain vocal phenomena are traceable in the transcripts. As the data bear out, *and you say* tends to co-occur with agreement tokens as well as with what Culpeper and Kytö (2010) refer to as "pragmatic noise" (i.e. filled pauses with *eh*, *oh*, *ah* or *ha, ha* and inserts like *yeah*, expressing surprise, agreement or evaluation), for instance: *ok, and you say . . .* ; *and you say right*, *right, uh huh*. *Okay. All right*; *and you say yeah. And Detective says . . .*).[9] These items, I believe, can be looked at as instances of supportive feedback and non-coercive clarification requests, rather than hostility.

Similarly to the examples cited above, *when you say* clusters serve to elicit a response from the respondent, too. As already reported in Szczyrbak (forthcoming) analyzing trial and police interview data, *when you say* is favored by participants controlling discourse and it tends to co-occur with reformulations. These, it was proposed, are realized along the following schemata: *when-you-say-A,-(do)-you-mean-B?*; *when-you-say-A,-are-you-saying-B-or-C?* and *are-you-saying-A,-when-you-say-B?*, the last of which (the most coercive) was found only in the trial data. In agreement with these findings, (3) and (4) below illustrate the first two of the above patterns. In (3a), the questioner seeks clarification or requests the basis for the claim made by the respondent, trying to establish whether the latter actually heard the shots himself. In (3b)—where the *when-you-say-A,-are-you-saying-B-or-C?* pattern is at play—the police officer asks whether the interviewee actually *saw* the boy shot in the back or whether he only *assumed* that the boy was shot in the back.

(3) (a) A: **When you say** what you heard, you mean?
B: Based on the physical shootin' of a gun.
A: Like you could hear the shots?
B: Right, exactly.

(b) A: **And when you say** you had jumped, I know you also said that you know he was shot in the back because he jumped. Did you actually see him be shot, did you actually see him shot in the back or did you just assume he was shot in the back because he jumped?
B: Assume.

9 It is likely, however, that not all instances of "pragmatic noise" are marked in the interview transcripts.

Finally, it needs to be added that in the dataset analyzed, *when you say* was often used whenever the questioner wanted to disambiguate the reference of personal pronouns, as in (4) or to establish the exact position or directions, as in (5).

(4) A: **When you say** he grabbed him who are you saying . . .
B: I don't know if he grabbed him or not.
A: I know **when you say** he who is he?
B: The police officer.
A: Okay.

(5) A: . . . you see two males out there right?
B: Right-right one standing back there and one . . .
A: Wait, **when you say** standing back there . . .
B: Right, he was all like—like say this the truck, he's standing like back by the truck like about like this on the—on the passenger side.
A: Okay. On the passenger side of the truck.
B: Right, he standing back like this, yeah on the passenger side.

The last two patterns with *say* to be discussed here include the progressive form *saying*. As rightly noted by Johnson (2014, 545), "interviewing and cross-examination has a preference for present tense, present progressive and non-finite forms of SAY and this makes the focus of saying what is said rooted in the primary reality of the questioning activity." This is probably why *you're saying* turned out to be one of the most frequent *say* forms in the data. Needless to say, this cluster appeared in various configurations, including: *what you're saying* (19), *you're saying that* (9), *you're saying is* (5), *so you're saying* (5), *and you're saying* (3) and *now you're saying* (3), which may all be subsumed under the general schema (*and*/*now*/*so*/*what*) *you're saying* (*is*/*that*).

By analogy to *and*/*when you say*, *and*-/*now*-/*so*-/*what*-prefaced patterns with the progressive *saying* resemble confirmation-seeking (grammatical) questions, as illustrated in (6).[10] During the interviews, the respondent's confirmation was also elicited with the more forcible grammatical question *Is that what you're saying?* (attested by seven tokens), as shown in (7). In this instance, the interviewer reformulates the respondent's original wording "he has his shirt now" by asking "So, he's grabbing the sleeve of the shirt?" and awaiting confirmation.

10 Cf. Jones's ([2008, 65] quoted in Eades [2010, 181]) distinction into information seeking *so*-prefaced questions (i.e. open questions, *wh*- questions, yes/no questions) and confirmation seeking *so*-prefaced questions. The latter include *gists* which summarize the prior talk and *upshots* which "draw out a relevant implication which [the interviewer] is expected to ratify."

(6) A: But the police officer was on the driver's side?
B: He was
A: Well he was driving the car
B: Like, well he hit the one on the passenger side then, cause when they cut him off they leant in and got him
A: Ok, so **you're saying** that there was, with the guy that you referred to the police officer as Ears, **you're saying** that he had a partner in the car with him
B: Yeah
A: Ok, so, that's right **cause you're saying** that there was four (4) officers
B: Yeah
A: Ok, so whoever he's struggling with is the guy in the passenger side. . . Yeah . . . - but not the driver's side?
B: Uh mmm (no)

(7) A: He has his shirt now.
B: So, he's grabbing the sleeve of the shirt? **Is that what you're saying?**
A: Yes, correct.

Unlike the coercive confirmation-seeking devices referred to above, *you know what I'm saying*, which was found only in the respondent's speech, conversely, seemed to betray the speaker's uncertainty rather than dominance, as demonstrated in (8). Here, the interviewee repeats several times the words *you know* which have a monitoring function and which encourage the hearer to consider the sense of what has just been said (Schiffrin 1987, 310).

(8) A: So at the point that-that he starts walking back or moving back towards the officer is the point where you are now focused on your vehicle turning your vehicle around is that a fair assessment?
B: Oh yeah, I was then I was looking you know like to see whose, anybody else seeing this **you know what I'm saying** you know. That's when I saw all the people and this white car behind me I was trying not to hit that 'cause I was trying to, ya know, get out of there **you know what I'm saying**?
A: Okay, alright and so you turned around and you went back towards—
B: I went back to West Florissant . . .

As the above examples demonstrate, in the police interview context, the recruitment of *say* is everything but "random sprinkling" (cf. Fox Tree and Schrock 2002) and it is related to the moment of use and speaker status in interaction. Regarding the more frequent *you*-oriented *say* forms, they, as was established, are used by the questioners to frame and assess the accounts provided by the respondents. With respect to the less

common *I*-oriented items, it was noted in turn that they organize the respondents' talk, pointing to their insistence or uncertainty.

3.4 "Saying" Routines in Police Interview vs. Trial Data

Unlike the police interview material discussed above, trial data seem to suggest different *say* patterns. In the study reported in Szczyrbak (2016), based on an analysis of transcripts from a libel trial,[11] it was revealed that *I*-oriented forms are preferred over *you*-oriented ones, although, admittedly, *you are saying* proved to be the most frequent of all the analyzed items. Another observation which emerges when the normed scores from the two datasets (police interview and trial data) are compared (Table 3) is that, slightly surprisingly, there is more "saying" in police interviews than there is in the courtroom,[12] where, however, high frequencies of other communication verbs can be predicted.[13] Interestingly, while *you are saying* is the most frequent choice in the trial data, in the police interview material *you're saying* ranks fifth and *you are saying* ranks seventh among the most frequent items. However, it should be noted, when the respective frequencies are examined, it turns out that, when counted together, *you are saying* and *you're saying* are almost twice as frequent in police interviews as *you are saying* in the trial data (where the contracted form is absent).[14] This, in turn, may indicate greater emphasis on intersubjectivity in the interview context, with the questioners projecting subjectivity to the respondents. In the courtroom setting, on the other hand, the *I* perspective is more pronounced, which is understandable given that competing accounts and interpretations are provided and, subsequently, assessed. In such a confrontational context, the need to "promote" one's own standpoint and to construct one's credibility and authority is much stronger than is the case in police interviews. Again, the use of *say* reflects these interactional goals and supports the view that saying "what has been and is being said in prior texts, in present texts and across texts and contexts" is central in evidence construction (Holt and Johnson 2010, 35). Naturally, that is not to say that other communication verbs do not play a role. What it does show, however, is that the use of *say* is context-sensitive and that it is linked to power and the speaker's immediate interactional goals.

11 The analysis was based on 32 transcripts (totalling approx. 1.5m words) documenting a high profile libel trial. The transcripts were downloaded from Holocaust Denial on Trial (2013) .

12 Of course, at this point no generalizations should be made, as the analysis is based on transcripts from one trial.

13 It may also be speculated that *say*, as the primary verb of speaking, is more frequent in police interviews, since these are linked to less formal language than courtroom examinations, where a greater degree of formality is expected.

14 The pragmatics of the progressive forms of verbs of speaking, such as *saying*, *talking* or *suggesting*, is another interesting issue which, due to space constraints, cannot be addressed here. For a discussion of such forms in hostile examination, see, e.g., Taylor (2009).

Police Interviews	Normed score (per million words)	Trial	Normed score (per million words)
like I said	309.34	*you are saying*	92.28
and you say	202.55	*I would say*	86.89
and you said	198.86	*I have to say*	76.12
when you say	184.13	*if I may say so*	67.36
you're saying	165.72	*what you say*	51.87
you know what I'm saying	147.31	*as I say*	48.50
you are saying	103.11	*what I said*	48.50

Table 3. Distribution of *say* in police interview vs. trial data

4. Conclusions

From the foregoing discussion several points emerge. Firstly, as already stated, both in police interviews and in trial discourse, the verb *say* is not used randomly. Secondly, there is a visible correlation between the speaker's role in interaction (questioner, respondent) and their preference for individual *say* patterns. Thirdly, in police interviews, hearer-oriented forms are favored, whereas in courtroom talk (or at least in the trial under scrutiny) speaker-oriented patterns are more frequently selected. It should also be observed that regardless of the above general patterns, in the interview data, interestingly, the speaker-oriented form *like I said* was the most frequent choice. In the trial data, on the other hand, the hearer-oriented item *you are saying* ranked as the most common. Further, as regards the speaker-form correlation in the interview material, the respondents opted for *like I said* and *you know what I'm saying*, while the questioners deployed *and you say*/*said, when you say* and (*and*/*now*/*so*/*what*) *you're saying* (*is*/*that*). In sum, the study aimed to demonstrate that in the legal context, *say* is strategically deployed by speakers and that it reveals the interactional asymmetry between them. Put differently, it was argued that the most common communication verb indexes stance and the perspective from which institutional and non-institutional speakers view and evaluate their own narratives as well as those of their interlocutors.

Works Cited

Baldwin, John. 1993. "Police Interview Techniques: Establishing Truth or Proof." *British Journal of Criminology* 33 (3): 325–52.

Berk-Seligson, Susan. 1999. "The Impact of Court Interpreting on the Coerciveness of Leading Questions." *Forensic Linguistics* 6 (1): 30–56.

du Bois, John W. 2007. "The Stance Triangle." *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, edited by Robert Englebretson, 139–82. Amsterdama: John Benjamins.

Culpeper, Jonathan, and Merja Kytö. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Studies in English Language. Cambridge: Cambridge University Press.

Eades, Diana. 2010. *Sociolinguistics and the Legal Process*. Bristol: Multilingual Matters.

Fox Tree, Jean E., and Josef C. Schrock. 2002. "Basic Meanings of *You Know* and *I Mean*." *Journal of Pragmatics* 34: 727–47.

Gaines, Philip. 2016. "Nonconsent and Discursive Resistance. Radical Reformulation in a Post-sting Police Interview." *Discursive Constructions of Consent in the Legal Process*, edited by Susan Ehrlich, Diana Eades, and Janet Ainsworth, 213–38. Oxford: Oxford University Press.

Gawlik, Oskar. 2010. "Basic *Verba Dicendi* in Academic Spoken English." PhD diss., Uniwersytet Śląski, Katowice.

Gibbons, John. 2005. *Forensic Linguistics: An Introduction to Language in the Justice System*. Malden: Blackwell.

Haworth, Kate. 2009. "An Analysis of Police Interview Discourse and Its Role(s) in the Judicial Process." PhD diss., University of Nottingham, Nottingham.

Holocaust Denial on Trial. 2013. "Trial Transcripts," Accessed January 31, 2013. https://www.hdot.org/trial-materials/trial-transcripts/.

Holt, Elizabeth and Alison Johnson. 2010. "Socio-pragmatic Aspects of Legal Talk: Police Interviews and Trial Discourse." *The Routledge Handbook of Forensic Linguistics*, edited by Malcolm Coulthard and Alison Johnson, 21–36. London: Routledge.

Johnson, Alison. 2014. "Legal Discourse: Processes of Making Evidence in Specialised Legal Corpora." *Pragmatics of Discourse*. Vol. 3 of the *Handbooks of Pragmatics*, edited by Klaus P. Schneider and Anne Barron, 525–54. Berlin: De Gruyter Mouton.

Jones, Claire. 2008. "UK Police Interviews: A Linguistic Analysis of Afro-Caribbean and White British Suspect Interviews." PhD diss., University of Essex, Colchester.

Partington, Alan, Alison Duguid, and Charlotte Taylor. 2013. *Patterns and Meanings in Discourse. Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins.

Schiffrin, Deborah. 1987. *Discourse Markers.* Cambridge: Cambridge University Press.

Scott, Mike. 2012. WordSmith Tools. Version 6. Oxford: Lexical Analysis Software.

Shuy, Roger W. 1998. *The Language of Confession, Interrogation, and Deception*. Thousand Oaks: Sage.

St. Louis Post-Dispatch. 2014. "Transcripts of Police and FBI Interviews." Accessed June 10, 2015. http://www.stltoday.com/news/multimedia/special/transcripts-of-police-and-fbi-interviews/html_0c5bbc4c-c0b9-5aec-821d-937846293161.html.

Szczyrbak, Magdalena. 2016. "*Say* and Stancetaking in Courtroom Talk: A Corpus-Assisted Study." *Corpora* 11 (2): 143–68.

Szczyrbak, Magdalena. Forthcoming. "*When You Say Over Here, You Mean* . . . Reformulation Strategies in Confrontational Institutional Talk."

Taylor, Charlotte. 2009. "Interacting with Conflicting Goals." *Corpus-Assisted Discourse Studies on the Iraq Conflict. Wording the War*, edited by John Morley and Paul Bayley, 208–33. New York: Routledge.

Language Use and Linguistic Structure
Proceedings of the Olomouc Linguistics Colloquium 2016

June 9–11, 2016
Faculty of Arts, Palacký University Olomouc, Czech Republic
http://olinco.upol.cz
e-mail: olinco@upol.cz

Edited by Joseph Emonds and Markéta Janebová

Series: Olomouc Modern Language Series
Executive editor: doc. Mgr. Jiří Špička, Ph.D.
In-house editor: Mgr. Kristýna Bátorová
Typesetting and Cover Design: Gobak DTP

The publication did not pass the language editing review of the publishing house.

Published by Palacký University Olomouc
Křížkovského 8, 771 47 Olomouc
www.vydavatelstvi.upol.cz
www.e-shop.upol.cz
vup@upol.cz

First Edition

Olomouc 2017

ISBN 978-80-244-5173-2
(online: PDF; available at http://anglistika.upol.cz/olinco2016proceedings/)

ISBN 978-80-244-5172-5 (print)

VUP 2017/0171