RAILS

Rational Approaches in Language Science

13th-15th February 2025

IDeaL SFB 1102

DFG Deutsche Forschungsgemeinschaft

UNIVERSITÄT DES SAARLANDES

RAILS Rational Approaches in Language Science Saarbrücken 2025

Contents

Organizing Committee

Regine Bader
Stefania Degaetano-Ortlieb
Katja Haeuser
Robin Lemke
Ivan Yuen

Program/Review Committee

Diego Alves, Bistra Andreeva, Tania Avgustinova, Regine Bader, Peter Bourgonje, Matthew W. Crocker, Stefania Degaetano-Ortlieb, Francesca Delogu, Vera Demberg, Stefanie Dipper, Heiner Drenhaus, Koel Dutta Chowdhury, Emilia Ellsiepen, Cristina España-Bonet, Katja Haeuser, Bozhidara Hristova, Omnia Ibrahim, Torsten Jachmann, Benedict Krieger, Marie-Pauline Krielke, Maria Kunilovskaya, Ekaterina Lapshinova-Koltunski, Robin Lemke, Marian Marchal, Kate McCurdy, Julia Meßmer, Bernd Möbius, Ingo Reich, Margarita Ryzhova, Lisa Schäfer, Linda Sommerfeld, Luigi Talamo, Elke Teich, Annemarie Verkerk, Sophia Voigtmann, Lena Wieland, Wei Xue, Ivan Yuen, Heike Zinsmeister, Doruntinë Zogaj

Steering Committee

Matthew W. Crocker
Marie-Ann Kühne
Sabine Loskyll
Heike Przybyl
Elke Teich

Welcome

RAILS 25 followed the footsteps of its predecessor to bring together researchers from a wide range of disciplines, who are interested in the idea that language and its use can be better understood by considering rational explanations. The goal is to benefit from and share research from a wide range of disciplines using diverse methodologies that explore the idea that language users continuously strive to optimize their means of communication to effectively convey their intended messages. Rational communication not only influences how recipients encode and remember information, but also shapes language variation and change over time. We are delighted that the scientific contributions to RAILS 2025 reflect such diversity of disciplines and methodologies: from speech to discourse, from online processing through corpus-based investigation to computational modelling, and from information updating through aspects of short- and long-term mnemonic processes to language change and typology. Scientific and financial support for this conference comes from **SFB1102** “Information Density and Linguistic Encoding”, a Collaborative Research Center funded by the German Research Foundation (DFG). We are grateful for your contribution to and participation in our event, and look forward to fruitful exchanges and insights from all across the language sciences.

Welcome to Saarbrücken!

Schedule

Thursday 13 February	
09:00 – 09:30	Registration + Coffee
09:30 – 09:45	Opening
09:45 – 10:45	**Adele Goldberg (Keynote)** Rational Productivity: Using a System of Learned Constructions to Express New Messages
10:45 – 11:15	**John Duff, Delaney Gomez-Jackson, Fe Silva Robles, Maziar Toosarvandani and Matthew Wagers** Crosslinguistic Variation in Structural Prediction as Learned Behavior
11:15 – 11:45	Coffee
11:45 – 12:15	**Markus Bader and Michael Meng** Constraints on Word Exchanges during Noisy-Channel Inference
12:15 – 12:45	**Anna Kapron-King, Lauren Fletcher, Aida Tarighat, Radina Dobreva, Stephanie Droop, Chris Cummins and Hannah Rohde** Tell It Like It (Usually) Is(n't): Speakers' Mention of Instruments in Reddit Corpus Reflects Instrument (A)typicality
12:45 – 14:00	Lunch
14:00 – 14:30	**Vsevolod Kapatsinski** Probabilistic Inference and Frequency Effects in Language Change
14:30 – 15:00	**Wei Xue, Julius Steuer, Dietrich Klakow and Bernd Möbius** Investigating the Correlation between Human Predictability Judgements and Computational Estimates from Text- and Audio-Based Models
15:00 – 15:30	**Regine Bader, Samira Engels and Axel Mecklinger** Confirmed and Violated Predictions Benefit Long-Term-Memory
15:30 – 16:00	Coffee
16:00 – 17:00	**Rachel Ryskin (Keynote online)** Language Comprehension Adapted to the Environment
19:00 – late	Dinner at Casino am Staden

Friday 14 February	
09:00 – 10:00	**Richard Futrell (Keynote)** Prediction and Locality in Language and Language Models
10:00 – 10:30	**Kate McCurdy and Michael Hahn** Lossy Context Surprisal Predicts Task Differences in Relative Clause Processing
10:30 – 11:00	Coffee
11:00 – 11:30	**Xinyue Jia, Christoph Aurnhammer, Torsten Jachmann, Francesca Delogu, Heiner Drenhaus and Matthew W. Crocker** Lossy Context Surprisal: Influence of Linearization on Expectation
11:30 – 12:00	**Diego Alves, Stefan Fischer and Elke Teich** Paradigmatic Variability of Multi-Word Expressions in Scientific English
12:00 – 12:30	**Simon Wehrle and Malin Spaniol** Effects of Conversational Context on Turn-Timing in (Non-)Autistic Dyads
12:30 – 13:30	Lunch (provided)
13:30 – 14:30	*Poster Session 1*
14:30 – 15:00	**Robin Lemke** Testing a Rational Account of Fragment Usage with Crowd-Sourced Production Data
15:00 – 15:30	**Chris Cummins and Hannah Rohde** Joint Inference about Pragmatically-Relevant Contextual Features
15:30 – 16:00	Coffee
16:00 – 16:30	**Snecha Chetani, Iza Škrjanec and Vera Demberg** Analyzing the Effects of Temperature-Scaled Surprisal for Subword Reading Times
16:30 – 17:00	**Ellis Cain, Alton Chao and Rachel Ryskin (online)** Diachronic Language Change Explains Apparent Age-Related Differences in Information-Theoretic Efficiency

Saturday February 15	
09:00 – 10:00	**Marc Dingemanse (Keynote)** Reasoning in Interaction
10:00 – 11:00	*Poster Session 2*
11:00 – 11:30	Coffee
11:30 – 12:00	**Jingwen Cao, Lawrence Yam-Leung Cheung** A Multifactorial Corpus-Based Analysis of Classifier Positioning in Mandarin Relative Clauses
12:00 – 12:30	**Sophia Voigtmann** It's All in the Past. An Experimental and Rational Approach to the Influence of the Sentence Onset on Past Tense Choice in German
12:30 – 13:00	**Julius Steuer, Marie-Pauline Krielke, Stefania Degaetano-Ortlieb, Elke Teich and Dietrich Klakow** Quantifying the Development of Communicative Efficiency in Scientific English
13:00 – 13:15	Closing

Abstracts

Reasoning in Interaction

Mark Dingemanse (Radboud University)
mark.dingemanse@ru.nl

In rational communication, cognitive constraints and considerations of efficiency conspire to explain properties of message formulation and interpretation. Given the context of RAILS I will consider the attractions of quantitative and probabilistic approaches to language use evident, so I will use my time to go off the beaten path. In particular, I will argue for the utility of moving beyond single-minded views of rationality and communication, and I will explore opportunities for productive exchange between empirical work on human interaction and computational models of language use (Dingemanse & Enfield 2024). I will show how phenomena like interactive repair, parental scaffolding and liminal signs allow us to transcend individual resource limitations, distribute cognitive processes, and ride the coattails of ambiguity. The stark elegance of strategic one-shot communication may be the limiting case of processes that are better understood by studying language use as an interactional achievement.

Prediction and Locality in Language and Language Models

Richard Futrell (University of California, Irvine)
rfutrell@uci.edu

I argue that human language is shaped by constraints on memory in online language comprehension and production. One way the bottleneck shows up as a preference for locality in the order of elements. Using cross-linguistic corpora of 55+ languages, I show evidence for dependency locality, a pressure for syntactically related words in sentences to be close to each other. I show how a more sophisticated model of language processing, based on incremental probabilistic prediction under resource constraints, yields a generalization of dependency locality called information locality, which I show correctly predicts adjective order across languages. Next, I formulate a general information-theoretic measure of the complexity of sequential prediction, and show cross-linguistic corpus evidence that phonological forms and morphological paradigms are structured in a way that minimizes this complexity. Finally, I present evidence that modern large language models also have a bias towards information locality, and that this may partially explain their successes in learning human language.

Rational Productivity: Using a System of Learned Constructions to Express New Messages

Adele Goldberg (Princeton University)
adele@princeton.edu

In order to communicate, we each learn a complex, dynamic system of constructions, a ConstructionNet. Mismatches between what is expected and what is witnessed fine-tune our network of learned constructions via *competition-driven* learning (statistical preemption). To express novel messages, we must combine familiar constructions in new ways; such productive combinations have given us wugs; tweeted; humble brag; Ok, Boomer, and is (not) a thing. Productive combinations of constructions also allow us to talk about a period three hairstyles ago or explain that we napped our way across the Atlantic. Granted certain caveats, evidence is reviewed that novel combinations are generally judged less acceptable to the extent that there exists a “better” (conventional) way of expressing the same intended message-in-context (e.g., say to me > ?say me; succeeded in doing > ?succeeded to do).

Language Comprehension Adapted to the Environment

Rachel Ryskin, (University of California, Merced)
rryskin@ucmerced.edu

In order to understand each other across diverse contexts, humans must continuously adapt their linguistic expectations. Yet, the core of their language knowledge must remain stable. My research aims to understand how humans balance flexibility and stability in language comprehension in order to efficiently exchange information in the face of variability and noise. I will first review evidence that comprehenders learn from their environment at multiple levels including adapting to the probability of syntactic structures, the kinds of errors the speaker makes, and the noise in the input. I will then discuss work investigating the constraints on this continuous learning. For instance, studies with individuals across the lifespan indicate that word meanings and syntactic biases are learned on different timescales. And work with individuals with aphasia — a language disorder caused by stroke — suggests that they may not update their representations of errors in the environment as rapidly as healthy language users.

Constraints on Word Exchanges During Noisy-Channel Inference

Markus Bader (University Frankfurt) & Michael Meng (Merseburg University)
bader@em.uni-frankfurt.de

According to the Noisy Channel Model of Gibson et al. (2013), communication can succeed even when the input is corrupted because comprehenders rationally infer the speaker's intended meaning based on the a-priori probability of the literal interpretation and the probability that the input has been corrupted by noise. A major point of debate concerns what kind of corruptions comprehenders take into account. Whereas there is consensus that insertions and deletions are considered a possible source of noise, the status of word exchanges is less clear (Poppels and Levy, 2016).

To test whether and under which conditions word exchanges can be observed, we ran four online experiments on processing three types of simple German sentences: subject-before-object sentences (SO), object-before-subject sentences (OS), and passive sentences (see (1)). SO, OS and passive sentences provide an interesting test case because implausible sentences can be ”repaired” by exchanging function words or by exchanging nouns (see (2) for SO). As in Gibson et al. (2013), sentences were presented in full along with a yes-no question to probe interpretation. Exp. 1 (N=48) tested plausible and implausible SO and OS sentences and varied whether a word exchange would cross a main verb or an auxiliary. Exp. 2 (N=74) included plausible and implausible passive sentences in addition to SO and OS sentences. Exp. 3 (N=78) tested implausible SO, OS and passive sentences and varied the proportion of implausible sentences in the total stimulus set (high: 50% vs. low: 15%). Exp. 4 (N=36) tested implausible SO, OS and passive sentences but required explicit corrections of implausible sentences in addition to answering yes-no questions.

Results are shown in Figure 1. The results were analysed using Baysian mixed-effect modeling. We consistently found that implausible SO and passive sentences elicit few non-literal interpretations whereas the rate of non-literal interpretations is high for implausible OS sentences. This holds regardless of whether word exchanges have to cross a main verb or an auxiliary (Exp. 1) and, as predicted by the Noisy Channel Model, is more pronounced if the overall proportion of implausible sentences is low (Exp. 3). Thus, exchanges of function words of the same syntactic category are considered, but not noun exchanges. Moreover, word exchanges are considered only when resulting in a more likely syntactic structure, supporting the idea that comprehenders' noise model is structure-sensitive (Poppels and Levy, 2016). This prevents function word exchanges to be applied to SO and passive sentences. Finally, Exp. 4 showed that comprehenders use noun exchanges to a much higher extent when asked to provide explicit corrections, in line with Ryskin et al. (2018). This suggests that constraints on word exchanges depend on whether or not sentences are corrected consciously.

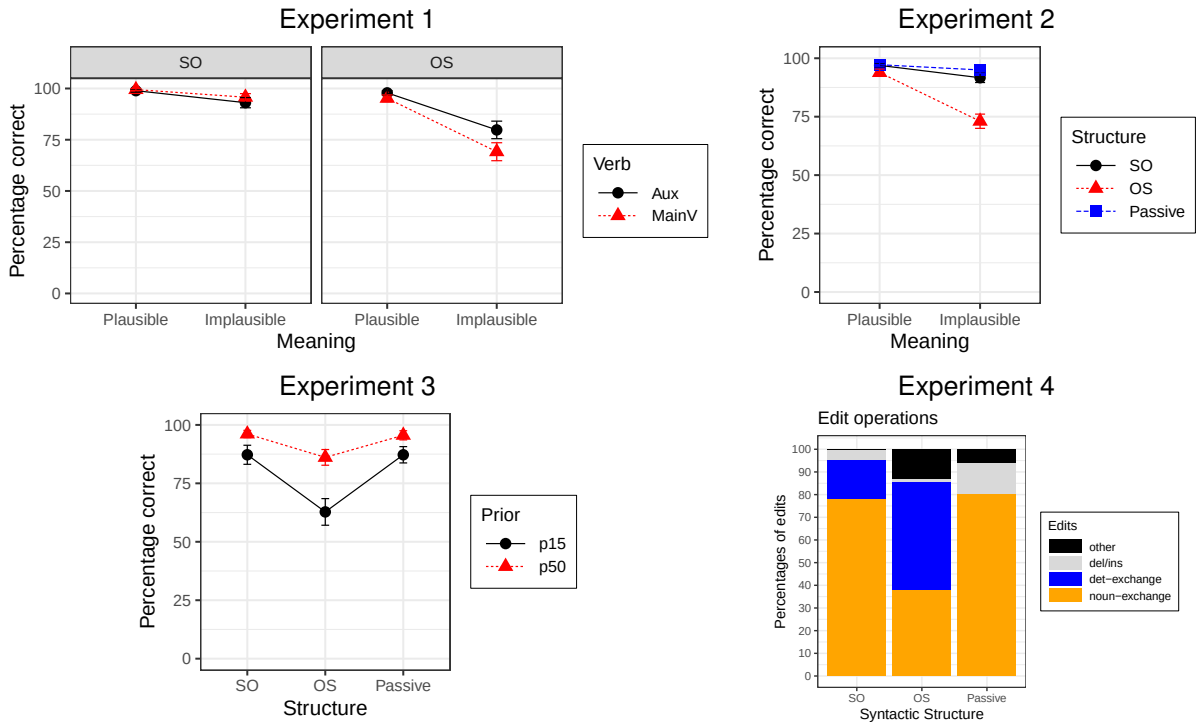


Figure 1: Percentages of correct answers in Exp. 1–3 and distribution of edit operations in Exp. 4.

(1) Implausible versions of the experimental sentences (plausible versions are obtained by exchanging nouns)

a. [SO:] Der Knochen hat den Hund gegessen.
the$_{\text{NOM}}$ bone has the$_{\text{ACC}}$ dog eaten

b. [OS:] Den Hund hat der Knochen gegessen.
the$_{\text{ACC}}$ dog has the$_{\text{NOM}}$ bone eaten

c. [Passive:] Der Hund wurde vom Knochen gegessen.
the.NOM dog was by-the bone eaten

(2) [Presented – SO:] Der Knochen hat den Hund gegessen.
the$_{\text{NOM}}$ bone has the$_{\text{ACC}}$ dog eaten

a. [Noun exchange – SO:] Der **Hund** hat den **Knochen** gegessen.
the$_{\text{NOM}}$ dog has the$_{\text{ACC}}$ bone eaten

b. [Det exchange – OS:] **Den** Knochen hat **der** Hund gegessen.
the$_{\text{ACC}}$ bone has the$_{\text{NOM}}$ dog eaten

References

Gibson, E., Bergen, L., and Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

Poppels, T. and Levy, R. P. (2016). Structure-sensitive noise inference: Comprehenders expect exchange errors. In Papafragou, A., Grodner, D., Mirman, D., and Trueswell, J. C., editors, *Proceedingsof the 38th Annual Meeting of the Cognitive Science Society*, pages 378–383, Austin, TX. Cognitive Science Society.

Ryskin, R., Futrell, R., Kiran, S., and Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, 181:141–150.

Confirmed and Violated Predictions Benefit Long-Term-Memory

Regine Bader (Saarland University), Samira Engels (Saarland University), & Axel Mecklinger (Saarland University)
regine.bader@mx.uni-saarland.de

It is widely acknowledged that predicting up-coming information plays a pivotal role in online language comprehension. However, the down-stream effects on the retention of this information on the long run have been less thoroughly explored. On one hand, schema-based memory theories suggest that predicted information is remembered more effectively as it aligns with prior knowledge, i.e. existing schemas (van Kesteren et al., 2012). Alternatively, if information merely confirms a prediction, it may be processed more superficially leading to weaker retention (Hubbard et al., 2024). On the other hand, unpredicted information may be particularly memorable because it generates prediction errors (PE; van Kesteren et al., 2012). PEs signal a shift in the processing context, potentially prompting an update of the current situation model, thereby serving as a cue for learning. Research on the mnemonic consequences of confirmed and disconfirmed predictions in language comprehension has yielded mixed findings (Haeuser & Kray, 2022; Höltje & Mecklinger, 2022; Hubbard et al., 2024). These inconsistencies may be addressed by comparing memory for both predicted and unpredicted information with an appropriate baseline condition. In such a baseline, only minimal predictions should be possible, thus avoiding both confirmation and violation. Moreover, for long-term retention it might make a difference whether prediction violations are still plausible or anomalous as only the former might lead to an update of the situation model. In our study, participants engaged in two study-test blocks. In the study phases, they read brief two-sentence statements with the sentence-final word of the second sentence (target) being either expected, unexpected but plausible, or completely anomalous. Of note, the sentences also varied in their degree of constraint: some were strongly constraining (1), while others were only weakly constraining (2). In the weakly constraining sentences, participants were unlikely to form predictions about the target. Therefore, these sentences constitute an appropriate baseline.

(1) (a) The birthday party was over, and Helene wanted to go home quickly. She ordered herself a taxi. (expected)
(b) The birthday party was over, and Helene wanted to go home quickly. She ordered herself some food. (unexpected)
(c) The birthday party was over, and Helene wanted to go home quickly. She ordered herself a pillow. (anomalous)

(2) (a) Mathilde knew exactly what she wanted to do next. She ordered herself a taxi. (matched to SC expected)
(b) Mathilde knew exactly what she wanted to do next. She ordered herself some food. (matched to SC unexpected)

After a 3 minutes retention interval, participants discriminated between previously presented target words and new words. Preliminary analyses (n=33) (Fig. 1) indicate that both predicted and unpredicted targets, whether plausible or anomalous, were remembered better than unexpected targets presented in weakly constraining sentences. These results align with the idea that both schema congruency and PE contribute to the long-term retention of information encountered during language comprehension.

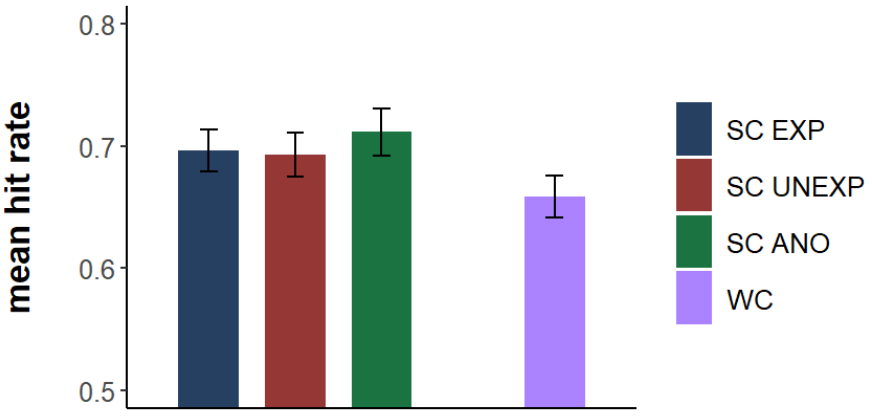


Figure 1. Mean hit rates for targets in the strong constraining expected (SC EXP), strong constraining unexpected (SC UNEXP), strong constraining anomalous (SC ANO), and weak constraining (WC) conditions. WC hit rates are averaged across both types of targets. Error bars depict the standard error of the mean difference.

References

Haeuser, K. I., & Kray, J. (2022). How odd: Diverging effects of predictability and plausibility violations on sentence reading and word memory. *Applied Psycholinguistics*, *43*(5), 1193–1220. https://doi.org/10.1017/S0142716422000364

Höltje, G., & Mecklinger, A. (2022). Benefits and costs of predictive processing: How sentential constraint and word expectedness affect memory formation. *Brain Research*, *1788*, 147942. https://doi.org/10.1016/j.brainres.2022.147942

Hubbard, R. J., & Federmeier, K. D. (2024). The Impact of Linguistic Prediction Violations on Downstream Recognition Memory and Sentence Recall. *Journal of Cognitive Neuroscience*, *36*(1), 1–23. https://doi.org/10.1162/jocn_a_02078

van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. Trends in Neurosciences, 35, 211–219. doi:10.1016/j.tins.2012.02.001

Diachronic Language Change Explains Apparent Age-Related Differences in Information-Theoretic Efficiency

Ellis Cain, Alton Chao, Rachel Ryskin (University of California, Merced)
ecain@ucmerced.edu

During production, speakers make choices which appear consistent with a pressure toward information-theoretic efficiency (ITE). For instance, some meanings can be expressed using multiple forms which differ in length (e.g., *A/C* and *air-conditioning*). Mahowald et al. (2013) previously showed that shorter forms were preferred (relative to long forms) in supportive than in neutral contexts. Previous work has argued that the ability to engage in prediction during language processing is reduced in older adults (Federmeier, 2010; Wlotko & Federmeier, 2012), which may influence their ITE during communication. Alternatively, age-related differences may be explained by exposure to changing language statistics (Ryskin & Nieuwland, 2023). Here, we examine how preferences differ across the lifespan and further explore how diachronic change in usage patterns (Michel et al., 2011) may explain these differences.

Methods: We recruited 126 English-speaking participants through Prolific (Age range: 20-60 years old, M = 39.64 y.o., 49% female). As in Mahowald et al. (2013), participants were presented with either a supportive or neutral sentence stem and had to choose between a short and long version of a word (e.g., 'A/C' or 'air-conditioning'; n = 40 word pairs). The context type of each newly written stem was verified using GPT-2 (Radford et al., 2019), such that the average surprisal of the full sentences for both short and long forms was lower in supportive (M = 3.71) than in neutral contexts (M = 5.66).

Results: Overall, the short form was more likely to be chosen in the supportive context, relative to the long form (Fig. 1). The frequencies of both the short and long forms varied over time (Fig. 2). We capture this change by the difference in short form frequencies between the 1960s and 2000s for each pair. Figure 3 shows the proportion of choosing short across the age range, grouped by change in short form frequency. Older adults (OA) appear less likely to choose the short form particularly for larger changes in frequency. We trained a Bayesian multilevel model[1] to predict whether participants chose the short form based on context, age, and change in frequency. Replicating Mahowald et al. (2013), participants were more likely to choose the short form when the context was supportive relative to neutral ($\beta_{context}$ = 0.38, 95% CrI = [-0.04, 0.79]), though the credible interval of the effect includes zero. The model also confirmed that OA were less likely to choose the short form (β_{age} = -0.51, [-0.79, -0.25]). There was a modest interaction between age and context, with the effect of age being reduced in supportive relative to neutral contexts ($\beta_{age*context}$ = 0.08, [-0.05, 0.20]). There was no main effect of frequency change, but there was an interaction with age ($\beta_{age*change}$ = -0.09, [-0.19, 0.01]), such that OA became even less likely to choose the short form for larger changes in short form frequency. The 95% CrI for the interactions included zero. There was no evidence of a 3-way interaction.

Conclusion: Replicating and extending Mahowald et al. (2013), we found that lexical choices are driven by a pressure for information-theoretic efficiency, but the tendency to use shorter forms is increased in YA. We also found that this increased tendency may be tied to patterns of language change. When the short form was previously much lower in frequency than the long form, OA were less likely to use the short form, suggesting that OA demonstrate a similar pressure toward information-theoretic efficiency in communication as young adults but the lexical choices of YA may be influenced by usage statistics from recent decades. We plan to collect more data to replicate these findings.

References:

Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. Brain and language, 115(3), 149–161.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. Cognition, 126(2), 313–318.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. Science (New York, N.Y.), 331(6014), 176–182.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: what is next?. *Trends in cognitive sciences*, *27*(11), 1032–1052. https://doi.org/10.1016/j.tics.2023.08.003

Wlotko, E. W., & Federmeier, K. D. (2012). Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension. Psychophysiology, 49(6), 770–785.

Models:

1. Chose short ~ context * age * change + (1 + context * age | pair) + (1 + context * change | subject)

Figures:

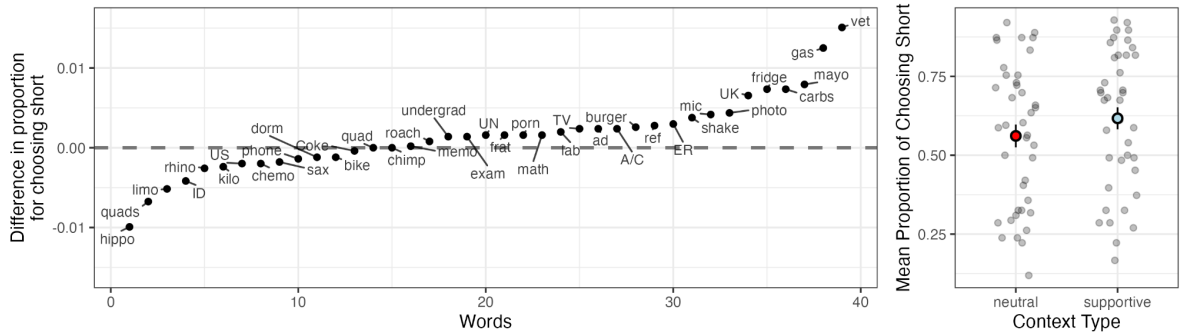


Figure 1: Left shows the difference in the proportion of choosing the short form in the supportive context relative to the neutral context, across the different word pairs. Right shows the overall average proportion of choosing the short form for each context type.

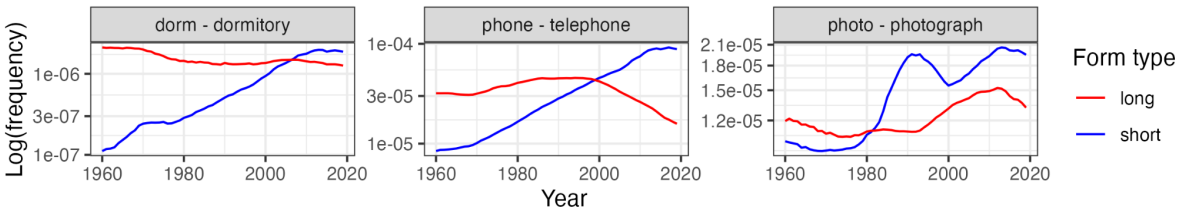


Figure 2: Diachronic trends in long and short form frequency for a small subset of the word pairs, gathered from the Google books corpus. For 31 out of 40 word pairs the short form was less frequent than the long form in the 1960s. 25 of the short forms increased in frequency between the 1960s and 2000s and 3 surpassed their long forms by the 2000s.

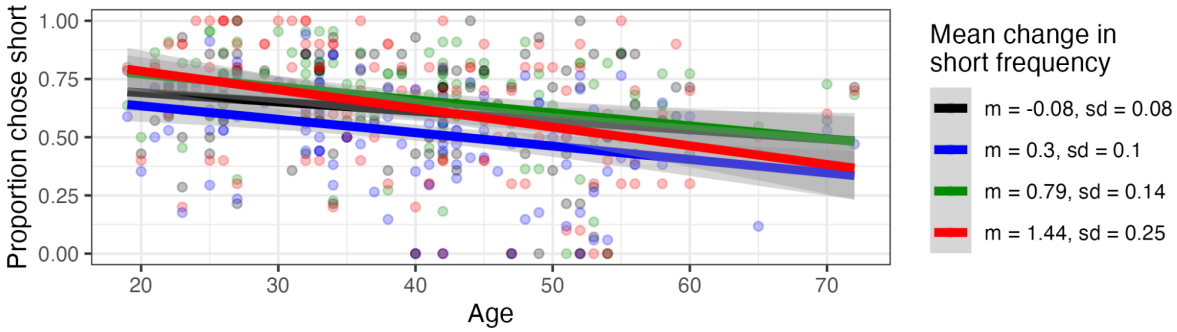


Figure 3: Proportion of choosing the short form in the supportive context, relative to the neutral context, by participant age. Color represents the bins for the average change in short form frequency.

Crosslinguistic Variation in Structural Prediction as Learned Behavior

John Duff (Saarland University), Delaney Gomez-Jackson (Motorola Mobility), Fe Silva Robles (Senderos), Maziar Toosarvandani (UC Santa Cruz), and Matt Wagers (UC Santa Cruz)
jduff@lst.uni-saarland.de

Cross-linguistic variation in sentence processing behavior has provided critical evidence for the source of that behavior, e.g. [1,2]. To this literature, we add a puzzling observation from a visual-world eye-tracking study on the incremental comprehension of relative clauses (RCs) in Santiago Laxopa Zapotec (SLZ). Despite rapid and accurate RC interpretation, sensitive to expected effects of similarity-based interference [3], participants showed no sign of structural prediction based on the animacy of the head noun, an effect familiar in other languages [4-6]. We argue that this variation can be best explained if language-specific experience determines whether comprehenders engage in procedural strategies like structural prediction. We see this as a natural result of treating processing as a learned human skill [7].

SLZ is an Oto-Manguean language of southern Mexico with VSO word order. Transitive RCs feature one pre-verbal argument (the head) and one post-verbal co-argument. They are ambiguous between interpretations where the head serves as the RC subject vs. the RC object (SRC/ORC) (1), unless they feature a grammatical resumptive pronoun (RP) marking the subject or object dependency explicitly (2-3). RPs are highly productive even in simple RCs. RPs and other pronouns mark animacy, e.g. HU(man) vs. IN(animate).

Methods Eye movements were recorded from 62 native speakers of SLZ (after exclusions) as they listened to stimuli (1) with relative clauses specifying which of two pictures to select, including in 24 critical trials (Table 1) crossing *Dependency Type* (Gap/ObjRP), *Head Animacy* (HU/IN), and *Co-Argument Animacy* (±Match).

Results We analyzed likelihood of new fixations on target images binarized by region in logistic m.-e. models in brms. Gap conditions received equibiased responses regardless of head animacy, and new fixations to SRC images (Fig. 2) were no less likely for IN heads, $\beta=_{.95}$(-0.63, 0.28). Comparing trials with ORC responses in Gap and ObjRP conditions (Fig. 3), ObjRPs cued rapid reduction in SRC looks in the following region, and this interpretation was slower when co-arguments matched in animacy, $\beta=_{.95}$(0.06, 0.72), consistent with the presence of similarity-based interference, and also trended slower when the *co-argument* was IN, $\beta=_{.95}$(-0.04, 1.06), consistent with a preference to take HU co-arguments as subjects.

Discussion The absence of animacy-based SRC predictions in SLZ is a problem for any universalist account of this behavior, despite existing cross-linguistic evidence [6]. Even a [2]-like account using experience-based biases would struggle to explain a subject bias for HU co-arguments, but not HU heads. Instead, we hypothesize that predictive dependency resolution as a whole is a learned behavior which is not motivated in SLZ RCs. Indeed, even in English, these predictions are not intrinsic to comprehension, but emerge over development [8], perhaps because SRC predictions can help avoid overlaps between lexical processing and dependency resolution. In contrast, in SLZ, even without predictions, either an RP will provide a dedicated cue to the dependency tail, or else a gap can be chosen flexibly later.

Although this hypothesis allows for substantial variation across languages, a prediction which needs much further testing and modeling, we see it as a promising idea which brings theories of sentence processing closer to theories of rational adaptive human behavior across other disciplines of cognitive science [7].

				Head	$_{RC}$**[**	**V**	(_$_?$)	**Co-Arg**	(_$_?$)	**…]**
(1)	Udanh	fotografia'nh	tse	bi'i xyage'nh		txube		coche'nh		
	touch	the.picture	of	the.boy		pull		the.car		

"Touch the picture of {**SRC** the boy who is pulling the car / **ORC** the boy who the car is pulling} …"

(2)	…	bi'i xyage'nh	txube	**=ba'**	coche'nh	(3)	…	bi'i xyage'nh	txube	coche'nh	**leba'**
		the.boy	pull	**=he**	the.car			the.boy	pull	the.car	**him**

"…the boy who (he) is pulling the car" (**SRC**) "…the boy who the car is pulling (him)" (**ORC**)

Dependency Type	Co-Argument Animacy			
N1 = HU	*Mismatch*		*Match*	
Argument Gap (Ambig.)	boy pull car	(HU V IN)	boy pull girl	(HU V HU)
Object RP	boy pull car him	(HU V IN RP)	boy pull girl him	(HU V HU RP)
N1 = IN	*Mismatch*		*Match*	
Argument Gap (Ambig.)	car pull boy	(IN V HU)	car pull truck	(IN V IN)
Object RP	car pull boy it	(IN V HU RP)	car pull truck it	(IN V IN RP)

Table 1: The eight conditions of one 2 x 2 x 2 item frame.

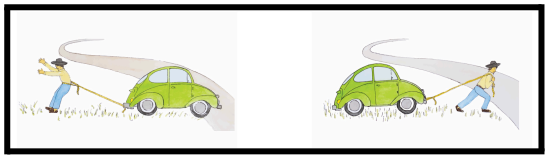


Figure 1: Example image choices for a Mismatch trial. For (1), L = ORC and R = SRC.

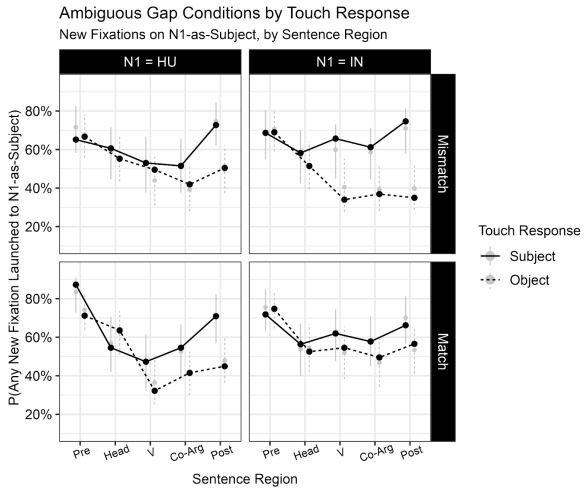


Figure 2: Gaze in ambiguous gap conditions.

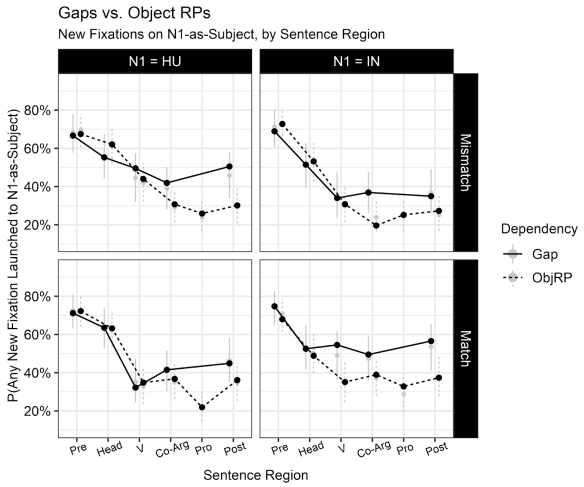


Figure 3: Comparing Gaps and ObjRPs.

Please see our anonymized OSF repository (link) for complete descriptions of methods, results, and analysis.

[1] Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition*, *30*, 73–105.
[2] Mitchell, D. C., & Cuetos, F. (1991). The origins of parsing strategies. In C. Smith (Ed.), *Current issues in natural language processing* (pp. 1–12). Center of Cognitive Science, University of Texas.
[3] Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1411–1423.
[4] Mak, W. M., Vonk, W., & Schriefers, H. (2002). The influence of animacy on relative clause processing. *Journal of Memory and Language*, *47*, 50–68.
[5] Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*, 69–90.
[6] Hammerly, C., Staub, A., & Dillon, B. (2022). Person-based prominence guides incremental interpretation: Evidence from obviation in Ojibwe. *Cognition*, *225*, 105122.
[7] Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*, 279–311.
[8] Atkinson, E., Wagers, M. W., Lidz, J., Phillips, C., & Omaki, A. (2018). Developing incrementality in filler-gap dependency processing. *Cognition*, *179*, 132–149.

Tell it like it (usually) is(n't): Speakers' mention of instruments in Reddit corpus reflects instrument (a)typicality

Kapron-King, A., Fletcher, L., Tarighat, A., Dobreva, R.,
Droop, S., Cummins, C., & Rohde, H. [University of Edinburgh]

If speakers use language to describe their world, one would expect to encounter language that describes the kinds of situations in which speakers are likely to find themselves, i.e., situations that are frequent and typical in the world (TELL IT LIKE IT IS). Prior work indeed affirms a role for real-world knowledge in people's expectations about what a speaker is going to say [1,2]. On the other hand, if one of the things that speakers do with language is to convey non-inferable information about particular situations, one would expect language that favors newsworthy content, particularly if speakers make rational decisions about the inclusion of syntactically optional elements (TELL IT LIKE IT ISN'T). Prior work on speakers' productions confirms a preference for mentioning the atypical: Event descriptions are more likely to include the instrument when it is atypical (e.g., an icepick rather than a knife for a stabbing [3]) and object descriptions are more likely to include a modifier when the property is atypical (e.g., blue rather than red strawberry [4]; wool rather than ceramic bowl [5]). Such work has primarily tested speakers' productions in psycholinguistic experiments. Here we use naturally occurring productions (extracted from social network forum www.reddit.com).

We target speakers' choice to talk about the presence of an instrument ("with") or its absence ("without"), see examples in Table 1. In order to compare the TELL IT LIKE IT IS versus -ISN'T accounts, we extract verb-object-instrument triplets that appear in Reddit comments. We test how commenters' choice to mention the presence or absence of an instrument varies with the typicality of the instrument (as indicated by a set of naïve participants). A rational strategy for instrument mention would predict "with" mentions for atypical instruments and "without" mentions for typical ones (TELLING IT LIKE IT ISN'T).

Methods. We automatically extracted mentions of instruments from 4 months of Reddit comments with templates for positive "with" and negative "without" instances, removing obscenities, brand names, and non-concrete nouns, yielding 9,621 cases. A manual filtering removed context-dependent instruments (*same water*) and cases where the noun was not enabling the action (*eat burger with cheese*), yielding 499 verb-object-instrument triplets. These triplets were presented in batches of 49-50 to Prolific participants (N=206, £1.25 payment) who gave typicality ratings on a scale of 1 to 11 (Fig 1). The resulting data consisted of 9826 typicality ratings after exclusions.

Results. Under the TELL IT LIKE IT IS account, we'd expect speakers to mention the presence of typical instruments and the absence of atypical ones. Instead, instruments in "with" mentions received only mid-range typicality ratings (mean rating 6.6) and those in "without" mentions received higher typicality ratings (mean 7.4). Fig 2 shows individual instruments' proportion of positive "with" mentions (out of all "with" and "without" mentions of that instrument). As can be seen, "with" mentions decrease as typicality increases. We used a logistic regression to model the binary outcome of each Reddit comment's polarity ("with" versus "without") with a fixed effect for typicality and a random effect for instrument. The results show a main effect of typicality (β: -0.180, SE=0.0707, p=0.01). Table 1 shows the pattern for the instrument *spoon* and an illustrative set of triplets about *teeth brushing*.

Our findings suggest speakers make rational choices to talk about uninferable aspects of a situation, i.e., the presence of atypical instruments or the absence of typical ones. We thus extend prior work on rational production, showing a pressure for the inclusion of informative content in the mention of instruments in naturally occurring language.

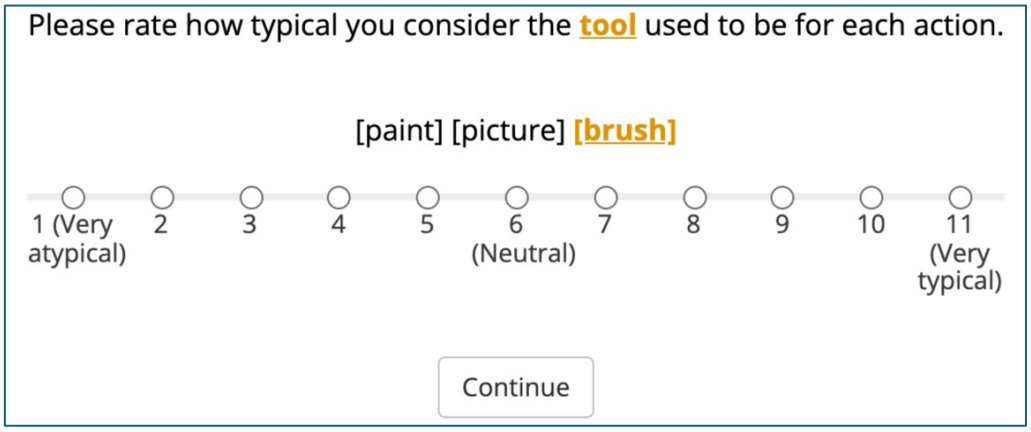


Figure 1: Task setup for elicitation of instrument typicality

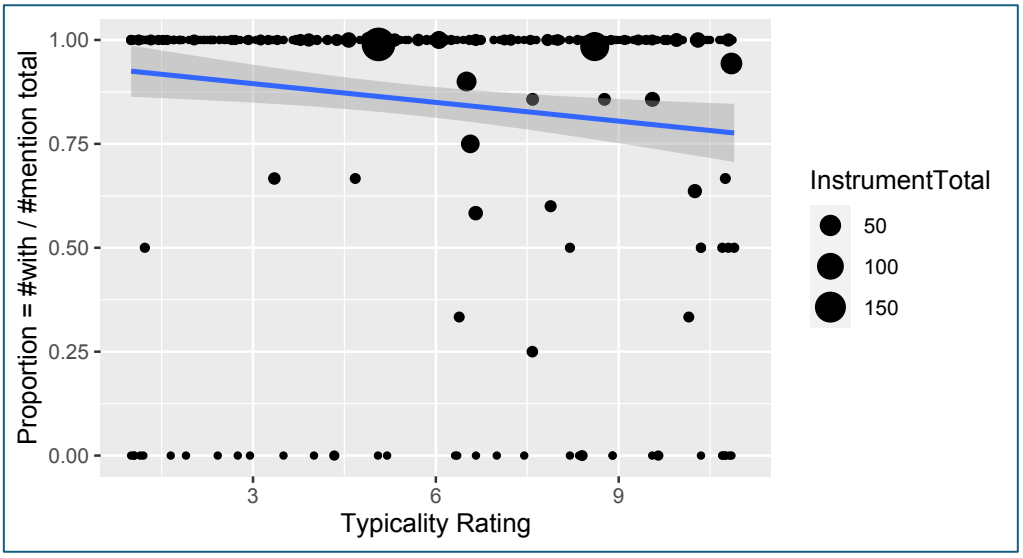


Figure 2: Mention of instruments using “with” vs “without” in Reddit corpus

verb-object-instrument triplet	Reddit probability	Typicality rating
eat chili spoon	p(“with”) = 0.0	9.6
eat spaghetti spoon	p(“with”) = 1.0	3.0
brush teeth toothpaste	p(“with”) = 0.4	10.3
brush teeth fingers	p(“with”) = 1.0	3.4
brush teeth wrong hand	p(“with”) = 1.0	3.3
brush teeth nail file	p(“with”) = 1.0	1.1
brush teeth toilet brush	p(“with”) = 1.0	1.0
brush teeth 12 gauge shotgun	p(“with”) = 1.0	1.0
brush teeth brick	p(“with”) = 1.0	1.0
brush teeth hairbrush	p(“with”) = 1.0	1.0

Table 1: Sample of Reddit usage (probability of “with”) and mean typicality rating (1-11 scale)

[1] Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language,* 49(1). 133–156. [2] Kutas, M. & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207. 203–205. [3] Brown, P. M. & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology* 19. 441–472. [4] Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research* 32. 3–23. [5] Mitchell, M., Reiter, R. & Van Deemter, K. (2013). Typicality and object reference. Proceedings of the Annual Meeting of the Cognitive Science Society 35. 3062–3067.

Investigating the Correlation Between Human Predictability Judgements and Computational Estimates from Text- and Audio-Based Models

Wei Xue, Julius Steuer, Dietrich Klakow, Bernd Möbius (Saarland University)

weixue@lst.uni-saarland.de

Previous research has shown that human language comprehension improves when an upcoming word is predictable within its context (Pickering and Garrod, 2007). Statistical language models (LM), trained to predict the next word in a sequence, offer probabilistic estimates of word predictability (de Varda et al., 2023). These estimates (e.g., surprisal) have been found to correlate well with human comprehension performance measures such as self-paced reading times (de Varda et al., 2023). In this study, we investigate whether the computational estimates from LMs and Automatic Speech Recognition (ASR) systems align with human judgments on the predictability of target words in sentence pairs. We focus on two estimates: surprisal and entropy. Surprisal reflects how unexpected a word is given its preceding context, while entropy quantifies the uncertainty in predicting the next word in a sequence.

To investigate the alignment, we first conducted a multiple-choice experiment where participants judged which context fit the target word best in paired sentences[1], resulting in binary predictability judgements for the target trigram of target words. An example of sentence pairs is shown in Table 1. We then compared the judgments to the surprisal and entropy estimates derived from the LM and ASR models. We hypothesized that a larger difference in estimates $\Delta_{\text{Estimates}}$ on the target word w given the two contexts correlate with the difference of preference in the human judgements. The difference in estimates is calculated following formula (1), where C_{easy} and C_{hard} refer to the contexts that make the target word easier or more difficult to predict, respectively. The difference of preference in the human judgements $\Delta_{\text{Preference}}$ is calculated following formula (2), where $P(w, C_{\bullet})$ represents the number of participants who judged the context to be more or less predictable. We then correlate $\Delta_{\text{Estimates}}$ and $\Delta_{\text{Preference}}$ over sentence pairs, namely trigrams.

Figure 1 shows a clear difference in surprisal and entropy estimates from LMs between predictable and unpredictable sentential context and type of trigrams. After excluding seven sentence pairs from the total thirty pairs for which there was no agreement in the human judgements (i.e., with $\Delta_{\text{Preference}}$ values smaller than 20), we found significant correlations (r = 0.50, p = 0.0152) of LM entropy from English translations of the stimuli with human judgments and of Dutch LM surprisal summed over whole sentences (r = 0.47, p = 0.022). During the conference, we aim to additionally present the correlation of surprisal and entropy with human predictability judgments in a cross-lingual setting. To this end, we would present native speakers of a language other than Dutch (i.e., German and English) with the Dutch stimuli and ask them to translate the target word. We hypothesize that lexical similarity affects the prediction of the target word if there is a high similarity between the Dutch context and a translated context, and therefore surprisal and entropy at the target word are low.

[1]We first selected 15 target words that are cognates in Germanic languages (i.e., Dutch, German, and English). Then we extracted one high-surprisal (i.e., only preposition phrases, PP) and one low-surprisal (i.e., only noun phrases; NP) trigram for each target word from trigram monolingual LMs trained on CGN (Schuurman et al., 2003), ukWaC, and deWaC (BARONI et al., 2009). Note that phrase type and trigram being high or low surprisal are tangled to ensure this setting is cross-lingual. For each trigram, we constructed two sentences where the target word is more predictable given the context in one sentence than the other, leading to four sentences per target word.

$$\Delta_{\text{Estimates}}(w, C_{\text{easy}}, C_{\text{hard}}) = |-\log_2 p(w|C_{\text{easy}}) + \log_2 p(w|C_{\text{hard}})| \quad (1)$$

$$\Delta_{\text{Preference}}(w, C_{\text{easy}}, C_{\text{hard}}) = \frac{|P(w, C_{\text{easy}}) - P(w, C_{\text{hard}})|}{P(w, C_{\text{easy}}) + P(w, C_{\text{hard}})} \quad (2)$$

Predictability	Trigram Surprisal	Sentence
low	high	De jongen raakte de bal met de **arm**. (English translation "The boy touched the ball with the **arm**.")
high	high	Hij maakte een mooie beweging met de **arm**. (English translation: "He made a nice movement with the **arm**.")
low	low	Hij masseerde zachtjes zijn andere **arm**. (English translation "He gently massaged his other **arm**.")
high	low	Ze toonde trots zijn andere **arm**. (English translation: "She proudly showed his other **arm**.")

Table 1: An example of a paired sentence given a selected trigram of the target word.

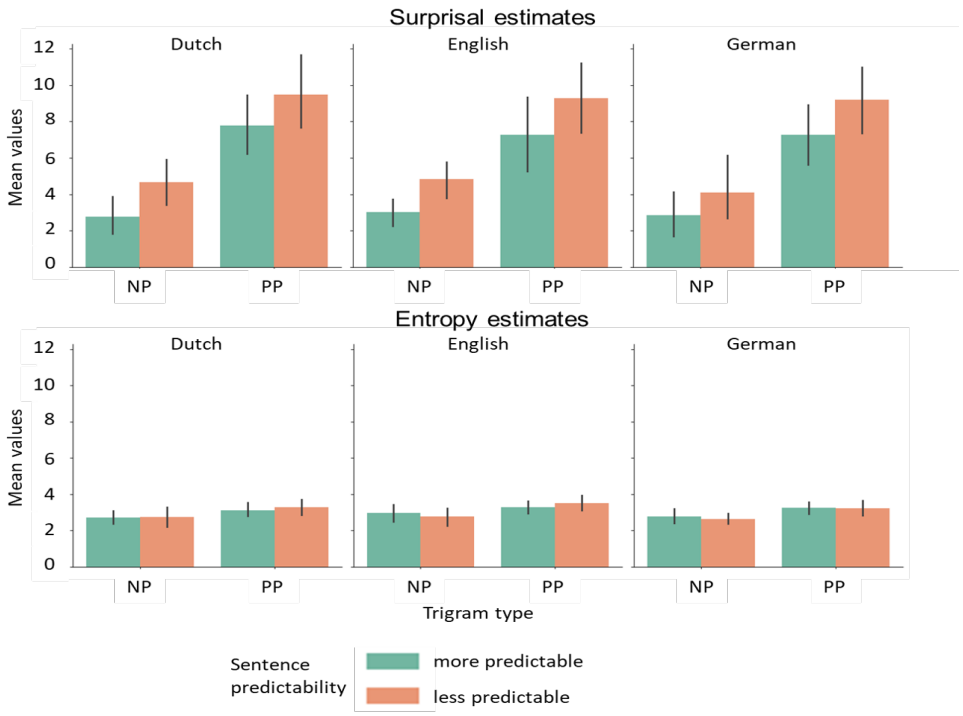


Figure 1: Mean surprisal and entropy estimates from Dutch, English, and German LMs.

References

BARONI, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43.3*, 209–226.

de Varda, A. G., Marelli, M., & Amenta, S. (2023). Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing eeg and reading time data. *Behavior Research Methods*, 1–24.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *rends in cognitive sciences*, *11*(3), 105–110.

Schuurman, I., Schouppe, M., Hoekstra, H., & van der Wouden, T. (2003). CGN, an annotated corpus of spoken Dutch. *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*. https://aclanthology.org/W03-2414

Paradigmatic Variability of Multi-word Expressions in Scientific English

Diego Alves (Saarland University), Stefan Fischer (Saarland University), Elke Teich (Saarland University)

diego.alves@uni-saarland.de

In this study, we analyze the paradigmatic variability (i.e., the sets of linguistic options available in a given or similar syntagmatic contexts) of different categories of multi-word expressions (MWEs) in the domain of scientific writing, inspecting diachronic changes from the mid-17th century to today. MWEs are sequences of words perceived either as wholes or with highly predictable transitions from one word to the next. Their use in scientific writing is particularly interesting because MWEs contribute to smoothing the information load over a message (Conklin & Schmitt, 2012). Teich et al. (2021), using embedding spaces and entropy measures to estimate paradigmatic variability, observed a reduction in this dimension for different parts-of-speech, indicating a continuous, diachronic process of conventionalization that serves to manage linguistic variability in the interest of cognitive resource efficiency. Our hypothesis is that different categories of MWEs present lower paradigmatic variability due to their semantic characteristics compared to analogous expressions, thus, contributing even more to conventionalization.

To test this hypothesis, we first extracted and classified the MWEs from an extensive diachronic dataset of English scientific texts, the Royal Society Corpus (RSC) into six categories following the work proposed by Alves et al. (2024): (1) compounds, composed of sequence of nouns (e.g., *orange juice, sea salt*); (2) flat, sequences of proper nouns and names of places and institutions (e.g., *Isaac Newton*); (3) phrasal verbs (e.g., *carry out*); (4) fixed, used for certain fixed grammaticalized expressions which tend to behave like function words (e.g., *in spite of*); (5) academic formulaic expressions, list of MWEs from the Academic Formulas List (Simpson-Vlach & Ellis, 2010) (e.g., *on the other hand, a kind of*); and (6) miscellaneous MWEs extracted from the RSC using the Partitioner tool (Williams, 2016) (e.g., at first sight, give rise). Then, we processed the RSC texts by connecting the tokens belonging to MWEs and proceeded with the calculation of the embedding space using structured skip-grams. The paradigmatic variability of a word over time was calculated following the method introduced by Teich et al., (2021), which defines it as the entropy over a probability distribution, based on the probability of a word from a specific neighbourhood being chosen instead of the other words in the same area.

Figure 1a shows that up to 1940, compounds have lower paradigmatic variability than nouns, with the same decreasing tendency, and flat MWEs present lower values than proper nouns, however, with peaks in 1810 and 1820. In Figure 1b, it is possible to notice that although phrasal verbs start with a higher paradigmatic variability when compared to other verbs, from 1750 on, the inverse is observed, with phrasal verbs presenting a considerable decreasing tendency regarding paradigmatic variability in the twentieth century. As shown in Figure 2a, academic formulaic expressions and fixed MWEs present a quite stable paradigmatic variability in time, with lower values when compared to adverbs and function words. Finally, Figure 2b shows that the other MWEs category presents similar behavior to function words and adverbs. Thus, overall, we can conclude that the conventionalization process throughout time regarding the lexicon in the scientific domain is even more evident when MWEs are considered as whole units.

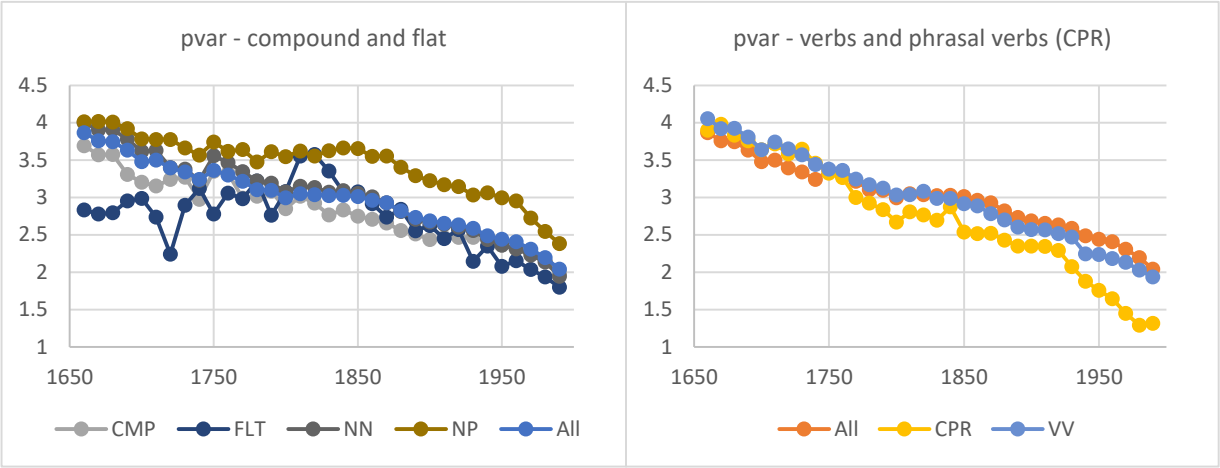


Figure 1. Paradigmatic variation per decade of: a) Compounds (CMP), Flat MWEs (FLT), Nouns (NN), Proper Nouns (NP), and All words in the embeddings space; and b) Phrasal Verbs (CPR), other verbs (VV), and All words in the embeddings space.

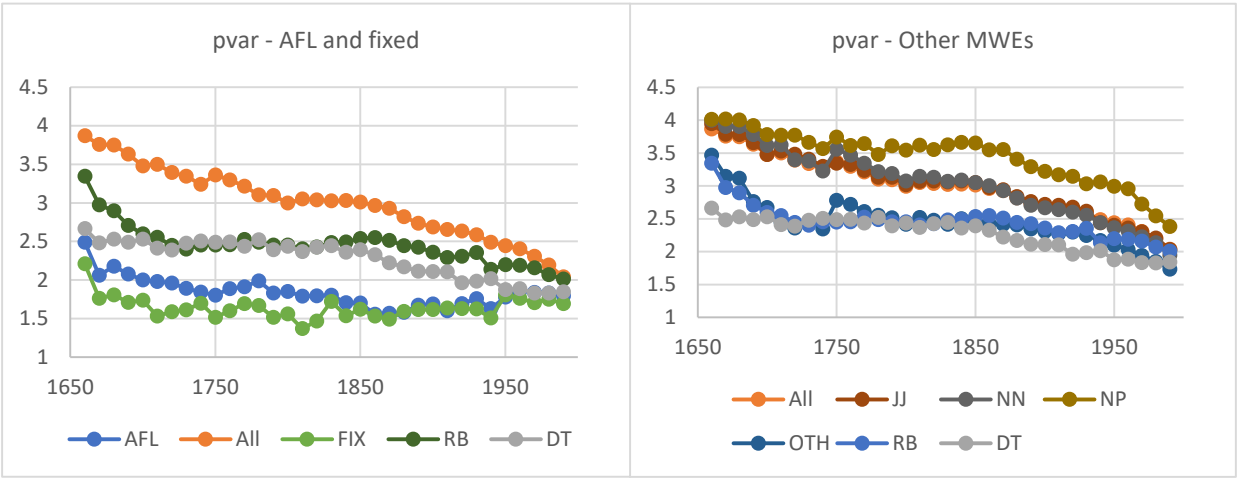


Figure 2. Paradigmatic variation per decade of: a) Academic formulaic expressions (AFL), fixed MWEs (FIX), Adverbs (RB), function words (DT), and All words in the embeddings space; and b) other MWEs (OTH), Adjectives (JJ), Nouns (NN), Proper Nouns (NP), Adverbs (RB), function words (DT), and All words in the embeddings space.

References:

Alves, D., Fischer, S., Degaetano-Ortlieb, S., & Teich, E. (2024, March). Multi-word expressions in English scientific writing. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)* (pp. 67-76).

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual review of applied linguistics*, *32*, 45-61.

Teich, E., Fankhauser, P., Degaetano-Ortlieb, S., & Bizzoni, Y. (2021). Less is more/more diverse: on the communicative utility of linguistic conventionalization. *Frontiers in Communication*, *5*, 620275.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied linguistics*, *31*(4), 487-512.

Williams, J. R. (2016). Boundary-based MWE segmentation with text partitioning. *arXiv preprint arXiv:1608.02025*.

Analyzing the Effects of Temperature-Scaled Surprisal for Subword Reading Times

Sneha Chetani[1], **Iza Škrjanec**[1], **Vera Demberg**[1]
[1]Saarland University, Germany
snehachetani45@gmail.com, {skrjanec, vera}@coli.uni-saarland.de

Surprisal [1, 2] measures predictability in context and has been accepted as a metric of per-word human processing effort [3-5] with surprisal estimated using large neural language models (LMs). Recent work indicates LMs with a higher quality (and a lower perplexity) do not necessarily correspond to a better fit to human reading times (RT) [6] and that they likely underestimate the surprisal of low-frequency words [7]. A way of adjusting for this is to use temperature scaling of LM outputs to make the probability distribution less certain. Liu et al. [8] show that surprisal estimates are closer to human reading times when surprisal is temperature-scaled. Additionally, they show this benefit is driven by words that are split into multiple subwords. This poses the question whether LMs are overestimating the processing difficulty due to subword tokenization and how this overestimation explains why temperature scaling differentially impacts RT of standalone versus split subwords.

In our study, we model reading times during naturalistic reading and calculate surprisal with the small GPT2 LM. We use the Dundee eye-tracking corpus [9], but instead of using word-level gaze duration measures, we re-calculate the measures for each subword based on the GPT2 tokenizer. We consider the log-transformed total reading time of each subword and fit a baseline mixed-effects regression with subword surprisal and length, word and subword frequency, position as predictors, including a binary split-indicator of whether a subword stands alone or is rather a part of a word. The experimental model included surprisal from GPT2, which has been temperature-scaled. We explore the range between 1 and 10 for scalar values, where a larger value results in a less certain probability distribution over the vocabulary and a higher surprisal for most words. To establish the effect of temperature scaling, we compare the log-likelihoods of the base and experimental model in the delta log-likelihood metrics, where a larger value indicates a better fit above the baseline.

Our results show that the split-indicator has a main effect above and beyond the length and frequency of the subword: words that are split are read more slowly. Our results also reveal that the effect of temperature scaling is not equally beneficial for all subwords. As shown in Figure 1, delta log-likelihood values indicate that single-subword words and the first subword of a split word profit, as their predicted reading times are closer to observed times after their surprisal is increased via temperature scaling. This contrasts the result for subwords that are in the middle or at the end of a word (e.g. *can**'t***, *four-year**ly***). We suspect that these are high-predictability continuations of the first subword. Increasing their surprisal does not correspond to human processing effort.

We plan to extend the analysis to synthetic languages (such as German or Finnish) with longer compound words and richer inflection. The analyses will add to the discussion of cognitive plausibility of subword tokenizers [10, 11].

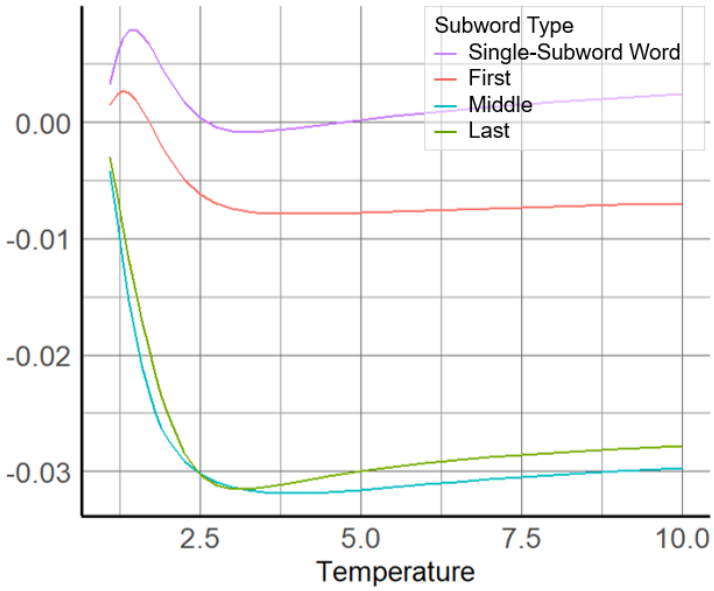


Figure 1: Effects of temperature scaling on delta log-likelihood values for different subword types across temperatures $T \in [1, 10]$.

References [1] Hale. Probabilistic Earley parser as a psycholinguistic model. NACL 2001. • [2] Levy. Expectation-based syntactic comprehension. Cognition 2008. • [3] Demberg, Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition 2008. • [4] Smith, Levy. The effect of word predictability on reading time is logarithmic. Cognition 2013. • [5] Wilcox, Gauthier, Hu, Qian, Levy. On the predictive power of neural language models for human real-time comprehension behavior. CogSci 2020. • [6] Oh, Schuler. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?. ACL 2023. • [7] Oh, Yue, Schuler. Frequency Explains the Inverse Correlation of Large Language Models' Size, Training Data Amount, and Surprisal's Fit to Reading Times. EACL 2024. • [8] Liu, Škrjanec, Demberg. 2024. Temperature-scaling surprisal estimates improve fit to human reading times – but does it do so for the "right reasons"?. ACL 2024. • [9] Kennedy, Hill, Pynte. The Dundee corpus. European conference on eye movement 2003. • [10] Beinborn, Pinter. Analyzing Cognitive Plausibility of Subword Tokenization. EMNLP 2023. • [11] Nair, Resnik. Words, Subwords, and Morphemes: What Really Matters in the Surprisal-Reading Time Relationship?. EMNLP 2023.

Joint inference about pragmatically-relevant contextual features

Chris Cummins and Hannah Rohde (University of Edinburgh)
c.r.cummins@gmail.com

In traditional approaches to pragmatics, inferences are taken to be underpinned by rich assumptions about the speaker and the context of utterance. For instance, on a Gricean account, quantity implicature (e.g. interpreting *some* as also conveying 'not all') depends on the stronger alternative being relevant to the current conversational needs, and the speaker being knowledgeable about the stronger proposition as well as broadly cooperative (in the sense of adhering to the Cooperative Principle; Grice 1989). Experimental work has documented how the availability of quantity implicature is modulated by these factors (Breheny et al. 2006; Goodman and Stuhlmüller 2013), in addition to other considerations such as whether the stronger proposition would be impolite or face-threatening to assert (Bonnefon et al. 2009).

However, while experimental research has typically proceeded by manipulating the above factors and exploring the effect on pragmatic inference, real-life interaction is more complicated: hearers typically lack prior information about the speaker's knowledge state, cooperativity, and so on. Rather, the content of the utterance may itself inform the hearer's understanding of these factors; and their understanding of these factors rationally should inform their pragmatic interpretation of the utterance.

On this view, rational pragmatic interpretation involves joint inference about the state of the speaker and of relevant contextual features as well as the state of the world given the utterance. Some progress has been made in examining joint inference processes in pragmatics (e.g. Kao et al. 2014 on identifying non-literal intention), but we argue that such processes are much more widespread and consequential than is typically acknowledged. Crucially, hearers typically lack certainty about multiple factors which bear on the speaker's utterance choice, in which case truly rational pragmatic interpretation involves evaluating an array of competing explanations for the utterance, and theories have yet to specify how a hearer might do this. Moreover, most research in this area has proceeded under the assumption that the speaker is fully cooperative, which is in practice atypical of human interaction and has been argued not to be essential for rich pragmatic inference (Asher and Lascarides 2013).

In this presentation, we outline a model of pragmatic joint inference which can encompass the full range of relevant factors, treating them as variables about which hearers have probabilistic beliefs which may receive Bayesian updates. We briefly discuss how this model allows us to draw new insights from existing experimental data in three domains: scalar diversity in quantity implicature (van Tiel et al. 2016), reference assignment for ambiguous singular 'they' (Arnold et al. 2021), and modified numerical expressions (Hesse and Benz 2020). In each case, we will argue that inferences about the speaker's knowledge state, cooperativity, and social disposition can and do bear upon interpretation.

We conclude by briefly discussing how specific novel predictions can be drawn from such a model and tested empirically, and how this could help us evaluate claims about the architecture of human pragmatic processing and the extent of rationality in pragmatic interpretation.

References

Arnold, J. E., Mayo, H. C., & Dong, L. (2021). My pronouns are they/them: Talking about pronouns changes how pronouns are understood. *Psychonomic Bulletin & Review*, *28*(5), 1688–1697.

Asher, N., & Lascarides, A. (2013). Strategic communication. *Semantics & Pragmatics*, 6(2), 1–62. https://doi.org/10.3765/sp.6.2

Bonnefon, J.-F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: scalar inferences in face-threatening contexts. *Cognition*, 112(2), 249–258.

Breheny, R., Katsos, N., and Williams, J. (2006). Are Generalised Scalar Implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100(3), 434–463.

Goodman, N.D., & Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science*, 5, 173–184.

Grice, H.P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Hesse, C., & Benz, A. (2020). Scalar bounds and expected values of comparatively modified numerals. *Journal of Memory and Language*, 111:104068. https://doi.org/10.1016/j.jml.2019.104068

Kao, J.T., Wu, J.Y., Bergen, L., & Goodman, N.D. (2014). Nonliteral understanding of number words. *PNAS*, 111(33), 12002–12007.

Van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, 33, 137–175.

Lossy Context Surprisal: Influence of Linearization on Expectation

Xinyue Jia, Christoph Aurnhammer, Torsten Kai Jachmann, Francesca Delogu, Heiner Drenhaus, Matthew W. Crocker
Language Science and Technology, Saarland University, Germany

The current SPR study investigates whether longer-distance dependencies exhibit attenuated surprisal effects due to an imperfect representation of context in memory, such that the processing of expected words is less facilitated, while unexpected words are less effortful, when the previous context is partially lost due to distance.

Expectation-based sentence processing is supported by a range of behavioral and neurophysiological evidence suggesting that additional input facilitates processing by reducing uncertainty about upcoming words (e.g., Hale, 2001, 2006; Levy, 2008). Studies on long-range dependencies, however, show that increasing the distance between dependent elements of a sentence increases processing difficulty, contradicting expectation-based accounts (for discussion see Futrell et al., 2020). By contrast, memory-based models have attributed such behaviour to working memory limitations (Lewis & Vasishth, 2005). Recent accounts, such as the lossy-context surprisal model, however, incorporate memory effects into an expectation-based framework by formally characterizing how surprisal is determined based on imperfect, or lossy, representations of the preceding context (Futrell et al., 2020; Hahn et al., 2022).

To investigate the interaction of expectation and memory, we conducted a reading time study in German using a 2×2 design based on materials adapted from Aurnhammer et al. (2021). We created 120 items which varied the linear position of an adverbial clause (*bevor der Holzfäller … stapelte*) to manipulate the distance (Long vs Short) between the main verb (*schärfte/aß*) and the object (*Axt*). The expectancy of the object (Expected vs Unexpected) is manipulated by the main verb in the preceding context (*schärfte die Axt vs. aß die Axt*, see Table 1). Cloze and plausibility pretests confirmed the differences in expectancy, but were unaffected by the linearization. In addition to main effects of expectation and distance, the lossy surprisal account crucially predicts an interaction: If the increased distance of the B&D conditions results in a lossy memory representation of the predictive context – namely the main verb (*schärfte/aß*) – surprisal effects are predicted to be attenuated compared to the short distance conditions (A&C). That is, we predict an interaction of expectancy and distance, such that stronger surprisal effects occur in the short distance conditions, and weaker effects in the long distance conditions. In a preliminary analysis of the reading times (N=68), we find precisely the predicted interaction in both the spill-over (*und*) and post-spillover regions (*hackte*), as well as the predicted main effect of expectancy in the post-spillover region (Table 2 and Table 3). We interpret these results as providing both clear support for the predictions of the lossy surprisal account, and broader evidence that linearization decisions influence the online processing effort of alternative encodings beyond offline predictors such as cloze and plausibility.

Conditions	
A Expected Short	Bevor der Holzfäller in den Wald ging und das Holz stapelte, schärfte er die **Axt** und hackte… *(Before the lumberjack in the forest went and the wood stacked, sharpened he the axe and chopped…)*
B Expected Long	Der Holzfäller schärfte, bevor er in den Wald ging und das Holz stapelte, die **Axt** und hackte… *(The lumberjack sharpened, before he in the forest went and the wood stacked, the axe and chopped…)*
C Unexpected Short	Bevor der Holzfäller in den Wald ging und das Holz stapelte, aß er die **Axt** und hackte… *(Before the lumberjack in the forest went and the wood stacked, ate he the axe and chopped…)*
D Unexpected Long	Der Holzfäller aß, bevor er in den Wald ging und das Holz stapelte, die **Axt** und hackte… *(The lumberjack ate, before he in the forest went and the wood stacked, the axe and chopped…)*

Table1. Example of Stimuli.

	Spillover			Post-Spillover		
Fixed effects	$\hat{\beta}$	*SE*	*p*	$\hat{\beta}$	*SE*	*p*
(Intercept)	5.6982	0.0253	0.0000	5.6629	0.0243	0.0000
Expectancy	-0.0021	0.0016	0.2057	**-0.0049**	**0.0017**	**0.0055**
Distance	-0.0003	0.0005	0.5834	0.0005	0.0005	0.3787
Expectancy:Distance	**0.0003**	**0.0001**	**0.0010**	**0.0003**	**0.0001**	**0.0040**

Table2. Summary of the linear mixed effects model of reading times.

Reading time (ms)	Spillover		Post-spillover	
	Short	Long	Short	Long
Expected	*M:* 303.14, *SE*: 1.81	*M:* 302.57 *SE:* 1.81	*M:* 288.59, *SE:* 1.75	*M:* 291.82, *SE:* 1.75
Unexpected	*M:* 308.59, *SE*: 1.93	*M:* 301.22, *SE:* 1.81	*M:* 296.79, *SE:* 1.84	*M:* 294.75, *SE:* 1.82

Table3. Descriptive statistics of reading times.

Selected References

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PloS* one, 16(9), e0257430.
Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3), Article e12814.
Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences of the United States of America*, 119(43), e2122602119.
Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. *North American Chapter of the Association for Computational Linguistics*.
Hale, J. T. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9), 397–412.
Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, 29(3), 375–419.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.

Probabilistic Inference and Frequency Effects in Language Change
Vsevolod Kapatsinski (University of Oregon)
vkapatsi@uoregon.edu

Language change often involves the gradual encroachment of an innovative variant form on the contexts previously occupied by another form. For example, in American English, *going to* is encroaching on *will* in the context of future marking, the pronunciation *-in'* is encroaching on the contexts that used to favor *-ing*, and the flap [ɾ] has encroached on [t] in post-tonic intervocalic contexts.

Much research has shown that words that often occur in contexts that favor the innovative variant become associated with that variant, so that the innovative variant becomes likely to be used with such words even outside of contexts that otherwise favor it (Bybee, 2002; Brown, 2004; Forrest, 2017; *inter alia*). For example, the [n] at the end of *-ing* is favored by informal speech style, and following coronal consonants. However, words that frequently occur in informal styles and before coronal consonants favor the [n] pronunciation even when they are used in a formal style and before a vowel. Following Brown, this is usually called the frequency in favorable contexts (FFC) effect. FFC effects are found to be stronger in frequent words (Forrest, 2017). The present work studies the conditions under which FFC and its interaction with frequency emerge from rational probabilistic inference.

We make use of a recently developed model of sound change in which online reductive pressures are combined with rational probabilistic hierarchical inference, which distributes credit for a pronunciation between the sublexical unit undergoing the change (the linguistic variable), and the larger lexical units (words and phrases) that contain it (Kapatsinski, 2021, in press). The model implements cycles of production and learning across generations. Production is assumed to have a bias in favor of the innovative variant, such that every token of a word's use increases the probability of selecting the innovative variant. The resulting productions constitute a corpus from which the next generation learns the language. Each learner is assumed to infer a hierarchical logistic regression model in which words, phrases and sublexical units are nested random effects (see also Vetchinnikova, 2024), alongside fixed effects of context and grammar. I report on an extension of the model in which 1) the production pressure for the innovative variant is stronger in innovation-favoring contexts and may be reversed in innovation-disfavoring contexts, 2) each word has a probability of occurring in a reduction-favoring context, sampled from a beta distribution, and 3) learners may not detect the relevant context in a particular token. As in the original model, word frequencies are Zipfian-distributed.

The FFC is shown to emerge only when learners misattribute (some of) the effect of context to lexical idiosyncrasy. That is, it emerges only if learners are prone to missing the context that an observed word token occurs in (faulty perception), or do not take the influence of context into account when building their mental model of variant choice (imperfect learning). The model also provides an interesting direction for future work on when FFC effect do and do not emerge. Specifically, the distributions of words across innovative and conservative contexts have yet to be studied. Yet, the model suggests that these distributions are crucial because the FFC effect emerges only when words have

polarized distributions across reductionfavoring and disfavoring contexts (top row of Figure 1 vs. bottom row).

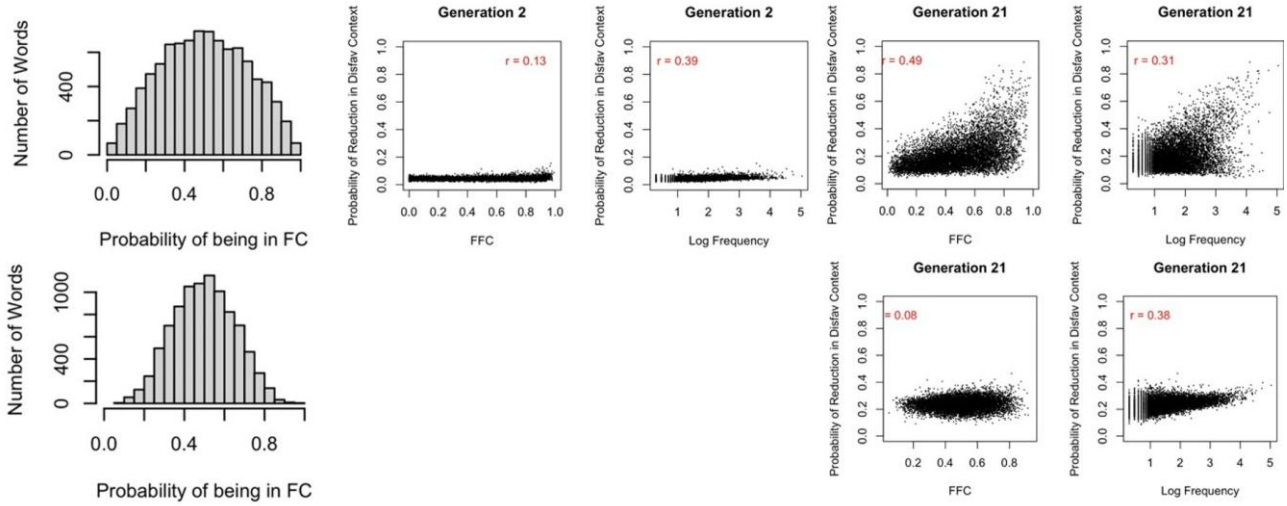


Figure 1. The effect of frequency in favorable context (FFC) emerges over the course of the sound change (by Generation 21) in the top row where the words' probabilities of being in a favorable context are variable enough. Starting from the same initial point (Generation 2), it does not emerge in the bottom row, where the words' distributions across contexts are less variable.

References:

Brown, E. L. (2004). The reduction of syllable-initial/s/in the Spanish of New Mexico and southern Colorado: A usagebased approach. Ph.D. Dissertation, University of New Mexico.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*(3), 261-290.

Forrest, J. (2017). The dynamic interaction between lexical and contextual frequency: A case study of (ING). *Language Variation and Change*, *29*(2), 129-156.

Kapatsinski, V. (2021). Hierarchical inference in sound change: Words, sounds, and frequency of use. *Frontiers in Psychology*, *12*, 652664.

Kapatsinski, V. (In press). Lexical frequency and diffusion. In A. Ledgeway, E. Aldridge, A. Breitbarth, K. É. Kiss, J. Salmons andA. Simonenko (Eds.), *The Wiley Blackwell Companion to Diachronic Linguistics*. Wiley.

Vetchinnikova, S. (2024). Idiosyncratic entrenchment: tracing change in constructional schematicity with nested random effects. *Corpus Linguistics and Linguistic Theory*.

Lossy Context Surprisal Predicts Task Differences in Relative Clause Processing

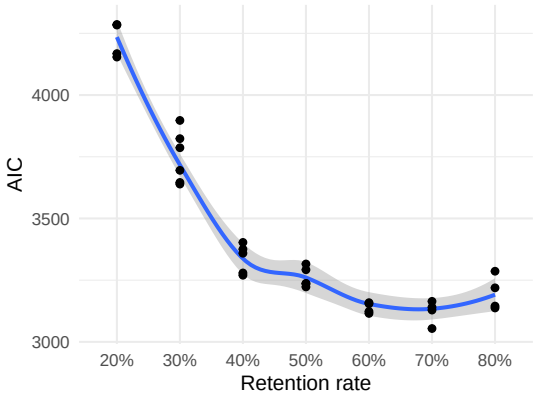
A fundamental goal of computational psycholinguistics is to predict and explain syntactic processing difficulty as manifested in reading times. English comprehenders take longer to read object relative clauses (ORCs), such as “The director that the dancer admired,” compared to equal-length subject relative clauses (SRCs), such as “The director that admired the dancer.” When do readers slow down, and why?

Expectation-based accounts (e.g. surprisal theory; Levy, 2008) predict that readers will slow down at the ORC noun phrase “the dancer.” SRCs are more frequent than ORCs (Roland et al., 2007); therefore, on seeing “The director that,” readers will expect a subject relative verb to follow. Vani et al. (2021) found that participants in a Maze task (Forster et al., 2009) showed the predicted slowdown at the ORC determiner “the.”

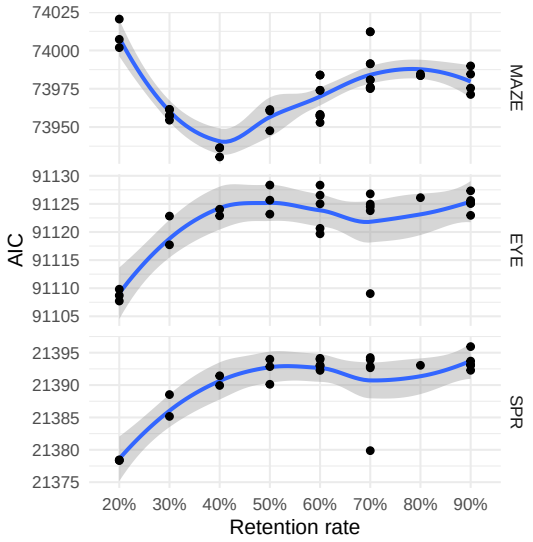
By contrast, memory-based accounts (e.g. Dependency Locality Theory; Gibson et al., 2000) predict that the ORC slowdown should instead appear at the verb “admired,” as readers integrate the dependency to the distant object “director.” This behavioral pattern has been reported in eye-tracking studies (Staub, 2010; Roland et al., 2021).

We argue that these discrepant empirical findings can be explained as task effects: the Maze task imposes higher memory demands, so readers systematically retain more of the preceding sentence context in Maze experiments compared to eye-tracking while reading. We support this account with computational evidence from the Resource-Rational Lossy Context Surprisal model (LCS; Hahn et al., 2022), which conceptually unifies expectation- and memory-based accounts.

We find that manipulating the LCS retention rate captures task-dependent differences observed in reading times (RTs) across experiments. Filler item RTs from the Maze task are best fit with a relatively high retention rate (e.g. 60%; Figure 1a), while lower retention (20%) better predicts eye-tracking RTs (Figure 1b). Using these task-dependent retention rates, LCS correctly predicts critical RT patterns observed for English relative clauses. In particular, low-retention (20%) LCS follows memory-based theories and predicts higher RTs for object relative verbs — an effect found in eye-tracking but not Maze studies (Figure 2). These results can explain the apparently contradictory behavioral evidence supporting both memory- and expectation-driven accounts: relative clause processing is likely modulated by the memory demands of the task, and we can model this phenomenon using Lossy Context Surprisal.



(a) Linear mixed-effects model fit for LCS to Maze RT data on filler items (Vani et al., 2021). Points are individual LCS model instances, line shows GAM smooth, x-axis shows retention rate, y-axis shows goodness of fit in AIC. Retention rate 60–70% achieves the best fit on average.



(b) Linear mixed-effects model fit for LCS to Maze (Hahn et al., 2022), eye-tracking (ET), and self-paced reading (SPR) data for filler items from Vasishth et al. (2010). Points are individual LCS model instances, line shows GAM smooth, x-axis shows retention rate, y-axis shows goodness of fit in AIC — lower is better. Maze data are better approximated by LCS with a higher retention rate (40%) compared to ET and SPR data (20%).

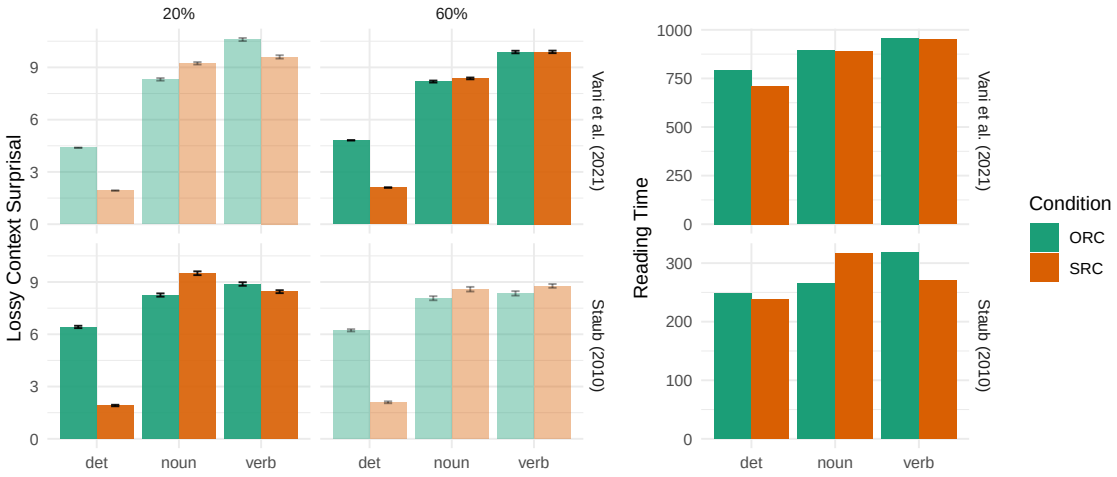


Figure 2. LCS predictions (left; error bars show standard error across model instances and items) and reading time data (right) for stimuli from Staub (2010, ET gaze duration, Experiment 1) and Vani et al. (2021, Maze, Experiment 1; cf. their Figs. 3 and 4). At the higher retention rate (60%), LCS predicts only the determiner slowdown observed in Maze data (top row). At the lower retention rate (20%), LCS also predicts the ORC verb slowdown observed in ET data (bottom row).

References

Forster, K. I., Guerrera, C., & Elliot, L. (2009, February). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, *41*(1), 163–171. Retrieved 2024-06-19, from https://doi.org/10.3758/BRM.41.1.163 doi: 10.3758/BRM.41.1.163

Gibson, E., et al. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, *2000*, 95–126.

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022, October). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), e2122602119. Retrieved 2023-04-23, from https://www.pnas.org/doi/10.1073/pnas.2122602119 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2122602119

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Roland, D., Dick, F., & Elman, J. L. (2007, October). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*(3), 348–379. Retrieved 2024-06-19, from https://www.sciencedirect.com/science/article/pii/S0749596X07000307 doi: 10.1016/j.jml.2007.03.002

Roland, D., Mauner, G., & Hirose, Y. (2021, August). The processing of pronominal relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *119*, 104244. Retrieved 2024-06-19, from https://www.sciencedirect.com/science/article/pii/S0749596X21000279 doi: 10.1016/j.jml.2021.104244

Staub, A. (2010, July). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*(1), 71–86. Retrieved 2024-04-19, from https://www.sciencedirect.com/science/article/pii/S001002771000082X doi: 10.1016/j.cognition.2010.04.002

Vani, P., Wilcox, E. G., & Levy, R. (2021). Using the Interpolated Maze Task to Assess Incremental Processing in English Relative Clauses. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43). Retrieved 2024-04-28, from https://escholarship.org/uc/item/3x34x7dz

Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, *25*(4), 533–567.

Testing a Rational Account of Fragment Usage with Crowd-sourced Production Data

Robin Lemke (Saarland University) robin.lemke@uni-saarland.de

Game-theoretic models have been applied to a range of pragmatic phenomena (Franke, 2009; Frank and Goodman, 2012), but their predictions have been tested at restricted and balanced sets of meanings. At the example of ellipsis, I test a rational account with a much more diverse and unbalanced data set collected with a crowd-sourced production task. I focus on fragments (1a) (Morgan, 1973), nonsentential utterances which are meaning-equivalent to sentences (1b) in an appropriate context. Previous research focused on the syntax of fragments, but why speakers actually use them is underexplored.

(1) [Passenger to conductor before entering the train:]

a. To Paris? b. Does this train go to Paris?

Account I hypothesize that speakers trade off the lower production cost for fragments (compared to sentences) with the risk of being misunderstood ((1a) could also mean *How long does it take to travel to Paris?*) and prefer fragments when the former outweighs the latter. To formally model this idea, following Franke (2009), I assume that the speaker sends a message $m \in M$ to the listener and selects the an utterance $u \in U$ to do so. The listener infers the meaning of u and if speaker and listener coordinate, both receive a reward. Therefore, the listener goes for the most likely interpretation (maximize $p(m|u)$), calculated as shown in equation 1. Sentences are unambiguous, but their cost is higher, so the speaker will prefer fragments when $p(m|u)$ is relatively high.

Method I evaluate the model with 3 pseudo-interactive utterance selection experiments. In each study, 60 subjects read a context story ($n = 15$) and select one out of 6 utterances to communicate one out of 3 messages (Fig. 1). The materials are based on a corpus of production data by Lemke (2021), from which M, U and the prior over messages $Pr(M)$ were estimated. The listener is simulated according to model predictions. In each trial, there is a fragment ambiguous between two messages: the *target* having a higher $p(m|u)$ than the *competitor*. There are 3 experimental conditions, which differ in whether the target, the competitor, or the third message (which is *unambiguous*ly encoded by the second fragment) is to be communicated. Utterances cost virtual coins and sentences are more expensive than fragments. Given the increasing $p(m|u)$ across the conditions (see Table 1), subjects should use fragments most frequently in the *unambiguous* condition than in the *target* and least often in the *competitor* condition.

Experiments and results Fig. 2 summarizes the data, which were analyzed with mixed effects logistic regressions (Bates et al., 2015). In exp. 1, $p(m|u)$ increased fragment ratio ($\chi^2 = 6.13, p < .05$), but some subjects produced only sentences, which yielded a net benefit given the cost structure. Therefore, in exp. 2 sentences were more costly, which increased fragment ratio further ($\chi^2 = 6.24, p < 0.05$). However, in exp. 1 and 2 there was no significant effect of $p(m|u)$ in the ambiguous conditions, so the effects found could be either evidence rational reasoning given the higher $p(m|u)$ in the unambiguous condition or that subjects just avoid ambiguity. Exp. 3 tested this by replacing the ambigous *competitor* condition by a further unambiguous one. This increases fragment ratio ($\chi^2 = 17.52, p < 0.001$), but the effect of $p(m|u)$ was replicated, too. Taken together, this supports the expected cost-accuracy tradeoff and shows that game-theoretic reasoning can also be applied to an unbalanced and diverse data set based on production data.

$$L_0(m,u) = \frac{Pr(m) \times [[u]]_m}{\sum_{m'} Pr(m') \times [[u]]_{m'}} \quad (1)$$

Condition	**Lowest** $p(m\|u)$	**Highest** $p(m\|u)$	**Mean** $p(m\|u)$
critical	0.12	0.69	0.36
competitor	0.03	0.17	0.08
unambiguous	0.15	1.0	0.76

Table 1 Range of $L_0(m|u)$ probabilities and means by conditions

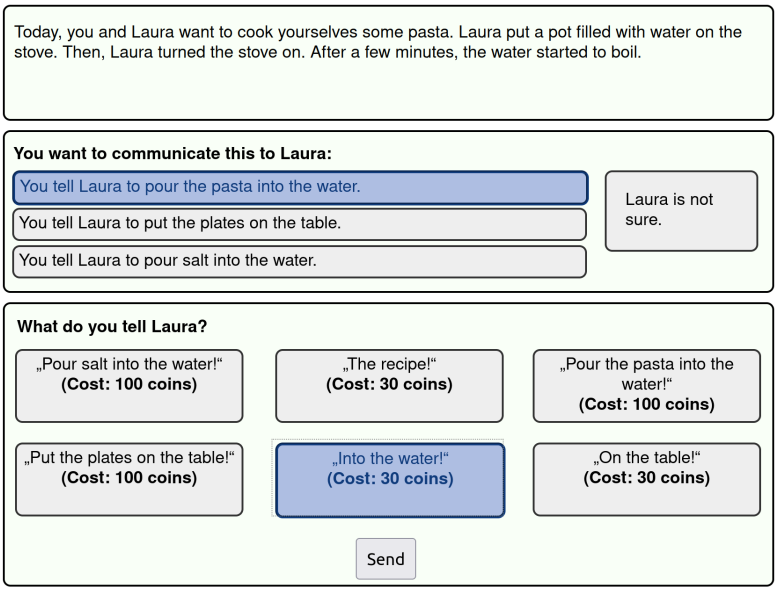


Figure 1 Screenshot of the experiment, translated to English for convenience.

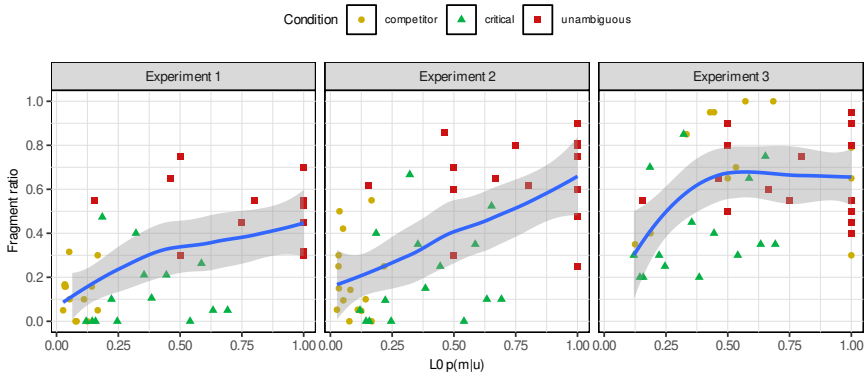


Figure 2 Ratio of fragments and sentences across the experiments and conditions.

Selected references •Lemke, R. (2021). Experimental Investigations on the Syntax and Usage of Fragments. Language Science Press. •Morgan, J. (1973). Sentence fragments and the notion ‘sentence’. In Kachru, B. et al., eds, Issues in Linguistics. 719–751. University of Illionois Press.

Effects of Conversational Context on Turn-Timing in (Non-)Autistic Dyads

Simon Wehrle[1] & Malin Spaniol[2]

[1]University of Cologne, Germany; [2]University Hospital Cologne, Germany

simon.wehrle@uni-koeln.de; malin.spaniol@uk-koeln.de

The rapid exchange of speaker turns is a foundational element of conversational interaction, with interlocutors optimising the speed of exchanges to maximize efficiency, and doing so despite the great cognitive demands on processing and prediction that this entails for each speaker–hearer. Rapid turn-timing, characterized by a preference for very short silent gaps between speakers—typically around 200 milliseconds—appears to be a near-universal phenomenon, although subtle differences in turn-timing have been observed for e.g. non-native [1] and autistic speakers [2,3], as well as patients on the schizophrenia spectrum [4], all of whom may face particular challenges in conversational interaction. Interestingly, the influence of conversational context on turn-taking remains understudied, and, to our awareness, no relevant systematic quantitative analysis has been published. Moreover, most quantitative analyses of conversational turn-timing, such as the highly influential work of [5, 6], have been restricted to specific kinds of interactions like question–answer pairs or telephone calls. In-depth analysis of turn-timing in naturalistic, multi-modal, face-to-face interaction remains scarce.

For the current work, we analysed a corpus composed of three distinct conversational contexts: small talk (Introduction), a cooperative task (Tangram), and an exchange about this same task (Discussion) (for details see [7]). We analysed data from 46 adults, 18 of whom had been diagnosed with autism spectrum disorder (ASD). The corpus has a total duration of over 11 hours. Dyads were grouped according to diagnostic status (e.g. ASD–ASD). Our primary focus was to examine differences in turn-timing based on context and diagnostic status. As in related work, we used the measure of floor transfer offset (FTO) to quantify turn-timing behaviour [3, 6]. Data and scripts are available on OSF. Bayesian inferential modelling was used for statistical analysis.

We found that turn-timing varied according to conversational context, with the Tangram context featuring more long silences, across groups (≥ 700 ms; [8]). Furthermore, autistic dyads consistently exhibited slower turn-timing across contexts, partly contradicting previous findings [3]; see Fig. 1. The non-ASD participants showed particularly fast turn-timing and a high proportion of overlaps in the introductory small talk (FTO mean = 91 ms; SD = 459) as compared to both the Tangram context (FTO mean = 316 ms; SD = 706) and to typical results in the previous literature, while autistic dyads were noteworthy for a generally higher proportion of long gaps; see Fig. 2.

In related work on the same data, we have observed that ASD participants engaged in less mutual gaze. Mutual gaze was less relevant in the Tangram context, as it was minimal across groups [9]. In this light, it is intriguing that turn-timing differences between groups were also less evident for the Tangram context, hinting at the important role of visual signals in the coordination of turn-timing in spontaneous interaction. Conversely, results from the Introduction (with strong group differences and many long gaps in ASD) recall the well-attested dispreference for small-talk situations in ASD. Overall, higher cognitive load seems to result in longer gap durations, thereby ultimately affecting communicative efficiency in more demanding conversational contexts.

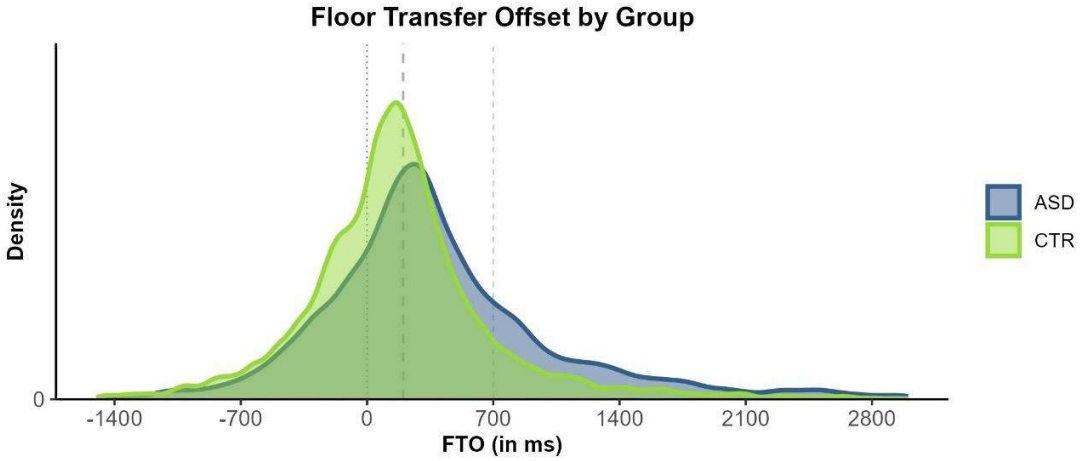


Figure 1: *Floor Transfer Offset (FTO) values by group (across contexts). Positive values represent gaps; negative values represent overlaps. The dotted line indicates the value of 0 ms FTO. Dashed lines indicate the values of 200 ms (expected typical transitions) and 700 ms (threshold for long gaps).*

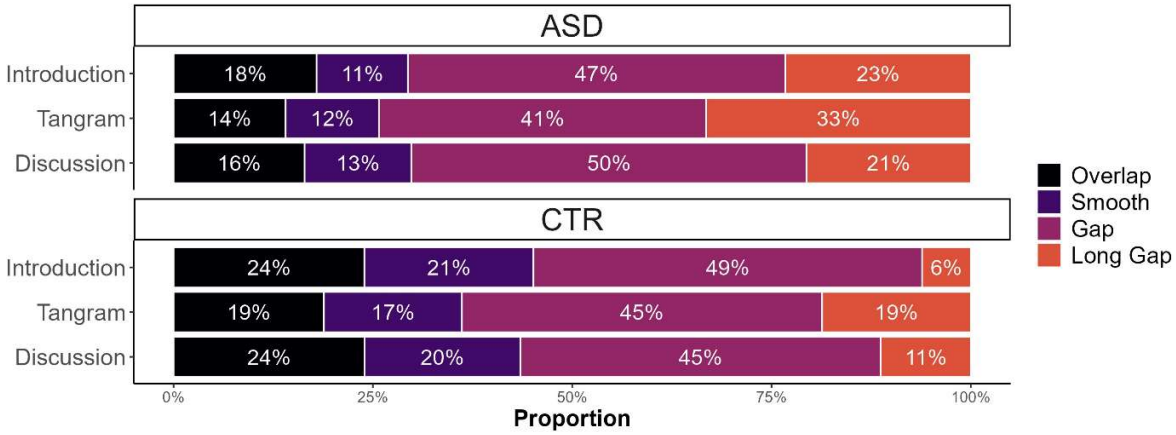


Figure 2: *Stacked bar charts by group and context, showing proportions of different transition types: overlaps (FTO ≤ -100 ms) in black, very short (smooth) transitions (FTO -99 – 99 ms) in dark purple, gaps (FTO 100 – 699 ms) in magenta, and long gaps (FTO ≥ 700 ms) in orange.*

[1] Wehrle, S., & Sbranna, S. (2023). Longer silence duration in L2 conversations is modulated by proficiency: Evidence from turn-timing in German and Italian. *Workshop on L2 Fluency, SP 2024*, Leiden, NL.
[2] Ochi, K., Ono, N., Owada, K., Kojima, M., Kuroda, M., Sagayama, S., & Yamasue, H. (2019). Quantification of speech and synchrony in the conversation of adults with autism spectrum disorder. *PLoS One*, *14*(12).
[3] Wehrle, S., Cangemi, F., Janz, A., Vogeley, K., & Grice, M. (2023). Turn-timing in conversations between autistic adults: Typical short-gap transitions are preferred, but not achieved instantly. *PLoS ONE*, *18*(4).
[4] Lucarini, V., Grice, M., Wehrle, S., Cangemi, F., Giustozzi, F., ... & Tonna, M. (2024). Language in interaction: turn-taking patterns in conversations involving individuals with schizophrenia. *Psychiatry Research*, *339*.
[5] Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *PNAS*, *106*(26).
[6] Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*.
[7] Spaniol, M., Wehrle, S., Janz, A., Vogeley, K. & Grice, M. (2024) The influence of conversational context on lexical and prosodic aspects of backchannels and gaze behaviour. *Proc. Speech Prosody 2024*.
[8] Roberts, F., & Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, *133*(6).
[9] Spaniol, M., Wehrle, S., Vogeley, K., & Grice, M. (2024). Interactions between autistic adults offer a new perspective on social gaze. *Proceedings of CogSci 2024*.

A Multifactorial Corpus-based Analysis of Classifier Positioning in Mandarin Relative Clauses

Jingwen CAO[1,*], Lawrence Yam-Leung Cheung[1]
[1]The Chinese University of Hong Kong
*Corresponding author email: jingwencao@cuhk.edu.hk

Background: In Mandarin, the modifying relative clause (RC) can either precede the head noun (1) or be separated by the demonstrative (Dem), numeral (Num) and classifier (CL) sequence (2). The two constructions differ in that while (1) is ambiguous in specificity, (2) exclusively conveys specificity (Zhang, 2006). Language users can choose either variant depending on the specific contextual conditions.
Gap: Previous studies (Sheng & Wu, 2013; Wu & Sheng, 2014) have demonstrated a correlation between the grammatical position of the head noun and positioning variation. They proposed processing-driven principles, namely the *Early Occurrence Strategy* and the *Semantic Clash Avoidance Strategy*, to explain the observed distribution. However, these studies primarily focused on individual factors, failing to capture the simultaneous contribution of various syntactic, semantic and cognitive constraints in real communicative contexts. This study adopts a multifactorial corpus-based approach to investigate the probabilistic factors constraining classifier positioning and to explore the communicative principles underlying the choice between two near-synonymous constructions.
Method: A total of N=498 observations across 4 genres (as detailed in Table 1) were extracted from the Beijing Language and Culture University Corpus Center corpus (BCC corpus) and annotated based on 8 explanatory variables (see Table 2) proposed in previous studies. Logistic regression was then conducted to evaluate the effects of predictors and their interactions.
Results: **First**, our results reveal that the occurrence of two constructions is constrained by various factors, with *RelGPos* (the grammatical position of the head NP in RCs) being the most significant. Specifically, classifiers tend to precede subject RCs but follow object RCs, consistent with previous studies. **Second**, *RCLength*, though received less empirical evidence in the literature on classifier positioning, has also been shown to be a significant predictor. Our results suggest that longer RCs increase the probability of the [RC-Dem/Num-Cl] sequence regardless of *RelGPos* (as shown in *Fig. 1*). This tendency shows some cross-linguistic parallelism, i.e., a peripheral placement of the longer and more complex dependents relative to the head (Futrell et al. 2020, Gibson et al. 2019). Language users tend to prepose RCs to minimize the length of classifier-noun dependencies, ensuring more efficient processing of language information. Besides, the interaction between *RCLength* and *Genre* is also observed in the model, which indicates that the strength and direction of *RCLength* on *Construction* varies depending on *Genre* (as shown in *Fig. 2*). Since Chinese RCs primarily provide background information due to their structural features (i.e., RCs preceding the head noun), language users, especially in formal speech contexts (e.g., newspapers and academic writing), tend to prepose RCs to a more prominent position to enhance message saliency.

Examples:

(1) san-ge [$_{RC}$ t_i dai yanjing de] xuesheng$_i$
3-CL t_i wear glass DE student$_i$
'three students who wear glasses'

(2) [$_{RC}$ t_i dai yanjing de] san-ge xuesheng$_i$
t_i wear glass DE 3-CL student$_i$
'three students who wear glasses'

Register	Dem/Num-Cl-RC	RC-Dem/Num-Cl	Total
Newspaper	20 (31.7%)	43 (68.3%)	63
Weblogs	60 (58.3%)	43 (41.7%)	103
Literature	50 (33.3%)	100 (66.7%)	150
Academic writing	100 (54.9%)	82 (45.1%)	182

Table 1. Distribution of constructions in the dataset by genre

Variable	Meaning and Levels (**Reference level**)
Genre	Source of the sentence (**newspapers**, weblogs, literature and academic writing)
DetType	Determiner type (**demonstrative,** numeral)
RelGPos	Grammatical position of head NP in relative clauses (**SRC**, ORC)
MatGPos	Grammatical position of head NP in the matrix clause (**subject**, object, others)
HeadNPAni	Animacy of head NP (**animate**, concrete inanimate, abstract inanimate)
EmbNPAni	Animacy of the argument in RC (**animate**, concrete inanimate, abstract inanimate, n/a)
HeadNPLength	Length of head NP in number of Chinese characters
RCLength	Length of RCin number of Chinese characters (incl. 'DE')
Construction	Construction of the sentence (**Dem/Num-Cl-RC**, RC-Dem/Num-Cl)

Table 2. Coding scheme

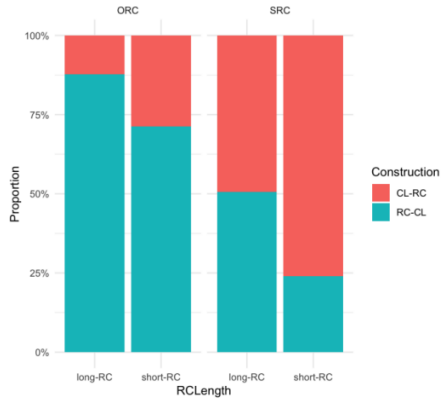


Figure 1. Distribution of construction by *RelGPos* and *RCLength*

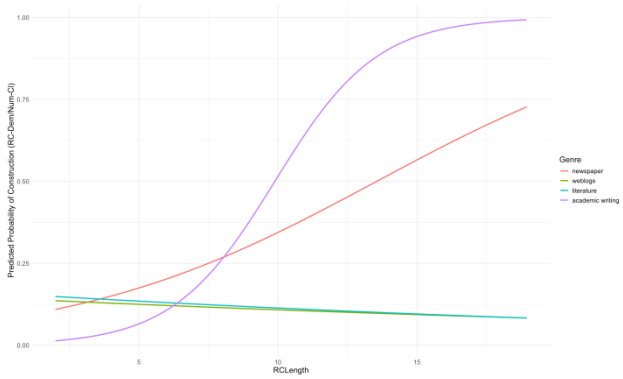


Figure 2. Interaction between *RCLength* and *Genre*

References:

Futrell, R., Levy, R., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language, 96*(2), 371–412.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389–407.

Sheng, Y., & Wu, F. (2013). Demonstrative-classifier positioning in Chinese relative clauses and its underlying reasons: A spoken corpus study*. Modern Foreign Languages, 2, 150—157.*

Wu, F., & Sheng, Y. (2014). Demonstrative-classifier positioning in Chinese relative clauses and its implication to language production models. *Journal of Foreign Languages, 37*(3), 49–58.

Zhang, N. (2006). Representing specificity by the internal order of indefinites. *Linguistics, 44*(1), 1–21.

Quantifying the Development of Communicative Efficiency in Scientific English

Julius Steuer, Marie-Pauline Krielke, Stefania Degaetano-Ortlieb, Elke Teich, Dietrich Klakow

jsteuer@lsv.uni-saarland.de

Scientific English is characterized by high informational density, technicality and abstractness, making it efficient for expert-to-expert communication (Banks, 2003; Biber & Gray, 2011, 2016). Over time, scientific English has evolved to balance lexical innovation (e.g., new technical terms) with grammatical conventionalization to ensure communicative efficiency, e.g., favoring nominal over verbal structures (Degaetano-Ortlieb & Teich, 2019; Teich et al., 2021). In this work, we explore the diachronic mechanism(s) of communicative efficiency focusing on sentence processing.

Incremental sentence processing is assumed to depend on two factors: working memory (Gibson, 1998; Lewis & Vasishth, 2005) and expectation (Hale, 2001; Levy, 2008). Both are involved in linguistic change in scientific English (Degaetano-Ortlieb & Teich, 2019; Juzek et al., 2020),but how they interact diachronically is still an open question. To address this, we use the Memory-Surprisal Tradeoff (MST; Hahn et al., 2021), which specifically models the interaction between these two factors. The MST indicates how much information a reader from a specific period needs to store in memory to reduce surprisal maximally compared to a reader from another period. We assume the MST to change over time as the linguistic code adapts to periods of innovation and conventionalization, that is, we expect the MST of some time periods to be less optimal than the MST in others, depending on the rate of innovation.

As a data set, we use the Royal Society Corpus (RSC; Fischer et al., 2020), covering scientific publications from the Royal Society from 1665 to 1996. We split each decade into a train and test section, and then estimated token-level surprisal on the test set from a language model trained on the train set using the base version of the OPT architecture (Zhang et al., 2022).

Figure 1 shows MST curves for four decades (each 100 years apart). The 17thc. shows the best MST, achieving with one bit of memory the lowest average surprisal ($<$7). In 1785-1795, the decade of the chemical revolution (Degaetano-Ortlieb & Teich, 2019), the MST deteriorates drastically: with the same amount of memory (1 bit), a much higher surprisal is needed on average (around 8 bits), possibly due to a vocabulary expansion resulting from the new discoveries at the time. In 1885-1895, the MST improves, which might be related to a period of conventionalization in the 19thc. (cf. Degaetano-Ortlieb & Teich, 2019). In 1985-1995, the MST deteriorates again, reflecting the immense increase in scientific activities in the 20thc. leading to the further expansion of a specialized vocabulary (Steuer et al., 2024) indicating specialization and diversification trends.

Overall, our findings suggest that during periods of innovation and specialization lexical expansion is rather disadvantageous to the MST. To obtain a more comprehensive picture, we want to compare (a) rather conventionalized patterns with a high degree of formulaicity (e.g., it is ADJECTIVE to/that, passive constructions), which should show an improvement of the MST, and (b) lexically productive nominal constructions (e.g. nominal compound, noun-of-noun pattern), which should show a comparatively less favourable MST. This comparison will allow us to further inspect the diachronic mechanisms of communicative efficiency at work over time.

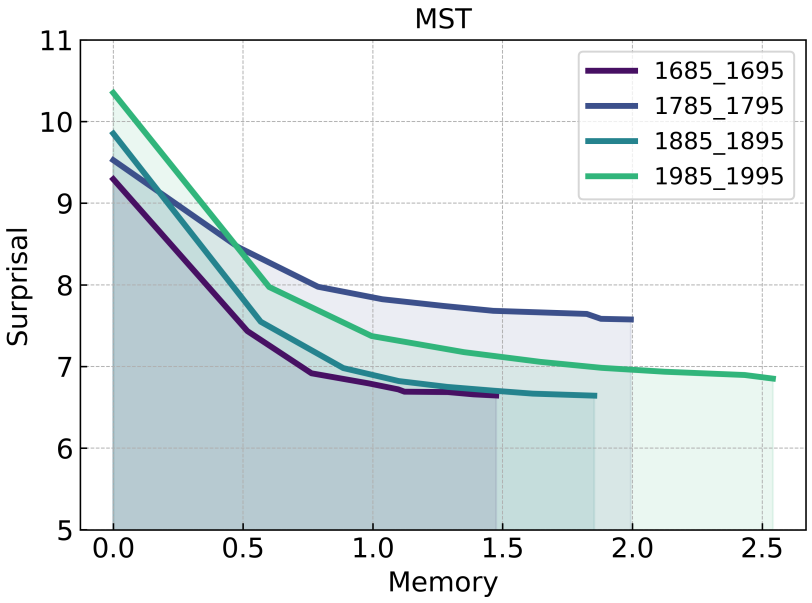


Figure 1: Memory-Surprisal Trade-Off (in bits) for four selected decades in the RSC, including the decade marking the end of the chemical revolution (1785-1795).

References

Banks, D. (2003). The evolution of grammatical metaphor in scientific writing. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, 127–148.

Biber, D., & Gray, B. (2011). The historical shift of scientific academic prose in English towards less explicit styles of expression. *Researching specialized languages*, *47*, 11.

Biber, D., & Gray, B. (2016). *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge University Press.

Degaetano-Ortlieb, S., & Teich, E. (2019). Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*.

Fischer, S., Knappen, J., Menzel, K., & Teich, E. (2020). The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 794–802.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Hahn, M., Degen, J., & Futrell, R. (2021). Modeling word and morpheme order in natural language as an efficient trade-off of memory and surprisal. *Psychological Review*, *128*(4), 726–756.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Juzek, T. S., Krielke, M.-P., & Teich, E. (2020). Exploring diachronic syntactic shifts with dependency length: The case of scientific English. *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, 109–119.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Steuer, J., Krielke, M.-P., Fischer, S., Degaetano-Ortlieb, S., Mosbach, M., & Klakow, D. (2024, May). Modeling diachronic change in English scientific writing over 300+ years with transformer-based language model surprisal. In P. Zweigenbaum, R. Rapp, & S. Sharoff (Eds.), *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024* (pp. 12–23). ELRA; ICCL.

Teich, E., Fankhauser, P., Degaetano-Ortlieb, S., & Bizzoni, Y. (2021). Less is more/more diverse: On the communicative utility of linguistic conventionalization (A. Benîtez-Burraco, Ed.). *Frontiers in Communication, Section Language Sciences*.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). OPT: Open Pre-trained Transformer Language Models [Publisher: arXiv Version Number: 4].

It's all in the Past.
An Experimental and Rational Approach to the Influence of the Sentence Onset on Past Tense Choice in German

Sophia Voigtmann (University of Kassel)
sophia.voigtmann@uni-kassel.de

Both German preterit (‘Präteritum’) and perfect (‘Perfekt’) tenses locate an event in the past (Rothstein, 2006; Thieroff, 1992) and are somewhat interchangeable, as evidenced by the loss of preterit in various dialects (Fischer, 2018). However, they differ in construction. The perfect (1a) is built with the auxiliary *haben* (‘have’) or *sein* (‘be’) plus the past participle while the preterit (1b) is morphologically marked on the full verb. Thus, the position of the lexical verb differs, appearing in the clause-final position in the perfect and the second position in the preterit tense in main clauses.

In contextless sentences, this can result in processing difficulties if the past tense is only anchored in the verb and not additionally in a temporal adverb (tAdvP), like *gestern* (‘yesterday’), at the sentence onset (1c/d). Using a tAdvP can spread the tense information more evenly. Spreading information as evenly as possible to avoid processing difficulties caused by peaks and troughs is the key concept of the Uniform Information Density Hypothesis (Fenk-Oczlon, 1983; Levy & Jaeger, 2007). This can play a role in the past tense choice in German for the following reasons:

In the preterit, conveying temporal and lexical information in a single word creates a potential peak without a tAdvP (1b) at the sentence onset, as both the lexical meaning and the tense information are processed simultaneously in the synthetic verb form. However, if the tAdvP already signals the necessity of past tense (1d), only the lexical information of the verb must be processed, reducing cognitive load. The same should be true for the perfect tense (1c). Additionally, an uncertainty on the auxiliary is resolved as *haben/sein* (‘have’/’be’) can no longer be mistaken for a full verb. However, since *haben/sein* (‘have’/’be’) is more commonly used as an auxiliary than a full verb and highly frequent, the effect of the tAdvP should be stronger for preterit.

These possible differences in the past tense choice dependent on the sentence onset are investigated in two experiments. First, we test whether context-free German sentences with perfect or preterit tenses are rated differently depending on whether they start with a tAdvP or a neutral AdvP like a sentence adverb, as another highly frequent frame setting constituent at the sentence onset (e.g. Speyer, 2008, 2009).

Thus, a rating study was conducted using a 7-point scale (7 = completely acceptable) to assess participants' judgments. The study employed a 2x2 experimental design, with factors of sentence onset (temporal (1c/d) or neutral adverb (1a/b)) and tense (preterit (1b/d) vs. perfect (1a/c)). 48 native speakers of German were recruited over prolific. 24 items as in (1) were presented to them using PCIbex.

The statistical analysis with CLMMs (Christensen, 2023)[1] shows only a main effect of the AdvP ($z= 3.89$, $p<0.001$). The tAdvPs are rated higher than sentence adverbials (tab. 1). This aligns with the proposed claim that the temporal adverbial phrase (tAdvP) distributes the tense information more uniformly throughout the sentence. Independent of the verb tense, the early tense information of the clause seems to help processing.

[1] The rating results are the dependent variable, AdvP and tense were sum-coded (+/- 0.5). Furthermore, the random slopes of AdvP and tense as well as the random intercept of the participants were included.

A second experiment, a reading time study using the same material as the rating study, is currently underway. It will test whether reading times on the finite verb differ across conditions and whether participants' regional background influences ratings, given the decline of the preterit in southern German dialects. (Fischer, 2018).

1) a) Vielleicht hat die Studentin ein neues Buch aus der Bibliothek geholt.
Perhaps has the student a new book from the library fetched.
'Perhaps, the student has fetched a new book from the library.'

b) Vielleicht holte die Studentin ein neues Buch aus der Bibliothek.
Perhaps fetched the Student a new book from the library.
'Perhaps, the student fetched a new book from the library.'

c) Gestern hat die Studentin ein neues Buch aus der Bibliothek geholt.
Yesterday has the student a new book from the library fetched.
'Yesterday, the student has fetched a new book from the library.'

d) Gestern holte die Studentin ein neues Buch aus der Bibliothek.
Yesterday fetched the Student a new book from the library.
'Yesterday, the student fetched a new book from the library.'

	Estimate	Std. Error	z-value	p-value
AdvP	0.68	0.18	3.89	<0.001
Tense	-0.23	0.18	-1.24	0.22
AdvP x Tense	0.54	0.34	1.59	0.11

Table 1 Results of the regression.

References

Christensen, R. H. B. (2023). *ordinal—Regression Models for Ordinal Data* [R package version 2023.12-4]. https://CRAN.R-project.org/package=ordinal

Fenk-Oczlon, G. (1983). Ist die SVO-Wortfolge die natürlichste? *Papiere Zur Linguistik*, *29*, 23–32.

Fischer, H. (2018). *Präteritumschwund im Deutschen: Dokumentation und Erklärung eines Verdrängungsprozesses* (Vol. 132). De Gruyter. https://doi.org/10.1515/9783110563818

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference* (pp. 849–856). The MIT Press. https://doi.org/10.7551/mitpress/7503.003.0111

Rothstein, B. M. (2006). *The perfect time span: On the present perfect in German, Swedish and English*. Universität Stuttgart. https://doi.org/10.18419/OPUS-2614

Speyer, A. (2008). German vorfeld-filling as constraint interaction. In A. Benz & P. Kühnlein (Eds.), *Pragmatics & Beyond New Series* (Vol. 172, pp. 267–290). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.172.13spe

Speyer, A. (2009). *Das Vorfeldranking und das Vorfeld-es*. *219*, 323–353.

Thieroff, R. (1992). *Das finite Verb im Deutschen. Modus—Tempus—Distanz*. Narr.

Domain expertise reduces reading times of multi-word expressions in academic texts

Sergei Bagdasarov (Saarland University), Marie-Pauline Krielke (Saarland University), Diego Alves (Saarland University)
sergeiba@lst.uni-saarland.de

Multi-word expressions (MWEs) are frequently co-occurring word combinations (Wahl & Gries, 2018) and represent a special case in cognitive processing. Due to their frequency and predictability, they are known to be processed faster than matched control phrases (Siyanova-Chanturia, 2013), while processing effort is also known to depend on intra-subject factors such as language proficiency: natives read MWEs faster than non-natives (Underwood et al., n.d.). Following rational communication principles assuming highest communicative efficiency with the lowest effort possible, MWEs contribute to language efficiency by representing highly predictable linguistic material with a clear processing advantage (Conklin & Schmitt, 2008). This processing advantage is especially relevant in scientific writing posing several cognitive challenges such as high information density, abstractness, constant lexical innovation, etc.

Considering the processing advantage of highly predictable linguistic material, our hypothesis is that domain-specific MWEs should be read faster by in-domain experts than by out-of-domain experts due to background knowledge and frequent exposure to certain types of expressions. We also expect to find different effects for different types of MWEs (e.g. discourse structure markers vs. multi-word terminology) and measures associated with processing effort such as reading times (RTs) should reflect this.

For the present study, we use the Potsdam Textbook Corpus (POTEC, Jäger et al., 2021), a naturalistic eye-tracking-while-reading corpus comprising eye-movement data from domain experts (physics and biology) and novices reading 12 German scientific texts. It follows a 2×2×2 fully-crossed factorial design, with the level of study and discipline as between-subject factors, and text domain as a within-subject factor.

We extract MWEs from the corpus using a novel 8-dimensional method proposed by Gries (2022). The method leverages both traditional frequency-based parameters and different information-theoretical measures like normalized and relative entropy. After manually filtering noisy output (e.g. chunks that only contain grammatical words or span phrase boudaries like *welche die* or *wird in der*), we obtained a list of 99 MWEs, e.g. *in der Nähe zum*, *extensiv ausgebildetes Wurzelsystem*, *qualitativ und quantitativ* etc.

To analyze the total fixation times of MWEs, we fitted a mixed-effects linear model with logged total fixation times as response and expertise level as well as the presence of general and domain-specific terminology as predictors, while controlling for variability due to differences among readers, texts, trials and MWEs. For this, we used the lmerTest package (Kuznetsova et al., 2017) available in RStudio (RStudio Team, 2020). Preliminary findings show that MWEs in general are read faster by in-domain experts (estimate $= -0.19$, SE $= 0.018$, $p < 2 \times 10^{-16}$, see Figure 1). If a MWE contains a domain-specific term, the fixation time increases (estimate $= 0.21$, SE $= 0.07$, $p = 0.0078$. However, as expected, in-domain experts read MWEs with domain-specific terminology slightly faster (estimate $= -0.058$, SE $= 0.0262$, $p = 0.0259$).

Due to the small size of the corpus, the quality of the extracted MWEs is rather limited. We thus aim to refine our MWE extraction method to improve the fit of the model. Moreover, we intend to analyse other reading time measures, comparing different classes of MWEs.

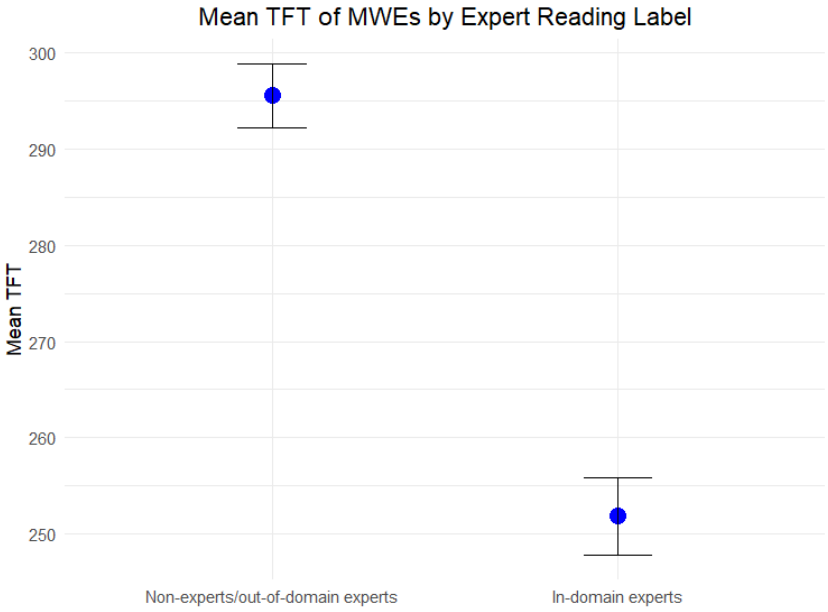


Figure 1: Mean total fixation time for in-domain experts and non-experts/out-of-domain experts (bars indicate standard error)

References

Conklin, Kathy & Norbert Schmitt (Mar. 2008). "Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers?" In: *Applied Linguistics* 29.1, pp. 72–89.

Gries, Stefan Th. (2022). "Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach". In: *Phraseology and Paremiology in English* 19.

Jäger, Lena A. et al. (Jan. 22, 2021). "Potsdam Textbook Corpus (PoTeC)". In: *Jäger, Lena A; Kern, Thomas; Haller, Patrick (2021). Potsdam Textbook Corpus (PoTeC). OSF: Open Science Framework.* Place: OSF Publisher: Open Science Framework.

Kuznetsova, Alexandra et al. (2017). "lmerTest Package: Tests in Linear Mixed Effects Models". In: *Journal of Statistical Software* 82.13, pp. 1–26.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA.

Siyanova-Chanturia, Anna (2013). "Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings". In: *The Mental Lexicon* 8.2. Publisher: John Benjamins, pp. 245–268.

Underwood, Geoffrey et al. (n.d.). "An eye-movement study into the processing of formulaic sequences". In: ().

Wahl, Alexander & Stefan Th. Gries (2018). "Multi-word Expressions: A Novel Computational Approach to Their Bottom-Up Statistical Extraction". In: *Lexical Collocation Analysis*. Ed. by Pascual Cantos-Gómez & Moisés Almela-Sánchez. Series Title: Quantitative Methods in the Humanities and Social Sciences. Cham: Springer International Publishing, pp. 85–109.

Predictability and surprisal as approximators of information status

Andrew Dyer (Language Science and Technology, Saarland University)
andrew.dyer@uni-saarland.de

Language model surprisal is often used as a rough measure of the difficulty of processing language (Goldstein et al., 2022; Wilcox et al., 2023), with more surprising tokens held to correspond to more difficult or contentful units of speech. This is often extended to “novel and unexpected” information (Xu and Futrell, 2024), with the implicit linking hypothesis that new information is more surprising. This posits a link to information status: the givenness or newness of entities and mentions in discourse (Chafe, 1976), which is itself known to affect the effort in processing words and sentences (Asahara, 2017).

Information status captures what speakers find predictable given previous context (Prince, 1981). This is mirrored by the learning objective of language models- maximising predictability of upcoming tokens- and the attention to long-range context in transformer based models. The implicit topic-modeling in such models also mirrors the view that given information corresponds to that which is topical (Givón, 1983). This accounts for bridging references in discourse, where entities not previously introduced are nonetheless unsurprising due to their semantic link with previous context (Clark, 1977; Clark and Haviland, 1977). From this perspective, it is credible that sufficiently context-aware language models’ surprisal values could approximate the information status of referents and mentions in discourse as experienced by human interlocutors.

On the other hand, this view contrasts with the more explicit view of information status usually evident in the design of information status and coreference corpora, whereby referents in discourse are explicitly assigned an information status attribute by the receiver at each mention in the discourse depending strictly on whether they have been mentioned, be that categorical (Gundel, Hedberg, and Zacharski, 1993) or gradient (Arnold and Griffin, 2007).

Despite the interest in these conflicting views, there has as yet been no direct corpus based study of the extent to which language model surprisal is correlated with, or is predictive of, information status- and vice-versa. A finding that language models approximate information status would be both a contribution to the debate on the nature of information status representation and effects (Arnold, 2016); and support for the practical approach of using language-model surprisal as a quantitative measure or stand-in for information status.

To study this, we will measure the link between information status and transformer based language model surprisal on the English and Portuguese portions of CiepInf (Dyer et al., 2024) a parallel multilingual corpus annotated for information status and coreference. We will compare the performance of a set of language models with different parameter sizes, architectures and context-sizes. We aim to shed light on the extent to which information status correlates with, or predicts, surprisal, and vice-versa; and the extent to which surprisal informs the actual forms of language use in new and given mentions. If such a pattern of interaction is found, we will have some more evidence to support the predictability-based view of information status.

References

Arnold, Jennifer E. (2016). “Explicit and Emergent Mechanisms of Information Status”. en. In: Topics in Cognitive Science 8.4, pp. 737–760. ISSN: 1756-8765. DOI: 10.1111/tops.12220. URL:https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12220 (visited on 09/25/2024).

Arnold, Jennifer E. and Zenzi M. Griffin (May 2007). “The effect of additional characters on choice of referring expression: Everyone counts”. eng. In: Journal of Memory and Language 56.4, pp. 521–536. ISSN: 0749-596X. DOI: 10.1016/j.jml.2006.09.007.

Asahara, Masayuki (Nov. 2017). “Between Reading Time and Information Structure”. In: Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation. Ed. by Rachel Edita Roxas. The National University (Philippines), pp. 15–24. URL: https: //aclanthology.org/Y17-1006.

Chafe, Wallace L. (1976). “Givenness, contrastiveness, definiteness, subjects, topics, and point of view”. In: Subject and Topic. Ed. by Charles N. Li. New York: Academic Press, pp. 25–55.

Clark, Herbert H (1977). “Bridging”. English. In: Thinking: Readings in Cognitive Science. Cambridge: Cambridge University Press, pp. 411–420.

Clark, Herbert H and Susan E Haviland (1977). “Comprehension and the Given-New Contract”. en. In: Discourse Production and Comprehension. DISCOURSE PROCESSES: ADVANCES IN RESEARCH AND THEORY 1. Norwood, New Jersey: ABLEX PUBLISHING CORPORATION, pp. 1–38.

Dyer, Andrew et al. (Dec. 2024). “A Multilingual Parallel Corpus for Coreference Resolution and Information Status in the Literary Domain”. In: Proceedings of the 22nd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2024). Hamburg, Germany: Association for Computational Linguistics.

Givón, Talmy (1983). “Topic continuity in discourse: An introduction”. In: Topic continuity in discourse: A quantitative cross-language study. Goldstein, Ariel et al. (Mar. 2022). “Shared computational principles for language processing in humans and deep language models”. In: Nature Neuroscience 25.3. Publisher: Nature Publishing Group, pp. 369–380. ISSN: 1546-1726. DOI: 10.1038/s41593-022-01026-4. URL: https://www.nature.com/articles/s41593-022-01026-4 (visited on 09/23/2024).

Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski (1993). “Cognitive Status and the Form of Referring Expressions in Discourse”. In: Language 69.2. Publisher: Linguistic Society of America, pp. 274– 307. ISSN: 0097-8507. DOI: 10.2307/416535. URL: https://www.jstor.org/stable/416535 (visited on 08/20/2024).

Prince, Ellen F. (1981). “Toward a taxonomy of given-new information”. In: Syntax and semantics: Vol. 14. Radical Pragmatics. Ed. by P. Cole. New York: Academic Press, pp. 223–255.

Wilcox, Ethan G. et al. (2023). “Testing the Predictions of Surprisal Theory in 11 Languages”. In: Transactions of the Association for Computational Linguistics 11, pp. 1451–1470. DOI: 10.1162/tacl_a_00612. URL: https://aclanthology.org/2023.tacl-1.82.

Xu, Weijie and Richard Futrell (Mar. 2024). “Syntactic dependency length shaped by strategic memory allocation”. In: Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP. Ed. by Michael Hahn et al. St. Julian’s, Malta: Association for Computational Linguistics, pp. 1–9. URL: https://aclanthology.org/ 2024.sigtyp-1.1 (visited on 03/26/2024).

Exploring Turn-Taking in People Who Do and Do Not Stutter

Lotte Eijk, Stefany Stankova, & Sophie Meekings
(Department of Psychology, University of York)
lotte.eijk@york.ac.uk

Speech most often occurs in interactions between people, where utterances seem to effortlessly flow from one into the next. Both interlocutors are able to time their utterances based on predictions about the other speaker's speech timings [1, 2] and gaps between turns have been found to be only between 0 and 300ms in many languages [3]. This turn-taking has quite extensively been investigated in typical speakers (e.g., [4, 5, 6]). However, turn-taking in conversations including populations with atypical speech such as people who stutter (PWS) has received less attention. PWS often experience involuntary syllable repetitions, prolongations, and so-called 'blocks' during which speakers are unable to produce sounds. This could lead to less predictable timings of their speech, which in turn might influence turn-taking in these conversations. Previous research (e.g., [7]) has demonstrated that typical speakers may be more likely to interrupt or complete the utterances of a conversational partner who stutters. Building on this, we aim to explore turn-taking in conversations with PWS in more detail, focussing on whether there are differences in turn-taking speed, whether PWS get a similar amount of speaking time as typical speakers, and whether PWS are more likely to be interrupted than typical speakers.

Twenty conversations were analysed. Half of the conversations were between two typical speakers (age: M = 29.7, SD = 10.5; gender: 6 F-F, 3 F-M, 1 M-M), and the other half consisted of typical-PWS pairs (age: M = 32.8, SD = 12.2; gender: 2 F-F, 7 F-M, 1 M-M). PWS were self-identified people who stutter.
Speakers participated in a Diapix spot-the-differences task [8] over Zoom. Each pair discussed two different pictures with 12 differences to be found in 10 minutes. For each round, one of the participants was the leader starting the description, and the other participant the follower. The participant who stutters always started as the leader in the first round, after which they switched.
Results were assessed using three mixed effects models with random effects for speaker and transcriber. Model 1 predicted gap duration by turn change type (PWS to typical, typical to PWS, or typical to typical). Model 2 predicted turn duration by speaker group (PWS, typical interacting with PWS, or with typical) and role (leader or follower), with an interaction between the two. The last model predicted the number of backchannels versus interruptions (automatically coded) by speaker group and role.

Preliminary results showed that leader's turns were longer and there was an influence of role on the type of overlap. We found no evidence for a difference in gap duration between the different turn changes, nor for a difference between turn duration or type of overlap between the different speaker groups. These results indicate that negative experiences by PWS could possibly be overcome by giving people clear roles in interactions. Future research could develop a more nuanced picture by using manual coding and investigating the relationship between stuttering severity and turn-taking behaviours.

References
[1] Roelofs, A., & Ferreira, V. S. (2019). The architecture of speaking. Human language: From genes and brains to behavior, 35-50.
[2] Meyer, A. S. (2023). Timing in conversation. *Journal of Cognition, 6*(1).
[3] Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J.P., Yoon, K-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587-10592.
[4] Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*(4), 555-568.
[5] Templeton, E. M., Chang, L. J., Reynolds, E. A., Cone LeBeaumont, M. D., & Wheatley, T. (2022). Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences, 119*(4), e2116915119.
[6] Bögels, S., Kendrick, K. H., & Levinson, S. C. (2015). Never say no… How the brain interprets the pregnant pause in conversation. *PloS one, 10*(12), e014547.
[7] Freud, D., Moria, L., Ezrati-Vinacour, R., & Amir, O. (2016). Turn-taking behaviors during interaction with adults-who-stutter. *Journal of Developmental and Physical Disabilities, 28*, 509-522.
[8] Baker, R., & Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods, 43*, 761-770.

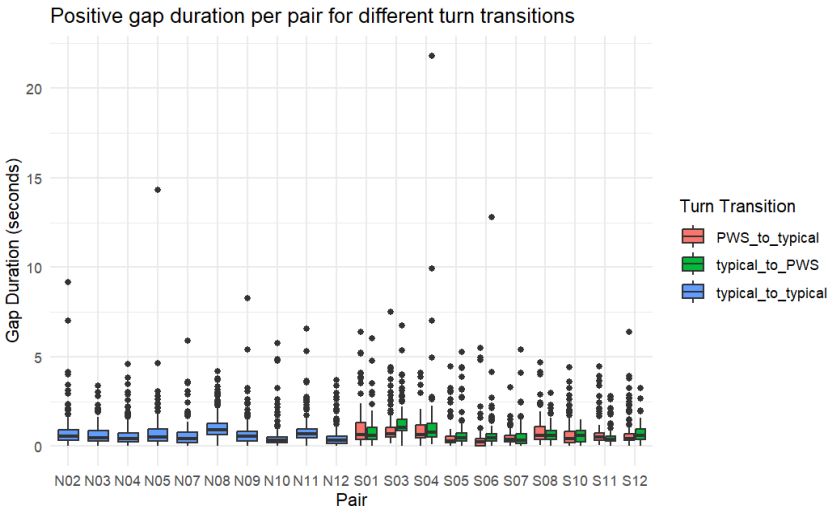


Figure 1. Positive gap duration per pair (S = PWS-typical pair, N = typical-typical pair)

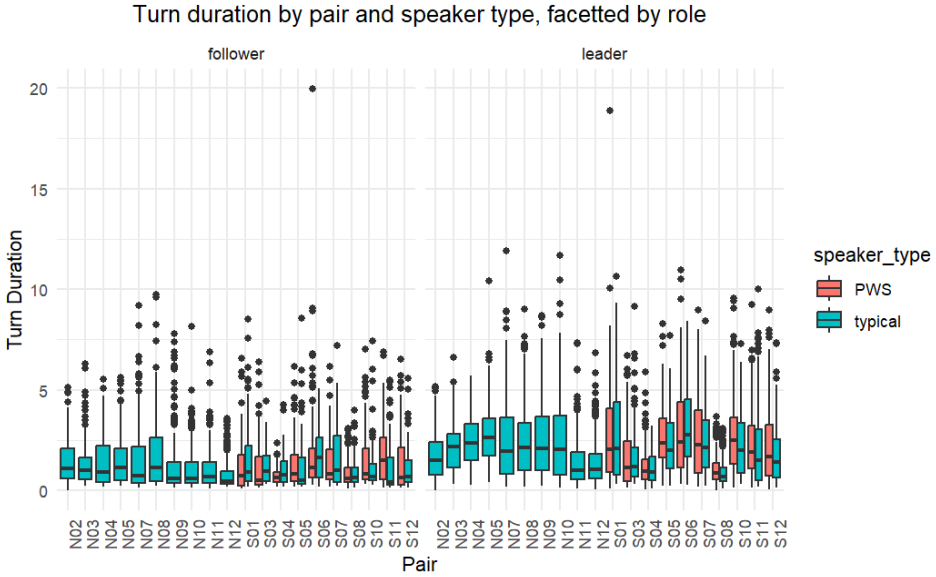


Figure 2. Turn duration by pair (S = PWS-typical pair, N = typical-typical pair), and speaker type (PWS and typical), facetted by speaker role (leader vs follower)

Colour and Discriminability Drive Over-informative Referential Expressions

Speakers are persistently *over-informative* in reference: they provide their listeners with redundant information (Pechmann, 1989). While redundancy seems to violate Grice's (1975) maxim of quantity, some propose that over-informative reference is *rational,* employed when the redundant information facilitates the perceptual processing of the listener (Rubio-Fernandez, 2021). This proposal tracks with two empirical observations: 1.) Speakers over-inform when the referred attributes are perceptually distinctive in a *visual* scene, and 2.) Speakers over-inform using colour attributes, which are held to be inherently perceptually distinctive relative to other attributes such as material constitution (Rubio-Fernandez, 2021; Kursat & Degen, 2021).

This *Perceptual Discriminability* account provides a plausible explanation of reference design as interacting with perceptual factors; however, prior studies have hitherto failed to disentangle whether (H1) the asymmetric use of colour is due to colours' high perceptual discriminability or (H2) colour is unique in reference over-and-above factors of discriminability. While the use of colour in reference declines when colour is made less perceptually discriminable (Viethan et al., 2017), it is unknown whether this reduction in discriminability eliminates speakers' preference for colour use relative to alternative attributes such as material constitution. Thus, it may be the case that perceptual discriminability only partially accounts for colours' asymmetric use.

We addressed this possibility by manipulating the perceptual discriminability of material constitution and colour in a language production experiment (N=72; see Fig. 1). We employed classic psychophysical methods of adaptive perceptual staircases to derive participant-calibrated high- and low- discriminability stimuli for colours and materials. While colour presentations were (necessarily) visual, we used audio for material presentations: the sound of wood or metal. This use of audio is two-fold: first, it allows us to investigate perceptual discriminability as a modality-general property, testing the strongest possible version of the Perceptual Discriminability account. Second, it allows us to overcome the difficulty of visually discriminating material constitution. In the language production experiment, participants were presented with coloured objects that generated an impact sound upon hitting an imaginary surface.

We investigated H1 and H2 using Bayesian logistic regressions with by-participant random intercepts, supporting both hypotheses (Figure 3). Across conditions, participants were more likely to over-inform using colour relative to material ($\beta = 1.41$, 95%CI = [1.19 – 1.64]), were more likely to over-inform in the presence of low-discriminability stimuli ($\beta = 0.28$, 95%CI = [0.12 – 0.43]), and crucially, were more likely to over-inform when redundancy was necessary to anchor reference in a high-discriminability attribute ($\beta = 1.13$, 95%CI = [0.88 – 1.39]). The model including these predictors ($\log_{10}$ Marginal Likelihood = -1191.11) far outperformed a null model's predictions ($\log_{10}$ Marginal Likelihood = -1308.62).

Confirming prior accounts (Rubio-Fernandez, 2021; Kursat & Degen, 2021), our results suggest that perceptual discriminability drives over-informative reference: speakers anchor their reference in easy-to-discriminate attributes across visual colour and auditory material. However, perceptual discriminability alone cannot account for the disproportionate use of colour: over and above such effects, colour is privileged in over-informative reference.

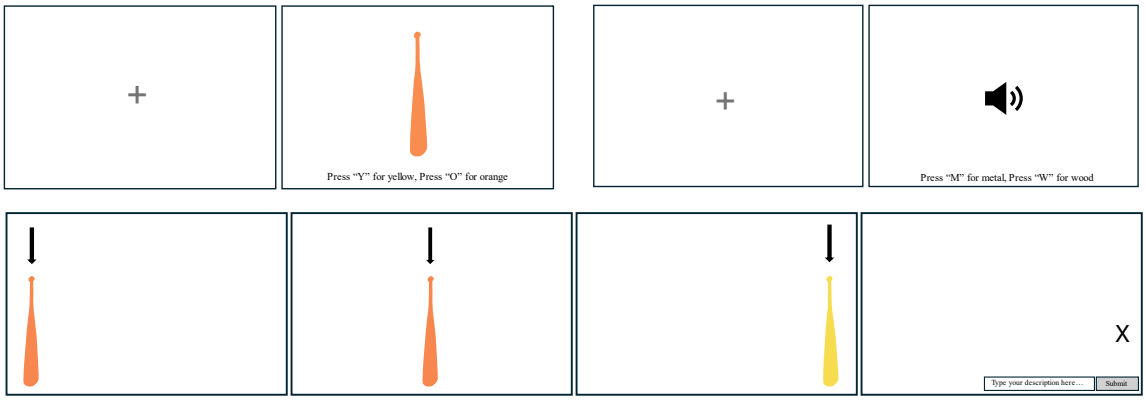


Figure 1. Task figures. The top panels show the psychophysical task, in which speakers are asked to label stimuli that vary in perceptual discriminability. Once stimuli of high- and low-discriminability stimuli are identified for colour and then material, participants move to the language-production experiment (Bottom Panel).

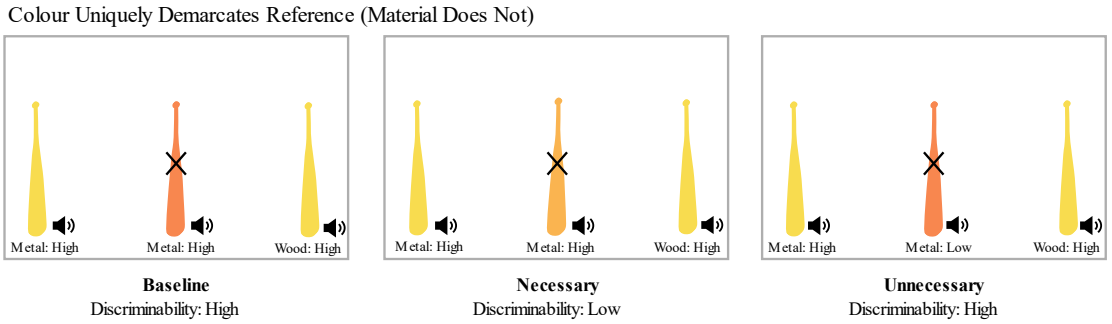


Figure 2. Illustration of experimental conditions, using colour-sufficient trials as an example case.

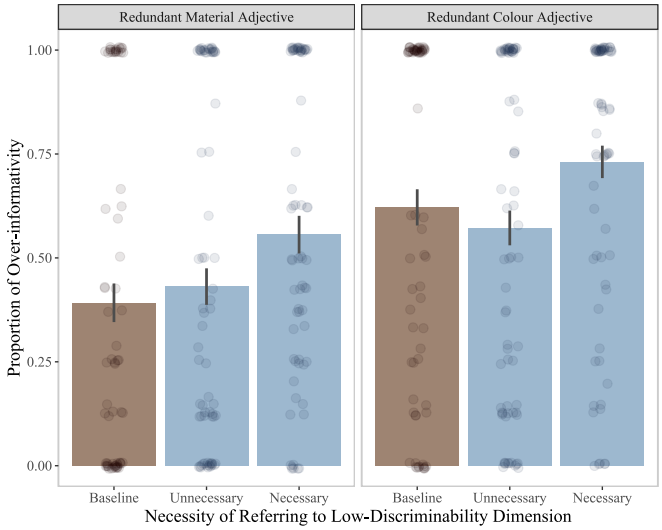


Figure 3. "Baseline" refers to high-discriminability trials. Mean proportion of over-informativity when object property discriminability is high (Base), when the property discriminability is low and not necessary to isolate the referent (Unnecessary), and when the property discriminability is low and necessary to establish reference (Necessary). Lines represent 95% bootstrapped CIs. Points represent participant means.

References

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), Syntax and semantics, Vol. 3, Speech Acts (pp. 41–58). Academic Press.

Kursat, L., & Degen, J. (2021). Perceptual difficulty differences predict asymmetry in redundant modification with color and material adjectives. Proceedings of the Linguistic Society of America, 6(1), 676-688.

Pechmann, T. (1989). Incremental speech production and referential overspecification. Linguistics, 27, 89–110.

Rubio-Fernandez, P. (2021). Color discriminability makes over-specification efficient: Theoretical analysis and empirical evidence. Humanities and Social Sciences Communications, *8*(147).

Viethen, J., van Vessem, T., Goudbeek, M., & Krahmer, E. (2016). Color in reference production: The role of color similarity and color codability. Cognitive Science, 41, 1493–1514.

Immediate Recall and Information Predictability in Reading and Listening Comprehension

Lucie Guštarová (Charles University), Jan Chromý (Charles University)
lucka.gustarova@gmail.com

Background: Recent studies [1,2] have documented systematic differences in the extent to what readers recall certain types of information immediately after reading a sentence. For example, information conveyed by direct objects tends to be recalled significantly better than information conveyed by temporal or locative adjuncts. This can be interpreted as a selective attention process: while reading, people drive their attention to information which they learned should be important/useful and they attend to a limited degree to information that is peripheral [3]. The present study examines immediate recall of information conveyed by subjects and locative adjuncts in Czech (Loc) and how it is influenced by adjunct predictability [4,5].
Method: First, predictability of Loc in combination with 57 transitive verbs was normed (N=115 Czech speakers). 24 of these combinations were then used for creating stimuli for further experiments. Two reading experiments were conducted using a self-paced reading paradigm with whole sentences appearing at once for the first experiment, and with sentences presented word-by-word for the second experiment. Once the sentence disappeared, an open-ended question was shown targeting either the subject (Who did it?), or the Loc (Where did it happen?). Then, two listening experiments were conducted. The first one used stimuli audio-recorded by native speakers of Czech with flat intonation. For the second experiment, stimuli generated by an artificial intelligence were used. The open-ended questions were visually presented, and participants responded by typing. All experiments use the same 24 experimental items and 72 fillers and manipulate word order, information targeted by the comprehension question and Loc predictability (see Table 1).
Results: Fig. 1 shows the differences in recall accuracy between the conditions in all experiments. The nested logit mixed-effects model showed a general recall difference for listening experiments (but not for the reading ones): subjects were recalled better than Loc. Moreover, the model yielded a significant effect of predictability for Loc recall in all experiments and for subject recall in experiment 1.
Discussion: Previous findings on Czech reading data [1,2] showed a tendency for better immediate recall of core information in sentences compared to accessory information. However, the present study replicated these results only in experiments involving spoken materials, not in reading. In Experiment 1 (reading, with the entire sentence presented at once), Loc predictability affected not only recall of the locative information, but also of the subject information (which stayed the same across the conditions). However, no such effect was found in the other experiments, likely due to the possibility to revisit sentences in Experiment 1. This may also explain why recall success in Experiment 1 was higher than in the other experiments. Finally, Experiment 4 (listening, AI-generated stimuli) showed no significant differences from Experiment 3. In consequence, we may conclude that stimuli generated by artificial intelligence could be suitable for use in similar experiments in the future.

Word order	Predictability	Sentence
ltvso	pred	V obchodě v neděli koupila Klára hrozně hezký pruhovaný tričko.
ltvso	unpred	V parku v neděli koupila Klára hrozně hezký pruhovaný tričko.
stvlo	pred	Klára v neděli koupila v obchodě hrozně hezký pruhovaný tričko.
stvlo	unpred	Klára v neděli koupila v parku hrozně hezký pruhovaný tričko.

Table 1: Item example. Word order values: ltvso = locative adjunct - temporal adjunct - verb - subject - object; stvlo = subject - temporal adjunct - verb - locative adjunct - object. Sentences have the same meaning and only differ in their word order and locative adjunct predictability: “Klára bought a really nice striped T-shirt in the store/in the park on Sunday.”

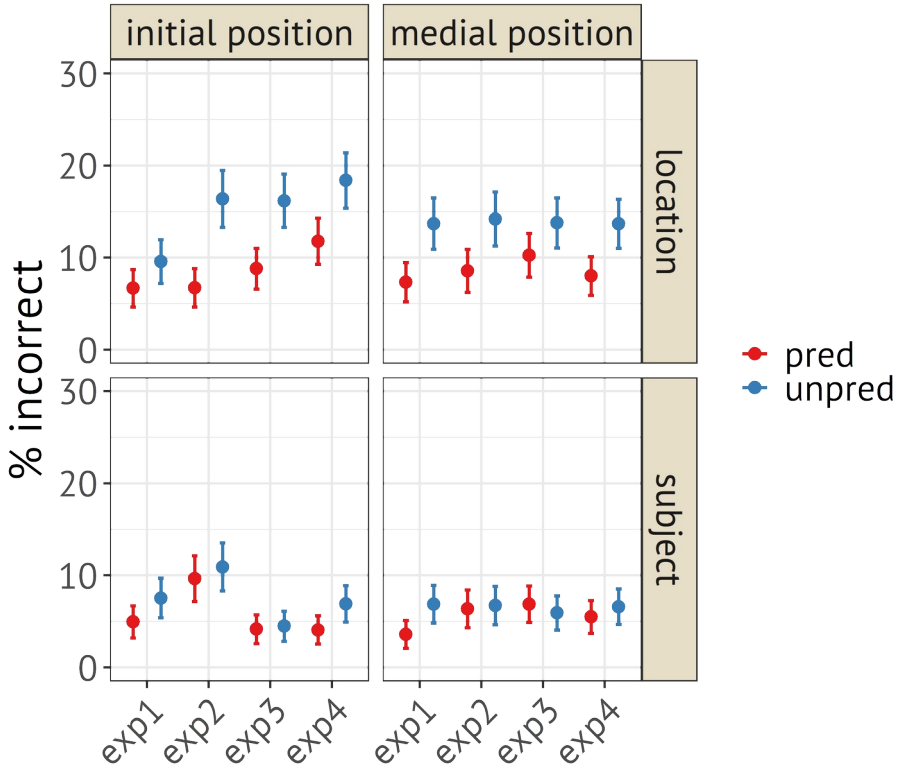


Figure 1: % of incorrect answers in all experiments. Pred = predictable locative adjunct, unpred = unpredictable locative adjunct, exp1 = reading experiment, sentence presented at once, exp2 = reading experiment, sentence presented word-by-word, exp3 = listening experiment, stimuli recorded by real speakers, exp4 = listening experiment, stimuli generated by AI.

References

[1] Chromý, J., & Vojvodić, S. (2024). When and where did it happen? Systematic differences in recall of core and optional sentence information. QJEP, 77(1), 111–132.
[2] Chromý, J. & Tomaschek, F. (submitted). Learning or boredom? Task adaptation effects in sentence processing experiments. Open Mind.
[3] Ferreira, F., & Yang, Z. (2019). The problem of comprehension in psycholinguistics. Discourse Processes, 56(7), 485–495.
[4] Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: what is next? Trends in CS, 27(11), 1032–1052.
[5] Haeuser, K. I., & Kray, J. (2023). Effects of prediction error on episodic memory retrieval: evidence from sentence reading and word recognition. LCN, 38(4), 558–574.

How Adaptive is Linguistic Prediction?

Katja Haeuser (Saarland University)
khaeuser@coli.uni-saarland.de

Language processing is known to be adaptive. For example, frequency of exposure to otherwise dispreferred syntactic structures can result in reduced processing costs for these structures over time, whereas more canonical syntactic structures become gradually dispreferred [1]. Crucially, this notion of adaptiveness has been extended to predictive processing [2,3], the rationale being that comprehenders are able to adapt their predictions depending on the likelihood that they will be fulfilled: Conditions which near-always meet linguistic predictions (i.e., high-validity conditions) should encourage comprehenders to continuously generate predictions, whereas conditions which disconfirm predictions frequently (i.e., low-validity conditions) should result in attenuated predictive processing. Unfortunately, however, to date there is rather mixed evidence on this claim. Whereas some studies support the idea that predictive processing is inherently adaptive [4,5,6], other studies have challenged these conclusions [7], or shown with new experimentation that adaptiveness of predictions is not supported by current psycholinguistic evidence [8].

However, a weakness of previous studies is that they manipulated adaptiveness of predictions by means of between-subject designs, such that subjects were either allocated to the high-or the low-validity conditions (but not both), resulting in low explanatory power and allowing for the possibility that between-subject individual differences confound the results. In addition, nearly all previous studies on prediction adaptation were likely underpowered, due to the inclusion of small sample sizes.

In this planned work, I will re-examine the adaptiveness of linguistic prediction by overcoming some of these limitations. First, I will use a within-subject design. Second, I will recruit a large sample of subjects, to be determined by power analyses. Third, I will explicitly take into account subject-related individual indifferences by measuring, for each participant, their performance in tests of working memory, inhibitory control and lexical-semantic abilities. Figure 1 illustrates the design of the experiment, which is split in two self-paced reading (SPR) blocks a 48 sentences each, separated from one another by means of the individual difference tests. Each SPR block consists of a training phrase and a test phase. In the training phase (32 items total), I train participants to rely or not rely on linguistic predictions, by presenting them with a large proportion of prediction-confirming or disconfirming- sentences (75% vs 25%, respectively; see figure caption for an example). In the subsequent test phase (16 items total), I measure predictability effects for equal proportions of predictable and unpredictable sentences.

I expect to find two critical effects. First, a predictability * trial interaction in the training blocks, suggesting that predictability effects become larger (or smaller) with repeated exposure to predictable (or unpredictable) sentences. Second, I expect to find a predictability * validity interaction in the test blocks, suggesting that predictability effects are larger (smaller) after high- (low-) validity training. The results of this study will be of interest to researchers who work on predictive processing and those who study adaptiveness of (linguistic) behavior.

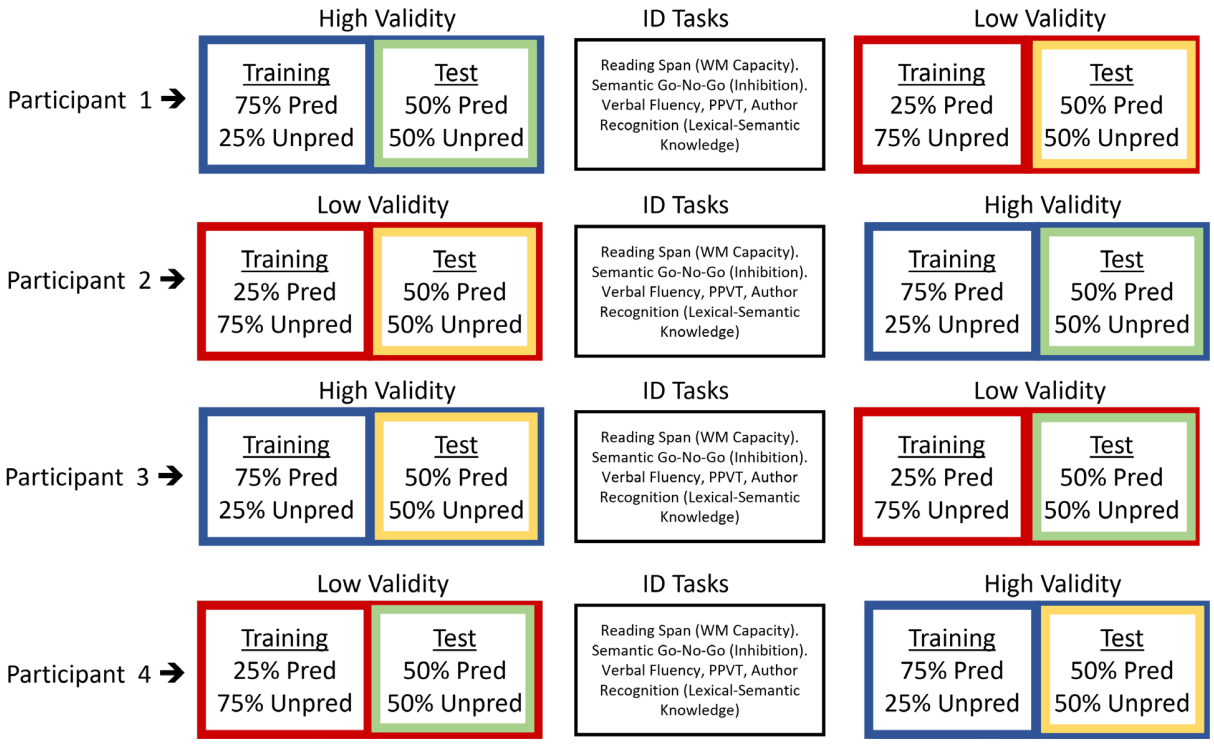


Figure 1
Experimental design. High- and low-validity blocks are marked in blue and red. The order of encountering high vs low-validity blocks first is counterbalanced over subjects, eliminating order effects. Green and yellow squares indicate different sets of test items which are crossed over validity and order sets, eliminating possible confounds related to presenting single items only in high- or low-validity conditions. Note that the training-test structure of the experiment is entirely implicit. From the perspective of the subjects, they simply read 48 sentences in one block without breaks or any other explicit or implicit indications of the training or test phase. An example of an experimental sentence is, (German original), "Als sie im Urlaub auf Mallorca waren, suchten Leo und Maja nach schönen Muscheln am $\text{Strand}_{\text{Predictable}}$ / $\text{Steg}_{\text{Unpredictable}}$ vor ihrer Ferienwohnung". Planned statistical analyses include LMER models on predictable/unpredictable nouns and the three-word spill-over region, with individual difference measures entered as interaction variables.

References

[1] Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS One*, *8*(10), e77661.
[2] Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences, 27(11),* 1032-1052.
[3] Ness, T., & Meltzer-Asscher, A. (2021). Rational adaptation in lexical prediction: The influence of prediction strength. *Frontiers in Psychology*, *12*, 622873.
[4] Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484-502.
[5] Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, *187*, 10-20.
[6] Baker-Kukona, A., & Hasshim, N. (in press). Mouse cursor trajectories capture the flexible adaptivity of predictive sentence processing. To appear in: *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
[7] Nieuwland, M. S. (2021). How 'rational' is semantic prediction? A critique and re-analysis of Delaney-Busch, Morgan, Lau, and Kuperberg (2019). *Cognition*, *215*, 104848.
[8] Van Wonderen, E., & Nieuwland, M. S. (2023). Lexical prediction does not rationally adapt to prediction error: ERP evidence from pre-nominal articles. *Journal of Memory and Language*, *132*, 104435.

Probing adaptive resource allocation in discourse: Evidence from foci and filled gaps

Morwenna Hoeks[1], Maziar Toosarvandani[2] & Amanda Rysling[2]

Given limited resources, one possible way to optimize the processing of linguistic material is to allocate fewer resources to those parts conveying less important information. Here, we test if readers utilize fewer resources processing discourse-given (≈repeated [1]) material, like the underlined part of (1b), than they do processing linguistically focused material, like *Lily* in (1b). Unlike foci, to which comprehenders generally allocate more processing resources [2-11], it may be that given material is *de*prioritized: It is often (but not always) defocused [11-12], and since it has already been parsed before, its (re)interpretation is not crucial for understanding the main message of an utterance. However, we show that reading slowdowns typically found on foci can still be observed when foci are given in E1, showing that additional resources to interpret foci are still being expended when they occur in a potentially deprioritized position. E2 extends this finding to filler-gap processing, showing filled-gap effects [13-14] even on given material, which suggests that readers do not sufficiently adapt their parsing strategies such that structure-building processes like these are absent in given material.

E1a (n=48) tested if focus slowdowns also arise on given (second-occurrence; SOF) foci, using 48 target sentences as in (2), presented in the Maze task [15]. To obtain reading time measures on given foci, different preceding contexts manipulated the Type (NEW, SOF) and Size of a focus bound by the particle *only* in those target sentences (held constant within each item). The |target| region in these sentences was always the first object NP as WIDE-NARROW RT differences there index focus marking (this word was focused in the WIDE but not the NARROW conditions). **Results.** Bayesian mixed effects models [16] revealed a main effect of Size (faster RTs in WIDE than NARROW conditions), Type (faster RTs on SOF than NEW foci), and a SizexType interaction, such that the Type effect was only reliable in WIDE conditions. E1 thus found given focus slowdowns even for SOF. **E1b** (n=42) extended E1a to foci not bound by a focus particle, to test if readers perhaps use a basic heuristic by which they always slow down on foci following those particles rather than using discourse properties like givenness to manage resource allocation. The particle was removed from E1a's SOF materials, creating conditions in which the |target| was either the second occurrence of a BOUND focus as in (3b) or that of a FREE focus as in (3d). Maze RTs were analyzed as in E1a. **Results.** revealed both a main effect of Size and Type, as well as an interaction, indicating slowdowns for both BOUND and FREE SOF. **E1c.** was identical to E1b, except that it used targets with a cleft construction as in (4) to overtly demarcate the target region as given. **Results.** revealed a Size effect, again indicating given slowdowns for foci of either type (BOUND and FREE).

E2 (n=42) compared embedded WH-clauses with an indirect object gap with embedded IF clauses, again in both NEW and GIVEN conditions as in (5). The object NP was the |target| region in all conditions, and a comparison between the WH and IF conditions at this region index a 'filled-gap effect'. **Results.** revealed a main effect of clause-Type, a main effect of Givenness and an interaction, indicating a general givenness speed-up as well as a reduced but reliable filled-gap effect in the GIVEN compared to NEW conditions.

In sum, these results do not support a view in which fewer resources are allocated to processing of material expressing less crucial information to the extent that the ramifications of focus marking (E1) and structure-building operations like filler-gap processing (E2) are not present. Future work should determine whether the obtained effects carry over to other types of deprioritized material and constructions as well.

[1] Universität Osnabrück; correspondence: mhoeks@uni-osnabrueck.de [2] UC Santa cruz

(1) **Speaker A:** Abby **read a book about penguins**.
Speaker B: No, it was LILY$_F$ (who [**read a book about penguins**]$_{GIVEN}$)

(2)
a. Abby read a book about **penguins$_F$**, Bob read a book about **gorillas$_F$** — NARROW NEW
b. Abby read **a report about penguins**, Bob read **an article about gorillas$_F$** — WIDE NEW
c. Abby read a book about **penguins$_F$**, but Bob only read a book about **bats$_F$** — NARROW SOF
d. Abby read **a report about penguins**, but Bob only read **a book about bats$_F$** — WIDE SOF

Target: {And (NEW) | No, (SOF)} Lily$_{F1}$ only read a |book| about bats

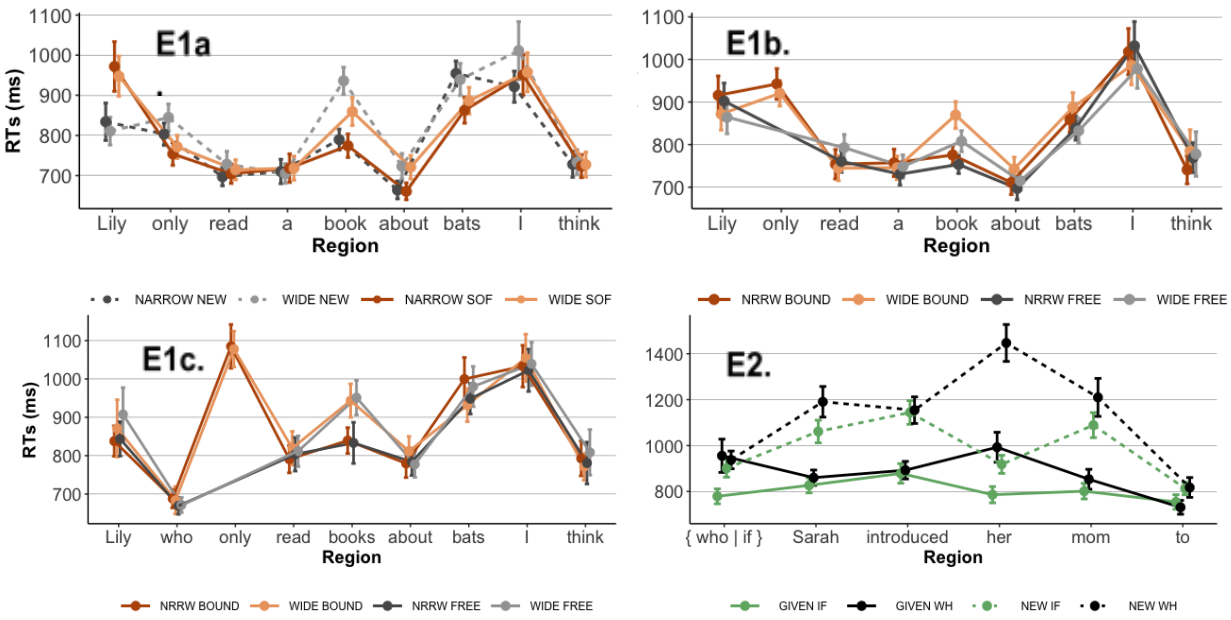
(3)
a. Abby read a book about **penguins$_F$**, but Bob only read a book about **bats$_F$** — NARROW BOUND
b. Abby read **a report about penguins**, Bob only read **a book about bats$_F$** — WIDE BOUND
c. Abby read a book about **penguins$_F$**, but Bob read a book about **bats$_F$** — NARROW FREE
d. Abby read **a report about penguins**, but Bob read **a book about bats$_F$** — WIDE FREE

Target: No, Lily$_{F1}$ { only (BOUND) | __ (FREE) } read a |book| about bats

(4) **Target:** No, it wasn't Bob, but Lily$_{F1}$ who { only (BOUND) | __ (FREE) } read a |book| about bats

(5)
a. My aunt asked who Fatima introduced my mother to at the party. — GIVEN WH
b. My aunt asked if Fatima introduced my mother to the author at the party. — GIVEN IF
c. My aunt wondered who Jonathan met at the reception. — NEW WH
d. My aunt wondered if Jonathan met the author at the reception. — NEW IF

Target: No, it was my uncle who asked {who | if} Fatima introduced |my mother| to { _$_i$ | the author} at the party.



	E1a			E1b			E1c				E2		
	β	(error)	95% Cr.I.	β	(error)	95% Cr.I.	β	(error)	95% Cr.I.		β	error	95% Cr.I.
Intcpt	2.90	(.01)	[2.87, 2.92]	2.88	(.01)	[2.85, 2.91]	2.92	(.01)	[2.89, 2.94]	**Intcpt**	2.96	(.02)	[2.93, 2.99]
Type	0.05	(.01)	[0.02, 0.04]	0.04	(.01)	[0.03, 0.05]	0.00	(.01)	[-.01, 0.02]	**C-Type**	0.11	(.01)	[0.09, 0.13]
Size	0.03	(.01)	[0.03, 0.07]	0.02	(.01)	[0.01, 0.03]	0.05	(.01)	[0.03, 0.07]	**Given**	0.12	(.02)	[0.08, 0.15]
Ty x Si	0.04	(.01)	[0.01, 0.06]	0.01	(.01)	[–.01, 0.03]	-0.01	(.01)	[-.04, 0.02]	**Ty x Giv**	0.10	(.02)	[0.06, 0.14]

Table 1: Posterior estimates E1a, E1b, E2 and E3 (logRTs) from Bayesian mixed effects models in `brms` [16] fit to log and raw RTs on all target regions (only effects reliable in both measures are reported here).

References **[1]** Schwarzschild (1999) **[2]** Cutler (1976) *Percep. Psychophys.* **[3]** Cutler & Fodor (1979) *Cognition.* **[4]** Bredart & Modolo (1988) *Acta Psych.* **[5]** Sanford & Sturt (2002) *Trends in Cog.Sci.* **[6]** Birch & Garnsey (1995) *JML.* **[7]** McKoon et al. (1993) *JML.* **[8]** Birch & Rayner (1997) *Mem. & Cog.* **[9]** Benatar & Clifton (2013) *JML.* **[10]** Lowder & Gordon (2015) *Psych. Bull.& Rev.* **[11]** Hoeks et al., (2023) *JML.* **[12]** Selkirk (2007) *Int. Stud. on IS* **[13]** Stowe (1986) *Lang. Cog. Proc.* **[14]** Omaki et al. (2015) *Frontiers.* **[15]** Boyce et al. (2020) *JML.* **[16]** Bürkner (2017) *J. Stat. Soft.*

Towards a Stochastic Model of the Human Word-finding Process Underlying Zipf's Law: A Crucial Role for Sample-Space Reduction?

Gerard Kempen (MPI for Psycholinguistics, Nijmegen, The Netherlands)
Karin Harbusch (Faculty of Computer Science, University of Koblenz, Germany)
Gerard.Kempen@MPI.NL

We propose a "bounded rationality" model of the emergence of Zipf's Law in word frequency distributions. It assumes that *Sample-Space Reduction* (SSR) as defined by Corominas-Murtra et al. (2015/16) and Thurner et al. (2015/18; henceforth CH&T) can model a key phenomenon of human language production: semantic precision being compromised in favor of easier lexical access. Zipf (1936/1949) himself conjectured a causal link between this "least effort" tendency and the frequency distributions he had observed: power-law distributions with slope parameter $\alpha \approx 1$ (Fig. 1). However, no-one has since proposed a cognitively plausible theory of why "least effort" yields distributions close to Zipf's Law (see review by Piantadosi 2014).

Selection of lexical items during language production is standardly depicted as a three-stage process: from reference delimitation via concept activation to lemma selection. (Lemmas correspond to citation forms of inflected wordforms.) In line with lexicographic practice, we assume that many concepts (meanings) are associated with one or more (synonymous) lemmas, and that concepts vary w.r.t. semantic complexity: the number of criteria determining whether the activated concept accurately covers the intended reference (denotation)—neither too broad nor too narrow. If such a lemma proves hard to access, producers will resort to *referentially "good enough" concepts* associated with more easily accessible lemmas. Options include (1) switching to a concept that delimits the reference by applying another set of criteria; (2) selecting a superordinate concept (a simpler, less precise meaning), and/or (3) splitting the delimitation criteria across multiple, simpler concepts and conceptual dependency links (thereby often restoring semantic precision). Crucially, these scenarios cause a *unidirectional frequency shift*: it boosts the frequencies of lemmas with relatively imprecise meanings, and of "function words" (many of them used to mark conceptual dependencies explicitly). This contributes to a negative correlation between the referential precision and the usage frequency of content and function lemmas.

The "good enough" ("satisficing") word-finding strategy generates ranked lemma-frequency distributions with heads densely populated by a small vocabulary of semantically imprecise but easily accessible content and function words, and with tails sparsely populated by a large set of more precise but harder to find lemmas. CH&T present mathematical proof and computer simulations of a remarkable result: SSR transforms large, relatively flat *input* power laws ($0 \leq \alpha' < 1$) into *output* power laws with $\alpha \approx 1$. This outcome obtains ("in the limit", and in the absence of external biases) with input distributions spanning large (including human) vocabularies, generalizing beyond power laws to many types of frequency distributions with zero or negatively accelerated decay. In the words of CH&T: *Zipf's Law acts as an attractor* (Fig. 2).

We propose to treat concept-frequency and lemma-frequency distributions as input and output distributions, respectively, hypothesizing that "good enough, easy-access" word finding tendencies will map the former onto the latter by emulating SSR. This presupposes that slope exponents of ranked concept-frequency distributions do not exceed 1. If this can be verified, and if additionally observed details of human word-finding turn out compatible with the assumptions underlying SSR, the proposed model will meet an important criterion put forward in Piantadosi's (2014) review: that any explanation of Zipf's Law should be founded on a plausible view of lexical processing.

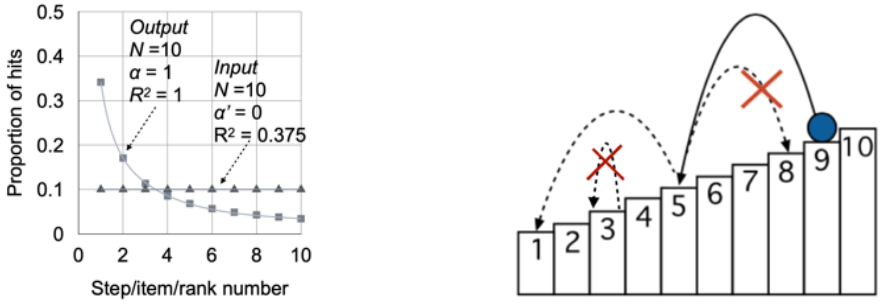


Fig. 1. "Zipf's Law" emerging from a uniform "input" distribution. *LEFT*: A "power law" is a ranked distribution of item probabilities in which the probabilities of rank i are proportional to those of a harmonic series: $p(r_i) = 1/i$ for $i = 1, \ldots, N$); e.g., the curve labeled "output". The slope of power-law curves can be adjusted by raising the denominators to a power α; $p(r_i) = 1/r_i^{\alpha}$. For $\alpha > 1$, decay is steeper, for $\alpha < 1$ it is flatter than that of a power law with $\alpha = 1$, i.e., the slope of "Zipf's Law" proper. *RIGHT*: This stairway (drawing slightly adapted from CH&T) illustrates the notion of *Sample-Space Reduction*. Imagine a ball is bouncing down the steps, never rebounding to a higher step (unidirectionality), hitting ("visiting", "sampling") the same step at most once, and halting at the lowest step. At the onset of each jump, the ball has a number of contiguous steps to chose from: the current "sample space". The probability of the ball visiting r_i during a jump equals 1 divided by the current sample space. This yields a harmonic series if the steps have equal widths, hence equal probabilities of being sampled. We refer to a distribution of step widths as "input distribution". In the left chart, the input distribution is *uniform* (which, analyzed as power law means $\alpha' = 0$, with low R^2). However, the steps may have wider and narrower widths, causing them to be visited with proportionately higher or lower probabilities (discussed in Fig. 2).

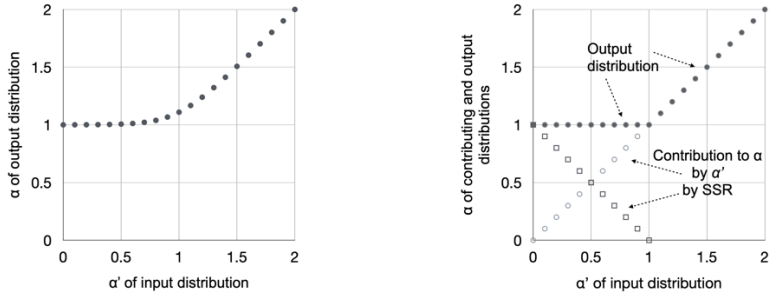


Fig. 2. Effects of applying SSR to input distributions decaying with varying slopes (α'). *LEFT*: SSR applied to *input* distributions with slope $0 \leq \alpha' \leq 2$, spanning N = 50,000 steps. This N value reflects common estimates of the active lemma vocabularies of adult natural-language users. The width distributions of the steps are power laws with either a flat input distribution ($\alpha' = 0$), a slow decay rate ($0 < \alpha' \leq 1$), or a rapid decay rate ($\alpha' > 1$). SSR tends to cause accumulation of probability mass at the head of the output distribution, thereby attenuating the probability mass occupied by the tail. With large and slowly decaying input distributions, the emerging output slope values remain within a very narrow bandwidth around $\alpha \approx 1$: "Zipf's Law as an attractor." *RIGHT*: These nearly invariant output slope values can be understood intuitively as *additive* contributions of α' and SSR to the slope of output distributions. The chart represents the situation expected when N is approaching infinity. Open and filled circles: contribution by α'; open squares: contribution by SSR; filled circles: slope values of the emerging output distributions. For instance, at $\alpha' = 0$ (uniform, horizontal input distribution), SSR is responsible for the entire output slope ($\alpha = 1$), yielding a harmonic series. For larger values of α', the SSR contributions decrease: a higher α' implies a thinner tail, hence lower probabilities of downward jumps from high-rank steps belonging to the tail. SSR runs dry at $\alpha' = 1$, meaning $\alpha = \alpha'$ for $\alpha' \geq 1$. For the proof see CH&T.

References

Corominas-Murtra, B., Hanel, R., & Thurner, S. (2015). Understanding scaling through history-dependent processes with collapsing sample space. *PNAS, 112,* 5348-5354.

Corominas-Murtra, B., Hanel, R., & Thurner, S. (2016). Extreme robustness of scaling in sample space reducing processes explains Zipf's law in diffusion on directed networks. *New J. Phys., 18*, 093010.

Piantadosi, S. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychon. Bull. Rev., 21*, 1112-1130.

Thurner, S., Hanel, R., Liu B., & Corominas-Murtra, B. (2015). Understanding Zipf's law of word frequencies through sample-space collapse in sentence formation. *J. R. Soc. Interface, 12*, 20150330.

Thurner, S., Hanel, R., & Klimek, P. (2018). *Introduction to the Theory of Complex Systems.* OUP.

Zipf, G.K. (1936). *The psycho-biology of language*. Routledge.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. AddisonWesley.

Cross-Lingual Account of Memory and Surprisal for Interpreting

Maria Kunilovskaya, Heike Przybyl, Christina Pollkläsener (University of Saarland),
Ekaterina Lapshinova-Koltunski (University of Hildesheim),
Elke Teich (University of Saarland)
maria.kunilovskaya@uni-saarland.de

Mediated language, a result of translation or interpreting, has been confirmed as identifiably distinct from comparable original production in the target language. The factors that trigger specific linguistic choices in translation and interpreting remain unclear. This computational study investigates the explanatory potential of the information-theoretical account of language processing for parallel simultaneous interpreting data. The data comes from EPIC-UdS (Przybyl et al., 2022), a corpus which contains manual transcriptions of recorded European parliament speeches and their interpreting in both directions for English-German language pair. Our implementation is grounded in memory and surprisal values computed from language-pair-specific MarianMT models, the pre-trained encoder-decoder machine translation Transformer models[1]. Interpreting data (transcripts of spoken language) is deemed more suitable than translation to model the memory component in a cross-lingual communicative process as interpreting reflects linear online processing that leaves little room for subsequent correction and editing typical for a translation product. Our aim is to build a model that would approximate available data by varying the amount of context available at the inference time. We expect that the model will return more optimal memory-surprisal trade-off (MST) for liberal translation strategy when more context is available, while more literal and conventionalised choices should achieve comparable MST when the context is limited. In this case, the specificity of production in the situation of cross-lingual mediation can be explained from the rational account of language use, which stipulates that speakers adapt their behaviour to the communicative conditions to keep the processing effort at the necessary minimum.

At the same time, the expected negative correlation of cross-lingual and monolingual surprisal (from a monolingual GPT2) would support the theoretical claim that production effort in interpreting is inversely proportional to the target audience comprehension effort, i.e., the more effort is invested into generating the target, the lower its comprehension cost. The memory component of the model will be represented by the size of the source language context (in sentences) available to generate each segment in interpreting. The document-level context is the premise for various pragmatic and discursive aspects in translation (deixis, ellipsis, lexical cohesion, word order and information flow), and its importance has been emphasized in translation studies and has recently become one of the directions for machine translation improvement (Voita et al., 2019; Läubli et al., 2020). In interpreting studies, the “upstream processing of the previous parts” is hypothesized to make comprehension “easier and faster through gradual construction of a mental model” (Gile, 2008, p. 62). In the context of this study, the translation difficulty of the subsequent sentences in a document is expected to be lower, potentially affecting the allocation of cognitive resources, and hence the outcomes of the mediation process. At the same time, some studies show that production cognitive load is often limited to the sentence boundaries and is not imported to the next sentence (Chmiel et al., 2023; Plevoets & Defrancq, 2020). While cognitive load in interpreting has been a subject of academic scrutiny, we are not aware of any other study where the interpreting data is modelled as directly conditioned by the source language input.

[1] https://huggingface.co/docs/transformers/en/model_doc/marian

References

Chmiel, A., Janikowski, P., Koržinek, D., Lijewska, A., Kajzer-Wietrzny, M., Jakubowski, D., & Plevoets, K. (2023). Lexical frequency modulates current cognitive load, but triggers no spillover effect in interpreting. *Perspectives*, *32*(5), 905–923. https://doi.org/10.1080/0907676X.2023.2218553

Gile, D. (2008). Local cognitive load in simultaneous interpreting and its implications for empirical research. *Forum*, 6(2), 59–77.

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., & Toral, A. (2020). A set of recommendations for assessing human-machine parity in language translation. *Journal of Artificial Intelligence Research*, 67, 653–672.

Plevoets, K., & Defrancq, B. (2020). Imported load in simultaneous interpreting: An assessment. In *R. Muñoz Martín & S. L. Halverson (Eds.), Multilingual mediated communication and cognition* (pp. 18–43). Routledge.

Przybyl, H., Lapshinova-Koltunski, E., Menzel, K., Fischer, S., & Teich, E. (2022, June). EPIC-UdS-creation and applications of a simultaneous interpreting corpus. In *Proceedings of the 13th Language Resources and Evaluation Conference* (pp. 1193-1200).

Voita, E., Sennrich, R., & Titov, I. (2019). When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1198-1212).

Exploring the Interaction of Linguistic and Visual Cues in Sentence Production: The Role of Information Structure

Ágnes Lukács, Anna Babarczy, Krisztina Sára Lukics, Péter Márton Rácz, Bálint József Ugrin
(Budapest University of Technology and Economics, Department of Cognitive Science, Budapest, Hungary)
lukacs.agnes@ttk.bme.hu

Structural priming refers to the reuse of previously encountered sentence structures both in language comprehension and production (Ziegler et al., 2019). We explored the priming effects of linguistic (word order) and nonlinguistic (changes highlighting different elements in a scene) information structure in Hungarian sentences. Hungarian serves as an ideal testing ground for examining such effects. Unlike many languages in which word order marks grammar relations and changes affect the fundamental meaning of the sentence, in Hungarian, such changes serve discourse-related functions, such as the novelty or emphasis of certain sentence elements (É Kiss, 2002). We looked at two word orders: Verb Subject Object (VSO) sentences highlight the action, while Object Verb Subject (OVS) sentences highlight the patient.

(1) *Mossák a gyerekek az autót*.
wash-3pl.indef the children the car-acc
“The children are washing the car.”

(2) *Az autót mossák a gyerekek*.
the car-acc wash-3pl.indef the children
“The children are washing THE CAR.”

We conducted a structural priming experiment with 70 participants and 64 trials per participant. Each trial displayed a prime picture and a target picture which depicted simple transitive activities. The agent, patient, and action varied between prime and target scenes. The prime picture was accompanied by either a VSO or an OSV auditory prime. Participants had to describe the target picture. We expected visual changes involving optimal contrast in the scene (paralleling the linguistic structure: V change for VSO primes, and O change for OSV sentences) to yield highest ratios of reuse of the respective word orders. We used Generalised Linear Mixed Models to predict change in response word order from prime sentence structure and visual change.

Results showed significant priming across all conditions relative to baseline (VSO and OSV sentences nearly absent without priming, $p < 0.05$). Single element changes in the visual scene between prime and target caused the highest reuse ratios, independent of sentence structure type: V change and O change conditions exhibited the highest reuse ratios, but this was equally true for both VSO and OVS structures. Priming, though less efficient, persisted with two changes in the scene and even when all visual elements changed. Findings suggest that word order structures alone induce priming, enhanced by lexical overlap (no overlap versus any degree of overlap). Overall, information structure coded by word order had a robust priming effect, independent of discourse functions (see e.g. Goldberg 2001; Bod 2006; Linzen & Jaeger 2016). Further studies are needed to disentangle linguistic and nonlinguistic information structure effects on language production.

References

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, 23(3), 291–320. https://doi.org/10.1515/TLR.2006.012

É. Kiss, K. (2002). *The syntax of Hungarian*. Cambridge University Press.

Goldberg, A. E. (2001). Patient arguments of causative verbs can be omitted: The role of information structure in argument distribution. *Language Sciences*, 23(4–5), 503–524. https://doi.org/10.1016/S0388-0001(00)00034-6

Linzen, T., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 40(6), 1382–1411. https://doi.org/10.1111/cogs.12274

Ziegler, J., Bencini, G., Goldberg, A., & Snedeker, J. (2019). How abstract is syntax? Evidence from structural priming. *Cognition*, 193, 104045. https://doi.org/10.1016/j.cognition.2019.104045

Predicting Discourse Relations: The Processing Benefit of a Connective

Marian Marchal (Saarland University), Merel Scholman (Utrecht University), Ted Sanders (Utrecht University), Vera Demberg (Saarland University)
marchal@coli.uni-saarland.de

Rationale. Comprehenders make predictions at various linguistic levels [e.g. 1, 2]. To illustrate, when hearing *Lizzy was tired,* one might make predictions about whether the speaker will next discuss the cause or the consequence of Lizzy's tiredness (a relation prediction), what a specific consequence might be (a semantic prediction; e.g. *drink coffee, go to bed*), and how a specific consequence will be formulated (a lexical-syntactic prediction; e.g. *make a cappuccino, took a sip of her coffee).* It is unclear, however, to what extent predictions of discourse relations (DRs) [3] influence processing beyond semantic or lexical-syntactic predictability, since these factors are confounded in previous work. Here, we examine (1) whether DR predictability explains processing difficulty *beyond* other levels of predictability and (2) whether the processing benefit provided by a connective [e.g. 4, 5] can be explained by enhanced prediction or has an additional effect.

Method. We operationalize DR predictability as **relation surprisal** (RS), the negative log probability of the DR type given the context, and take **semantic information value** (SIV) [6] as a measure of semantic predictability. These were calculated based on continuations in a human (n = 160) cloze task. **GPT2 surprisal** (GS) without context served as an estimate of lexical-syntactic predictability. We conducted a region-by-region self-paced reading (SPR) study (n=121) as well as an eye-tracking-while-reading (ET) study (n=79), in which native English speakers read 24 target stories containing cause-consequence sentence pairs as in Table 1. In the explicit but not the implicit condition, these DRs were marked with the connective *therefore*. We analyzed log-transformed response times (RT) from SPR and first-pass (FP) and total fixation (TF) duration from ET. Using mixed-effects piecewise structural equation modeling (pSEM, Figure 1) [7], we estimated the direct and indirect effects of the predictors of interest, while controlling for trial and length.

Results. First, we examine how connective presence influences predictability (see Table 2). As expected, RS is higher in the implicit condition. There was no significant effect of connective on GS, but RS predicts SIV, and as such the connective indirectly facilitates semantic predictions. With respect to processing difficulty, SIV positively predicted all three reading measures, providing evidence for semantic prediction (see Table 3). GS only predicts ET reading measures. Contrary to expected, RS negatively predicted TF, suggesting that more expected relations are read *slower* when accounting for facilitation through semantic prediction. Crucially, there was a significant effect of connective beyond predictability for all measures except for TF.

Conclusion. We show that the connective increases the predictability of upcoming material, and that predictability influences reading times, though sometimes in unexpected ways. The effects of predictability should thus be taken into account when analyzing the facilitating effect of the connective. We find that the connective facilitates processing beyond making upcoming material more predictable.

conn	pred	item text	RS	SIV	GS
Context		Angela used to live in a small flat in Atlanta.			
exp	high	She didn't pay rent for months. She **was evicted**	0.23	0.57	13.16
imp	high	She didn't pay rent for months. Therefore, she **was evicted**	0	0.72	13.82
exp	low	She had over fifteen cats. She **was evicted**	1.15	1.19	13.16
imp	low	She had over fifteen cats. Therefore, she **was evicted**	0	1.21	13.82
Context		… by her landlord. Angela decided to move to a rural area.			

Table 1. Example of an item in each condition, along with relation RS, SIV and GS estimates. Note that the manipulation of predictability was binary, but a continuous measure was included in the analysis.

	predictor	path	type	β	95% CI	
GS	conn	c	direct	.15	[-.01,.30]	
	length	f	direct	.62	[.48,.72]	*
SIV	conn	b	direct	.02	[-.20,.24]	
	length	e	direct	.27	[-.11,-.41]	*
	RS	d	direct	.32	[.18,.47]	
	conn	ad	indirect	-.16	[-.25,-.09]	*
RS	conn	a	direct	-.50	[-.58,-.41]	*

Table 2. (In)direct effects of the presence of a connective, the predictability measures and length on the different predictability measures. These are independent of the reading time measures. * indicates significance at the .05 level. Paths refer to Figure 1. Connective was deviation-coded (imp: -1; exp: 1).

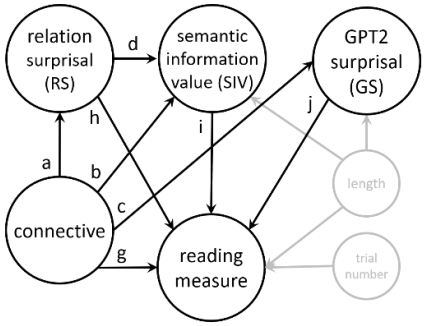


Figure 1. Structure of the pSEM. Note that in the model for the ET measures, there was an additional path between trial number and GS.

			RT			FP			TF		
predictor	path	type	β	95% CI		β	95% CI		β	95% CI	
conn	g	direct	-.06	[-.08,-.03]	*	-.10	[-.15,-.05]	*	-.06	[-.10,-.01]	*
GS	j	direct	.02	[-.01,.05]		.10	[.02,.15]	*	.07	[.03, .15]	*
SIV	i	direct	.03	[.01,.05]	*	.08	[.04,.13]	*	.14	[.10,.19]	*
RS	h	direct	.02	[-.01,.04]		-.07	[-.13,-.03]	*	-.02	[-.07,.03]	
conn	ah+adi+...	indirect	-.01	[-.02,.01]		.04	[.01,.08]	*	.00	[-.04,04]	
RS	di	indirect	.01	[.00,.02]	*	.03	[.01,.05]	*	.05	[.02,.08]	*
conn	g+ah+…	total	-.07	[-.09,-.05]	*	-.06	[-.10,-.01]	*	-.06	[-.10,-.01]	*
RS	h+di	total	.03	[.01,.05]	*	-.05	[-.10,-.00]	*	.03	[-.04,.08]	

Table 3. Direct, indirect and total effects of the predictors of interest on the various reading measures. The estimates for trial and length are not presented due to lack of space. Paths refer to Figure 1.

References

[1] Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31,* 32–59. **[2]** Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & De Lange, F. P. (2020). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32). **[3]** Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics,* 25(1), 1-44. **[4]** Cozijn, R., Noordman, L. G., & Vonk, W. (2011). Propositional integration and world-knowledge inference: Processes in understanding 'because' sentences. *Discourse Processes, 48*, 475–500. **[5]** Van Silfhout, G., Evers-Vermeul, J., & Sanders, T. (2015). Connectives as processing signals: How students benefit in processing narrative and expository texts. *Discourse Processes, 52,* 47–76. **[6]** Giulianelli, M., Wallbridge, S., & Fernández, R. (2023). Information value: Measuring utterance predictability as distance from plausible alternatives. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5633–5653). Singapore: ACL. **[7]** Lefcheck, J. S. (2016). piecewisesem: Piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, 7, 573–579.

A Revised Version of the German Author Recognition Test

Geraldine Müller (Universität des Saarlandes), Katja I. Haeuser (Universität des Saarlandes)
s9grmuel@stud.uni-saarland.de

Print Exposure reflects the extent of an individual's reading habits and has been shown to be a relevant predictor for verbal and cognitive abilities that involve language processing [1 - 3]. Since their inception in the late 1980s, author recognition tests (ARTs) have been used successfully to measure print exposure [4]. In ARTs, participants are tasked with discriminating authors from non-authors. A German version of the ART was first introduced and tested by Grolig and colleagues in 2020. Even though their test measured print exposure reliably (split half reliability of r = .95), it also included two problematic aspects. First, distractors were used that, in some cases, were not clearly distinguishable from the test items, as the names for the distractors were picked from the editorial boards of scientific papers and publications, making their status regarding authorship unclear. Second, the test was piloted on a sample composed largely of academics and visitors to the Frankfurt Book Fair, two groups for whom a higher level of print exposure can be expected compared to the general population. This aspect is clearly problematic as we know that author recognition tests may vary in their suitability for different target groups, providing more reliable results for individuals with higher educational backgrounds
[6 - 7].

Here, we introduce a new, improved version of the German ART. We developed more appropriate distractors for the test and thoroughly verified their potential authorship through extensive research. Additionally, the test results were analyzed separately for target groups with and without an academic degree. Furthermore, we compared the impact of two test-formats: the forced-choice format vs check-all. Previous research has demonstrated that the response format of a psychometric test can significantly influence participants' response behavior [5]. Earlier versions, including the specific predecessor by Grolig and colleagues, were primarily published in the check-all format. Finally, we correlated ART performance against two other normed measures of verbal abilities, the LexTale vocabulary test and a verbal fluency test, both testing for important components underlying effective communication. The moderate correlations we found align with prior research demonstrating ART's links to verbal abilities and highlight the relevance of print exposure for cognitive abilities related to language and communication.

The new test version comprises 120 items (80 authors, 40 non-authors). Participants completed the ART in either the traditional check-all format or the forced-choice format, which version was specifically devised to test the effects of test format. Results show that the improved version of the German ART exhibits robust reliability for both the check-all version (Cronbach's alpha $\alpha = 0.92$, split-half reliability $r = 0.93$) and the forced-choice version ($\alpha = 0.95$, $r = 0.89$). Additionally, as expected, significant performance differences were found between groups with and without a university degree, with subjects holding a university degree outperforming those without. The comparison of test formats revealed higher hit rates and false alarm rates for the forced-choice vs check-all format. In sum, our results indicate that both education level and response format play a crucial role in shaping test performance for ART, underlining the need for their careful consideration in future test designs.

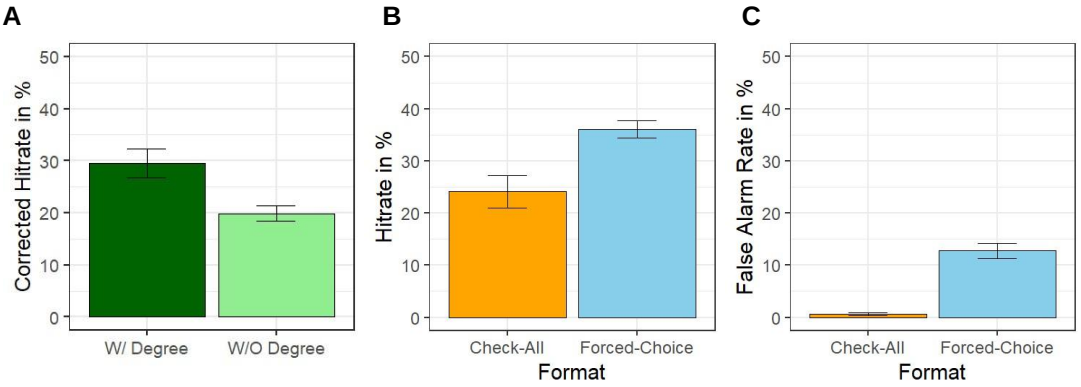


Figure 1. A: Barplots showing the ART performance (Hits-False Alarms) for participants with and without academic degree. **B-C:** Barplots showing the differences in Hit (B) and False-Alarm rates (C) between Check All and Forced Choice format.

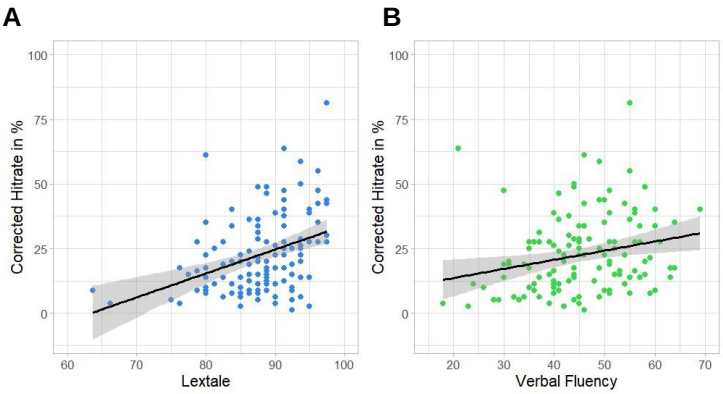


Figure 2. A-B: Scatterplots showing the correlations between ART performance and Lextale ($r = 0.34$, $p < .001$) and Verbal Fluency ($r = 0.24$, $p = .013$) performance.

References:

[1] Cunningham, A. E., & Stanovich, K. E. (1998b). What reading does for the mind. *The American Educator*, *22*(3), 8–15.

[2] Martin-Chang, S., & Gould, O. N. (2008). Revisiting print exposure: exploring differential links to vocabulary, comprehension and reading rate. *Journal of Research in Reading*, *31*(3), 273–284.

[3] Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, *137*(2), 267–296.

[4] Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and readingabilities in college students. *Behavior Research Methods*, *40*(1), 278–289.

[5] Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing Check-All and Forced-Choice Question Formats in Web Surveys. *Public Opinion Quarterly*, *70*(1), 66–77.

[6] Brysbaert, M., Sui, L., Dirix, N., & Hintz, F. (2020). Dutch Author Recognition Test. *Journal of Cognition*, *3*(1), 6.

[7] Grolig, L., Tiffin-Richards, S. P., & Schroeder, S. (2020). Print exposure across the reading life span. *Reading and Writing*, *33*(6), 1423–1441.

Rational Inference Underlies Judgments of Grammatical Well-Formedness

Moshe Poliak* (MIT), Aixiu An* (MIT), Roger Levy (MIT), Edward Gibson (MIT)
*indicated equal contribution; corresponding email moshepol@mit.edu

Background: Acceptability judgments are the main tool for investigating the grammar of a language, both with human subjects[1-4] and with large language models[5]. But what makes a sentence more or less acceptable? In this project, we evaluate three potential mechanisms across 8 languages. (1) Ever since Chomsky[6], it has been a standard assumption that grammaticality exists on a spectrum. Partially formalizing this idea, Pullum[7] proposed a framework in which the grammar of a language is a set of binary constraints, and the more constraints a sentence violates, the less grammatical that sentence is (**Equation 1**). (2) Another potential mechanism builds on mechanism 1 but also considers sentence length, such that longer sentences are less acceptable[8] (**Equation 2**). Whereas mechanisms 1-2 evaluate grammatical well-formedness from a linguistic structural perspective, we propose a different mechanism rooted in rational communication. (3) If the goal of language is to successfully exchange information, then **grammatical well-formedness should reflect how easy it is to infer the speaker's intention**. Therefore, mechanism 3 predicts that the **higher the percentage of uncorrupted information** in a sentence, the more acceptable the sentence will be (**Equation 3**), in addition to longer sentences being less acceptable. We operationalize this by dividing the number of corruptions in a sentence by the sentence's length.
Method: We evaluated the 3 mechanisms above using 8 experiments with the same design across Danish, English, French, German, Hindi, Korean, Mandarin, and Russian (Ns= 40, 40, 40, 40, 33, 36, 30, 41 respectively, after exclusions). For each language, we selected 72 sentences of different lengths (range: [4,43], median: 15), creating 4 conditions from each sentence: original, 1 transposition, 3 transpositions, and a shuffled word order (see **Table 1**). Participants were presented with all the sentences once, with semi-random assignment of condition to sentence such that each participant saw each condition the same number of times. Participants were asked to rate how natural each sentence is and then responded to a comprehension question about the sentence (inclusion criterion: >80% accuracy).
Results: The results from all languages are represented in **Figure 1**. We fit 3 cumulative Bayesian regressions with random intercepts for participants and for items within languages, varying the fixed effects according to **Equations 1-3**, and compared their predictive abilities using WAIC[9-11]. **Equation 3** had the best predictive ability by far, followed by **Equation 2** (ELPD Difference from Equation 3 = -1011, SD = 46.1), which was not substantially better than **Equation 1** (ELPD Difference from Equation 3 = -1045.9, SD = 48.7). Moreover, this inferential finding replicated within each language separately, as is also seen in the descriptive **Figure 1**, where sentences with 1-5 corruptions increase in acceptability the longer they are, for all languages.
Discussion: We find that the best explanation for the grammaticality of sentences is rooted in rational comprehension: the grammaticality of sentences reflects how easy it is to recover what the speaker intended, adding to the growing evidence that the goal of language comprehension is to understand the message that the speaker intended to communicate.

Equation 1: $grammaticality = \beta_0 + \beta_1 * corruptions$

Equation 2: $grammaticality = \beta_0 + \beta_1 * corruptions + \beta_2 * |s|$

Equation 3: $grammaticality = \beta_0 + \beta_1 * |s| + \beta_2 * \frac{corruptions}{|s|}$

Table 1. A sample item in English.

condition	sentence
original	A ball flying in the air can hurt.
1 transposition	A ball flying in air the can hurt.
3 transpositions	A ball in flying can the air hurt.
shuffled word order	In flying can ball a hurt air the.

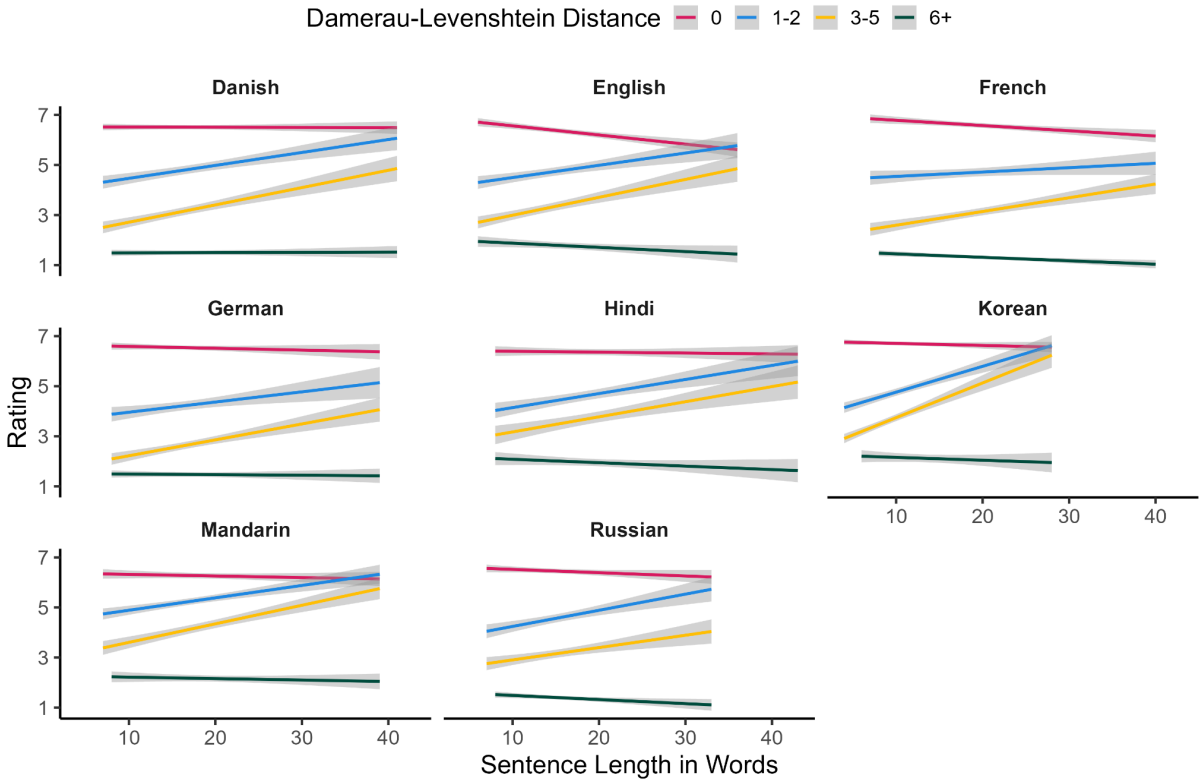


Figure 1. The line of best fit for acceptability rating as predicted by sentence length and the Damerau-Levenstein distance between the original and corrupted sentences, split by language. The distance is the minimal number of words that need to be deleted, inserted, substituted, or transposed with the neighboring word to arrive from the presented sentence to the original sentence that was collected from the UD treebank.

References: [1]Schütze, 1996; [2]Cowart, 1997; [3]Myers, 2009; [4]Sprouse et al., 2013; [5]Warstadt et al., 2019; [6]Chomsky (1964); [7]Pullum (2020); [8]Lau et al. (2017); [9]R core team (2024); [10]Wickham et al. (2024); [11]Bürkner (2017).

Limits of Dependency Length Minimization

Sidharth Ranjan (sidharth.ranjan03@gmail.com) and Titus von der Malsburg
University of Stuttgart

Dependency locality in the form of dependency length minimization (DLM) has been demonstrated as an explanatory principle behind word order preferences in natural languages (Futrell et al., 2020). This principle seeks to keep any pair of linked head-dependent words in a dependency tree as *close* as possible in their linear order within a sentence due to efficiency factors from limited memory capacity. It is unclear, meanwhile, how much DLM is employed in a specific language. Furthermore, the cognitive processes driving the minimization of dependency length is unknown. Following a recent study by Ranjan and von der Malsburg (2024), we hypothesize that placing a short preverbal constituent next to the main verb explains constituent ordering decisions better than the global minimization of dependency length across SOV languages. We refer to it as "least-effort" strategy, as it reduces the dependency lengths between the verb and all its preverbal dependencies but does so in a cost-effective manner by streamlining the search space of possible constituent orders. We substantiate our hypothesis using large-scale corpus evidence from Universal Dependency Treebank (Zeman et al., 2022). Finally, we argue that our findings can be situated within the frameworks of *good-enough* account of language processing (Ferreira et al., 2002), and *bounded rationality* in decision making (Gigerenzer et al., 2011), where *fast-but-frugal* heuristics hold precedence over extensive searches for optimal solutions.

Method. Our dataset includes sentences from the seven major SOV languages in the UD treebank: *Basque, Hindi, Japanese, Korean, Latin, Persian, and Turkish*. For each natural sentence (reference; 'ref') in the corpus, representative of human preferred choice, a large number of counterfactual variants ('var') were automatically created by randomly permuting the preverbal constituents in the sentence whose head was directly dependent on the root verb in the dependency tree (see Ex. 1 for an illustration with four preverbal constituents C_i in a sentence). We then examined the distribution of the length of preverbal constituents and dependency lengths within a sentence. Thereafter, we tested our main hypothesis by deploying these two predictors in a logistic regression model to distinguish reference sentences from the generated variants.

Results. As the preverbal constituents approach the main verb, the global DLM would predict a gradual decline in their lengths. On the other hand, the least-effort strategy would expect optimization mostly on the preverbal constituent next to the main-verb. Fig. 1 validates the prediction across SOV languages, implying that sentences in the natural corpus show a preference for either optimizing the length of preverbal constituent next to the main verb or at least prefer it. Next, if speakers employ least-effort strategy, the length of constituent closest to the verb ('CL Last') should be better at predicting correct choices *i.e.,* corpus reference sentences ('ref') against generated variants ('var') than total dependency length ('Total DL'). Fig. 2 presents the summary of our dataset and Table 1 the results of our models. Aligned with our prediction, we found that 'CL Last' consistently outperformed 'Total DL' in predicting corpus reference sentences, in terms of classification accuracy (% of correctly predicted reference sentences, 'ref'), except Basque and Japanese. 'Total DL' and 'CL Last' gave same outcome for ~70% cases across SOV languages with success cases (%Correct) indicated in parenthesis. The percentage of different outcomes can be inferred from the same column (100 minus %Same). Further, adding 'CL Last' feature over a baseline model with 'Total DL' feature, induced a significant increase in the accuracy ($p < 0.001$ using McNemar's test) for all SOV languages, including Basque and Japanese.

Conclusion. Overall, our findings show that SOV speakers minimize dependency length by considering only a limited search space of constituent orders, likely to conserve resources within the bounds of rationality and good-enough processing.

(1) a. C_1 C_2 C_3 C_4 V

b. C_3 C_2 C_1 C_4 V

c. C_4 C_1 C_2 C_3 V

d. C_1 C_2 C_4 C_3 V

Constituent length (CL) was calculated by counting words in a constituent (C_i), and the **total dependency length** (Total DL) of a sentence by summing the distances (in terms of words) between all head-dependent pairs in a dependency tree.

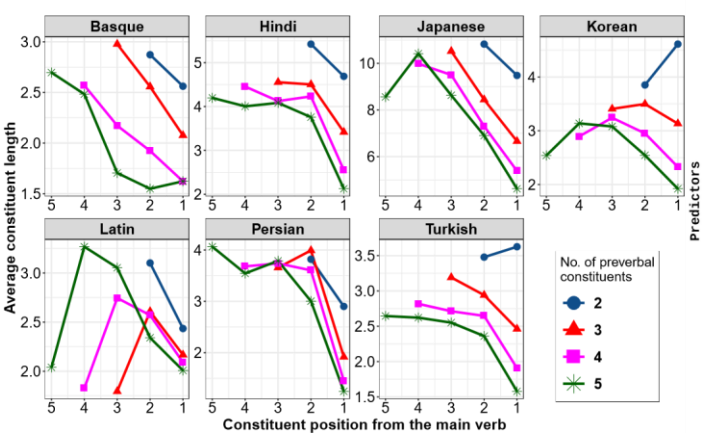


Figure 1: Average length of preverbal constituents in the corpus sentences containing only-2 to only-5 preverbal constituents

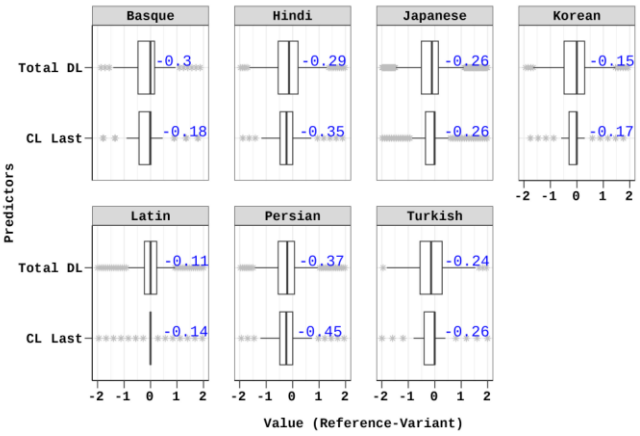


Figure 2: Summary of dataset denoting difference between predictor values of reference and the paired variant sentences; Mean difference values annotated inside the subplot

Language	CL Last	Total DL	Total DL + CL Last	%Same (%Correct)
	Classification Accuracy (%)			Prediction (CL Last *vs.* Total DL)
Basque	55.07	61.71	62.01	80.40 (48.59)
Hindi	69.49	63.39	69.23	75.03 (53.97)
Japanese	62.80	63.09	64.36	75.47 (50.68)
Korean	56.92	55.11	56.44	76.11 (44.08)
Latin	51.48	48.51	49.55	79.60 (39.79)
Persian	74.57	69.04	75.17	68.69 (56.16)
Turkish	61.72	60.00	62.02	77.44 (49.58)

Table 1: Classification accuracy (%) of various models (10-fold cross-validation) with constituent length of last preverbal constituent (CL Last) and total dependency length (Total DL) as predictors

References

Ferreira, F., Bailey, K. G., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, 11(1):11–15.

Futrell, R., Gibson, E., and Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Gigerenzer, G., Hertwig, R. E., and Pachur, T. E. (2011). *Heuristics: The foundations of adaptive behavior.* OUP.

Ranjan, S. and von der Malsburg, T. (2024). Work smarter...not harder: Efficient minimization of dependency length in sov languages. In Samuelson, L. K., Frank, S., Toneva, M., Mackey, A., and Hazeltine, E., editors, *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*, Rotterdam, Netherlands.

Zeman, D., Nivre, J., et al. (2022). Universal dependencies 2.11. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

False Remembering Elicited by Disconfirmed Predictions: Do Semantic and Word Form Features Linger in Memory?

Celina Rolgeiser & Katja I. Haeuser (Saarland University, Saarbrücken, Germany)
celina.rolgeiser@uni-saarland.de

Prediction during language comprehension involves the pre-activation of expected words (Huettig et al., 2022) and their semantically-related (Federmeier & Kutas, 1999) and possibly word form-related neighbors (DeLong et al., 2019). If predictions are disconfirmed, expected words linger in memory and elicit false remembering of expected words and of semantically-related words (Haeuser & Kray, 2024; Hubbard & Federmeier, 2024; Hubbard et al., 2019). False remembering of word form-related words has not been found yet. The lack of a word form-related effect might be due to the manipulation of word form similarity at word offset (Haeuser, 2022). Since there are inconsistent results regarding prediction-related pre-activation of word form features, which might be attributable to different manipulations of word form similarity (i.e., onset vs. offset, Li et al., 2022), manipulating word form similarity at word onset might more readily elicit false remembering. In line with this, the cohort-model of word recognition suggests stronger activation of onset- than offset-related neighbors (Simmons & Magnuson, 2018).

Here, participants (n = 142, m = 43, f = 96, nb = 3, *M* = 23.3 years old, range = 18 - 34 years old) read highly constraining sentences which ended with an unexpected word. After a 10-minute retention interval, participants were presented with single words and indicated whether the word was “old” or “new”. To additionally measure qualitative differences in recognition memory, participants indicated for old judgements whether they remembered details of the encoding phase (i.e., recollection) or just had a familiar feeling about having read this word (i.e., familiarity, Yonelinas, 2002). Presented words were old (e.g., “Uhr”, “clock”), new (e.g., “Fisch”, “fish”), expected but disconfirmed (e.g., “Reifen”, “tires”) and semantically- (e.g., “Auto”, “car”) and word form-related words (e.g., “Reihen”, “series”) to these expected words. Participants also completed a test battery of individual difference tests.

According to the results, expected and semantically-related words elicited higher levels of false remembering than new words (see Figure 1), and expected words elicited more recollection judgements than semantically-related words, suggesting a false memory effect over and above a backward semantic context association effect (see Figure 2). Surprisingly, word form-related words elicited less false remembering than new words (see Figure 1), especially in familiarity judgements (see Figure 2). We hypothesized that this effect was driven by prior suppression of onset-related neighbors (i.e., predicting “Reifen” inhibits “Reihen”, Haeuser & Borovsky, 2024). Indeed, in an exploratory analysis there was a correlation of the false alarm rate of word form-related words and inhibitory control (operationalized as *d’* which reflects the scaled hit rate minus the scaled false alarm rate of a go/nogo inhibition task). However, a similar correlation was found for new words, suggesting that the inhibitory control effect was not specific to word form-related words (see Figure 3).

In sum, we replicated effects of lingering predictions for expected and semantically-related words. Word form-related words do not elicit false remembering, possibly because they become suppressed during initial activation of the expected word.

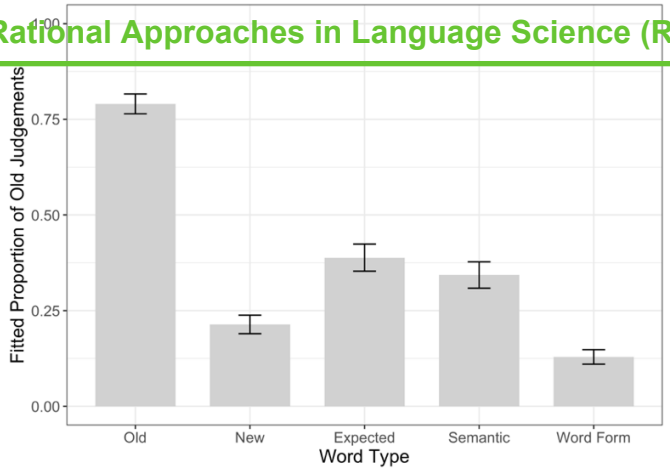


Figure 1. Fitted proportion of old judgements (aggregating over recollection and familiarity judgements) across the different word types

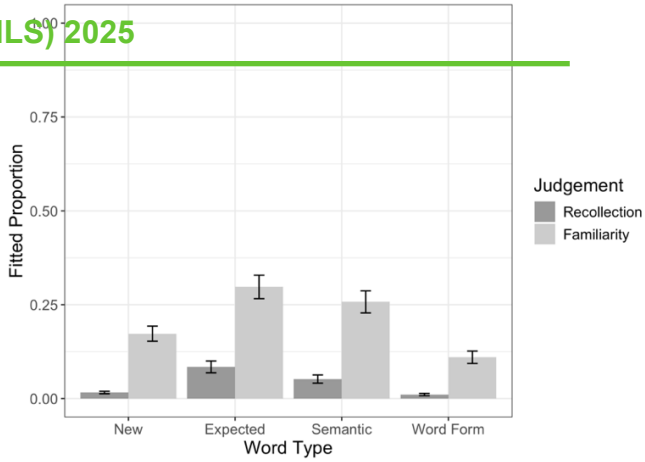


Figure 2. Fitted proportion of recollection and familiarity judgements across new, expected, semantically- and word form-related words

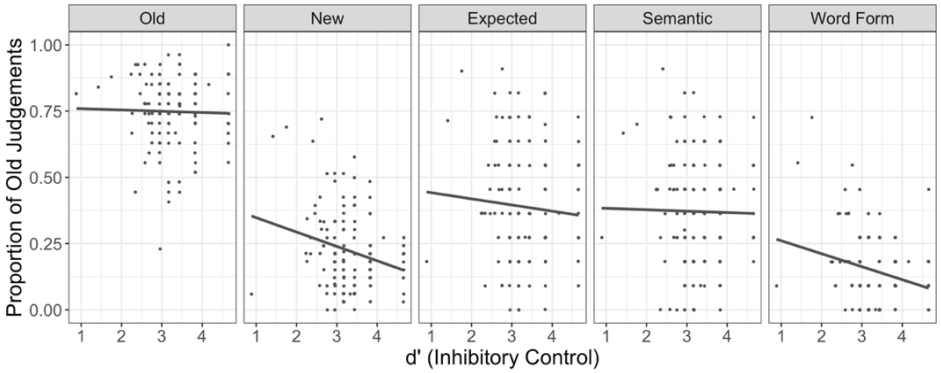


Figure 3. Correlation of the proportion of old judgements of each word type and inhibitory control
Note. Higher *d'* indicate better inhibitory control.

References

DeLong, K. A., Chan, W. H., & Kutas, M. (2019). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology*, *56*(4), e13312. https://doi.org/10.1111/psyp.13312

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469-495. https://doi.org/10.1006/jmla.1999.2660

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General*, *136*(3), 389-413. https://doi.org/10.1037/0096-3445.136.3.389

Haeuser, K. I., & Borovsky, A. (2024). Predictive processing suppresses form-related words with overlapping onsets. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*. https://escholarship.org/uc/item/95w210ck

Haeuser, K. I., & Kray, J. (2022). Uninvited and unwanted: False memories for words predicted but not seen. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*. https://escholarship.org/uc/item/7w22b8gm

Haeuser, K. I., & Kray, J. (2024). Age differences in context use during reading and downstream effects on recognition memory. *Psychology and Aging*, *39*(7), 715-730. https://doi.org/10.1037/pag0000845

Hubbard, R. J., & Federmeier, K. D. (2024). The Impact of Linguistic Prediction Violations on Downstream Recognition Memory and Sentence Recall. *Journal of Cognitive Neuroscience*, *36*(1), 1-23. https://doi.org/10.1162/jocn_a_02078

Hubbard, R. J., Rommers, J., Jacobs, C. L., & Federmeier, K. D. (2019). Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. *Frontiers in Human Neuroscience*, *13*, 291. https://doi.org/10.3389/fnhum.2019.00291

Huettig, F., Audring, J., & Jackendoff, R. (2022). A parallel architecture perspective on pre-activation and prediction in language processing. *Cognition*, *224*, 105050. https://doi.org/10.1016/j.cognition.2022.105050

Li, X., Li, X., & Qu, Q. (2022). Predicting phonology in language comprehension: Evidence from the visual world eye-tracking task in Mandarin Chinese. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(5), 531-547. https://doi.org/10.1037/xhp0000999

Simmons, E. S., & Magnuson, J. S. (2018). Word length, proportion of overlap, and phonological competition in spoken word recognition. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (p. 1062-1067). Madison, WI.

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*(3), 441-517. https://doi.org/10.1006/jmla.2002.2864

Effects of linguistic context and reading abilities on comprehension of unknown words

Margarita Ryzhova, Emilia Ellsiepen, Katharina Trinley, Iza Škrjanec, Vera Demberg (Saarland University, Germany)
mryzhova@lst.uni-saarland.de

Words that are unfamiliar to us can elicit processing difficulties. Word familiarity can be modulated by the intrinsic properties of the word, like frequency and length [2, 6]. However, the literature shows that the context also affects comprehension [3, 4, 7]. For example, scientific or technical texts may contain more specialized vocabulary that is unfamiliar to the general reader. In contrast, everyday texts such as newspapers or novels may contain more familiar language. In such common contexts, the reader can be surprised to encounter an unknown word or attribute it to a typo, while in a more scientific context, the reader might expect to encounter special domain terms they don't know. On the other hand, substantial evidence indicates that reading comprehension is influenced by the reader's literacy. More skilled readers are better at monitoring their comprehension, recognizing when additional processing is necessary, and are more motivated to fully understand the texts they read, thereby investing greater cognitive effort [1, 5].

In our studies on processing unknown words in German, we manipulate the type of context to explore its effect on readers' sensitivity to unfamiliar words. Additionally, we assess each participant's reading proficiency to investigate potential interactions between context and literacy (ART and vocabulary size). We conducted two self-paced reading experiments (with two sets of materials that were only partially overlapping) and asked participants to read texts for comprehension. Each text includes a target word: either a *real word* or a *pseudoword*. The target words were embedded into two types of contexts: *everyday* and *scientific*, making both studies follow a 2x2 design. Everyday stories concern familiar events from daily life (e.g., children playing in a park), while scientific stories occur in less common settings with characters with a specialized profession (e.g., researchers conducting experiments in a laboratory). The scientific stories themselves are not expository texts but rather narratives describing a less familiar scenario.

Our results confirm that readers are sensitive to pseudowords in *everyday* and *scientific* contexts, leading to increased reading times. However, evidence across the two studies is mixed regarding whether the context influences the processing of unknown words – see Table 1 for model specifications and results. Overall, the trend indicates that pseudowords are read more slowly in *everyday* than in *scientific* context, which may suggest that unknown words, despite their lack of a defined meaning, are more expected in domain-specific texts than in general narratives, resulting in faster reading. This effect, however, was only significant in one of the studies. We also find that high-literacy readers take more time to process pseudowords than low-literacy readers, regardless of context. This may reflect a greater effort by high-literacy readers to understand and integrate unfamiliar words into the context.

At the time of abstract submission, we are collecting data for an eye-tracking counterpart of this study. Our motivation is that eye-tracking can provide deeper insights into the processing of pseudowords and the nature of reading times through regressions and second-pass fixation durations. For literacy, early eye-tracking measures may reveal that high-literacy readers recognize pseudowords more quickly, while later measures may indicate that these readers invest more effort in integrating unknown words into the context.

	Critical region (Study 1)				Spillover region (Study 1)				Critical region (Study 2)				Spillover region (Study 2)			
Predictors	*Est*	*Std. Error*	*t-value*	*p*	*Est*	*Std. Error*	*t-value*	*p*	*Est*	*Std. Error*	*t-value*	*p*	*Est*	*Std. Error*	*t-value*	*p*
(Intercept)	795.91	24.76	32.15	**<0.001**	627.76	18.92	33.17	**<0.001**	843.46	19.80	42.61	**<0.001**	680.27	17.39	39.13	**<0.001**
Story (scientific)	-22.64	12.75	-1.78	0.077	-7.43	8.44	-0.88	0.379	12.61	14.81	0.85	0.395	-19.69	13.74	-1.43	0.152
Word (pseudo)	76.78	16.62	4.62	**<0.001**	92.68	11.95	7.76	**<0.001**	197.76	18.73	10.56	**<0.001**	233.57	17.20	13.58	**<0.001**
Story:Word	-12.89	13.18	-0.98	0.328	-18.28	8.08	-2.26	**0.024**	-23.66	22.36	-1.06	0.290	-28.28	18.54	-1.53	0.128
Trial	-32.84	13.55	-2.42	**0.016**	-7.31	8.28	-0.88	0.378	-35.95	9.03	-3.98	**<0.001**	-13.68	6.02	-2.27	**0.023**
Word:Trial	-27.09	13.60	-1.99	**0.047**	-2.44	8.73	-0.28	0.780	-51.99	16.24	-3.20	**0.001**	-0.92	11.06	-0.08	0.934
Chunk	-15.49	19.53	-0.79	0.428	-4.88	10.76	-0.45	0.650	-21.75	11.57	-1.88	0.061	-5.91	8.85	-0.67	0.504
log(Frequency)	-7.56	16.89	-0.45	0.655	-16.77	11.74	-1.43	0.154	-14.62	11.39	-1.28	0.200	-3.33	4.15	-0.80	0.423
Num. Chars Chunk	50.23	19.46	2.58	**0.010**	12.19	13.40	0.91	0.363	11.50	11.27	1.02	0.308	3.61	10.61	0.34	0.734
Num. Words Chunk					9.72	13.53	0.72	0.473					4.93	9.83	0.50	0.616
ART									41.05	17.36	2.37	**0.018**	12.19	13.99	0.87	0.384
Vocab									-24.77	20.48	-1.21	0.227	17.61	14.48	1.22	0.224
Story:Vocab									1.61	17.53	0.09	0.927	3.62	11.03	0.33	0.743
Word:Vocab									2.86	16.23	0.18	0.860	54.51	14.70	3.71	**<0.001**
Story:Word:Vocab									23.77	22.24	1.07	0.285	8.82	17.09	0.52	0.606
Story:ART									-6.46	15.96	-0.40	0.686	-15.40	12.10	-1.27	0.203
Word:ART									26.06	15.98	1.63	0.103	-23.66	16.26	-1.46	0.146
Story:Word:ART									-22.14	33.71	-0.66	0.511	-9.68	16.18	-0.60	0.550

Table 1. Regression coefficients and test statistics from the Generalized Gamma mixed-effects models (with identity link) of reading times in critical and spillover regions in Studies 1 and 2.

References:

1. **Broek, P. v. d., White, M. J., Kendeou, P., & Carlson, S. (2009).** Reading between the lines: Developmental and individual differences in cognitive processes in reading comprehension. In R. K. Wagner, C. Schatschneider, & C. Phythian-Sence (Eds.), *Beyond decoding: The behavioral and biological foundations of reading comprehension* (pp. 107–123). The Guilford Press.
2. **Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004).** Length, frequency, and predictability effects of words on eye movements in reading. European journal of cognitive psychology, 16 (1-2), 262–284. https://doi.org/10.1080/09541440340000213
3. **Lowell, R., Morris, R.K. (2014).** Word length effects on novel words: Evidence from eye movements. Atten Percept Psychophys 76, 179–189. https://doi.org/10.3758/s13414-013-0556-4
4. **Nieuwland, M. S., & Van Berkum, J. J. (2006).** When peanuts fall in love: N400 evidence for the power of discourse. Journal of cognitive neuroscience, 18(7), 1098–1111. https://doi.org/10.1162/jocn.2006.18.7.1098
5. **Oakhill, J.V., Hartt, J., & Samols, D. (2005).** Levels of Comprehension Monitoring and Working Memory in Good and Poor Comprehenders. Reading and Writing, 18, 657-686. https://doi.org/10.1007/s11145-005-3355-z
6. **Rayner, K. (1998)**. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372
7. **Williams, R. S., & Morris, R. K. (2004).** Eye movements, word familiarity, and vocabulary acquisition. European Journal of Cognitive Psychology, 16(1-2), 312–339. https://doi.org/10.1080/09541440340000196

Gaze and Pupil Indices of Rational Reference Production

Elli Tourtouri, Mitra Gholami, Nicole Gotzner (Osnabrück University)
Elli.tourtouri@uni-osnabrueck.de

Speakers often encode more information than necessary for specifying referents in the immediate visual context [1], defying Grice's Quantity [2]. Why speakers engage in such (seemingly) irrational behaviour has been the subject of much debate: Some studies suggest that production choices are mainly motivated by a concern to ease planning effort (Egocentric view [3-4]), while others support the view that speakers aim at producing utterances that are efficient given the conditions (Audience-design view [5-8]). Previous work has largely considered the influence of redundancy on listeners' comprehension as an index of production strategy, and linked audience design to increased cognitive effort for the speaker. In this study, we use eye-tracking to directly assess production strategies and cognitive effort. In a referential communication experiment, we examine whether speakers' gaze patterns before speech onset differ based on their production strategy, and whether audience design is associated with increased cognitive effort (measured as pupil dilations).

In a 2x3 within-participants design, we manipulate which adjective is *necessary* for specifying the target referent (colour or pattern), and which adjective *reduces* the uncertainty about the target referent (*referential entropy*) to a greater degree (colour, pattern, equal) (see [8]). We employ 36 experimental items comprising 6 visual displays each (one per condition). In each display, 6 objects are arranged in different configurations based on the condition (Fig.1), while the target object remains constant. In the fillers (N=108), either two or no adjectives are necessary. We control for the perceptibility of the distinguishing feature for each target object (see [9]), based on ratings obtained in an online norming study.

The procedure is as follows: Two participants, randomly assigned the roles of Speaker and Listener, engage in a referential communication game, while the Speaker's eye movements are tracked. Participants' task is to identify whether the objects are arranged in the same configuration on both their screens. The Speaker is instructed to ask questions about the horizontal position of an object marked as the target only on their screen (e.g., 'Is the blue ball on the left?' in Fig. 1a). The Listener has to respond with a 'yes' or 'no' by pressing a key in a keyboard in front of them. In half of the trials, the Listeners see a mirror version of the Speaker display. We plan to collect data from 48 pairs. The experiment is conducted in German.

Participants will be grouped based on their use of redundant adjectives (see [8]). We will analyse the proportions of referentially redundant utterances (binary coded, with GLMMs), and participants' pupil dilations in two time-windows: before and after the reveal of the target referent. We will also use LMMs to analyse proportions of inspections to the referents before and after the reveal of the target object, and log gaze probability ratios for fixations to the target (e.g., the blue ball in Fig.1a) vs. the contrast (e.g., the green ball in Fig.1a) or competitor (e.g., the blue mitt in Fig.1a) objects in the interval between the reveal of the target referent and speech onset. All models will include 'Necessary adjective' and 'Entropy-reducing adjective' as fixed effects, the perceptibility score as a control factor, and the maximal random effects structure [10]. Models analysing eye-tracking measures will additionally include 'Speaker Group' as a between-subjects factor. We generally expect that, compared to speakers using an audience-design strategy, egocentric speakers will scan the visual scene less broadly before the reveal of the target referent and expend less cognitive effort.

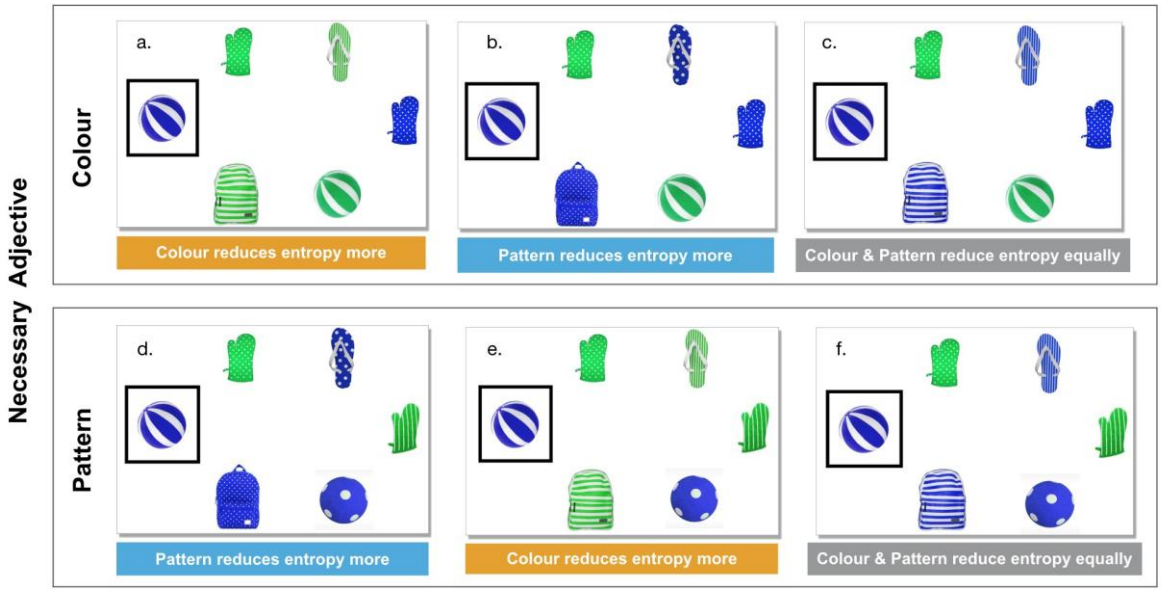


Figure1. Sample visual displays per condition. In the top panels, *colour* is necessary for specifying the target referent; in the bottom panels; *pattern* was the necessary feature. In the panels with the yellow label (a and e), *colour* is the more entropy reducing adjective, while in the panels with the blue label (b and d), *pattern* is the more entropy reducing adjective. In the panels with the grey label (c and f), both adjectives are equally entropy-reducing. The black frame indicates the target object and appears only on the Speaker's screen 2s picture onset. Figure is adapted from [8].

References

[1] Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics: Speech acts (Vol. III, pp. 41–58*). New York: Academic.

[2] Davies C. & J. E. Arnold (2019). Reference and informativeness. In C. Cummins and N. Katsos (Eds.), *The Oxford Handbook of Experimental Pragmatics and Semantics*, pp. 29–88. Oxford, UK: Oxford University Press.

[3] Keysar, B., D. J. Barr & W. S. Horton (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science*, 7(2), 46–49.

[4] Engelhardt, P., K. Bailey & F. Ferreira (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54(4), 554–573.

[5] Arts, A., A. Maes, L. Noordman & C. Jansen (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1), 361 – 374.

[6] Fukumura, K. & M. N. Carminati (2022). Overspecification and incremental referential processing: An eyetracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(5), 680–701.

[7] Rubio-Fernandez, P., F. Mollica & J. Jara-Ettinger (2021). Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *Journal of Experimental Psychology: General*, 150(3), 583–594.

[8] Tourtouri, E. N., F. Delogu, L. Sikos & M. W. Crocker (2019). Rational over-specification in visually-situated comprehension and production. *Journal of Cultural Cognitive Science,* 3(2), 175–202.

[9] Kursat, L. & J. Degen (2021). Perceptual Difficulty Differences Predict Asymmetry in Redundant Modification With Color and Material Adjectives. *Proceedings of the Linguistic Society of America,* 6(1), 676–688.

[10] Barr, D. J., R. Levy, C. Scheepers & H. J. Tily (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255-278.

Comprehension of Idiomatic Expressions in Low-Literacy Readers of Easy German: An Experimental Investigation

Lena Wieland (Saarland University), Ingo Reich (Saarland University)
lena.wieland@uni-saarland.de
i.reich@mx.uni-saarland.de

We present a planned experimental investigation into the comprehension of idiomatic expressions in *Leichte Sprache* (Easy German), a simplified variety of German aimed at enhancing accessibility by reducing the complexity of vocabulary, morphology, and syntax (Bredel and Maass, 2016). Easy German primarily supports low-literacy readers and individuals with cognitive impairments (Bock and Pappert, 2023), populations that often struggle with figurative language. Despite recommendations to minimize idiomatic expressions in Easy German, little empirical research has examined how these populations actually process idiomatic language in simplified contexts.

We address this gap by examining the effects of literacy, transparency, and literal plausibility on the comprehension of 24 idiomatic expressions, such as *jemandem einen Bären aufbinden* (fig.: to pull sb.'s leg, lit.: to tie a bear on someone). Idioms were selected from a prior rating study involving 30 participants without cognitive impairments, ensuring a balanced distribution across two key dimensions: transparency, the degree to which the idiomatic meaning can be inferred from its literal components (transparent vs. non-transparent), and plausibility, reflecting the extent to which the idiom appears contextually or logically reasonable (plausible vs. non-plausible).

In our main study, we aim to examine the factors that influence idiom comprehension in low-literacy readers within a simplified linguistic context. Additionally, we will explore whether participants exhibit a preference for literal over figurative expressions. We hypothesize that (i) higher literacy correlates with better comprehension; (ii) transparency impacts comprehension, making transparent idioms easier to interpret; (iii) literal plausibility affects preference, with highly plausible idioms ranked lower than literal alternatives; and (iv) interactions between transparency and plausibility modulate comprehension.

The main study will involve a cohort of 31 low-literacy readers, whose literacy levels are assessed using the *lea.diagnostik* online tool (Koppel and Wolf, 2014). The participants will complete two tasks per idiom: a multiple-choice task to assess recognition within context, followed by a ranking task to evaluate preferences for idiomatic versus literal interpretations. To ensure timely execution of the experiment before the conference, we will establish a timeline that includes finalizing the study design by November, completing data collection by early January, and preparing findings for presentation by late January.

Our preliminary findings from a GLMM analysis (Knudson, 2024) of low-literacy participants interpreting idioms without context in a sentence interpretation task indicate that literacy score significantly predicts accuracy, ($b = 0.0313$, $SE = 0.0084$, $z = 3.72$, $p < 001$), suggesting better comprehension with higher literacy (Fig. 1). Additionally, plausibility also predicts accuracy, ($b = 0.4367$, $SE = 0.1646$, $z = 2.65$, $p = .008$). A significant interaction between transparency and plausibility was observed, ($b = -1.3610$, $SE = 0.3304$, $z = -4.12$, $p < .001$), indicating that idioms characterized by high transparency and high plausibility are particularly challenging for participants (Fig. 2).

By examining idiom comprehension in Easy German, our research aims to offer novel insights into how figurative language can be adapted for low-literacy populations, informing educational and communicative strategies for cognitively impaired readers.

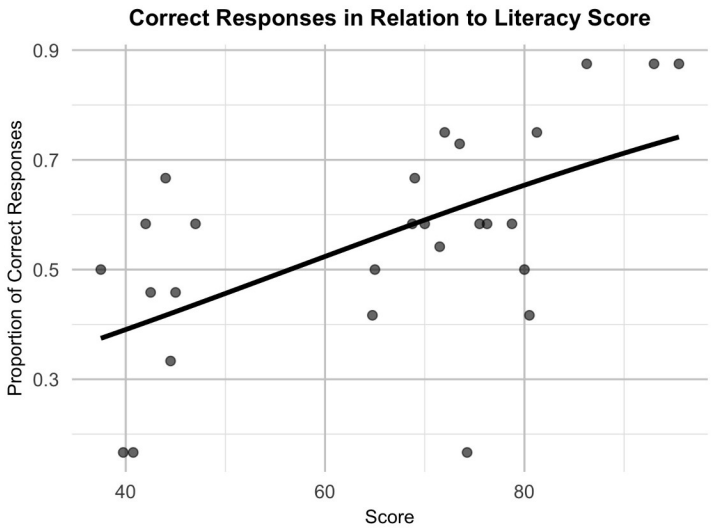


Figure 1

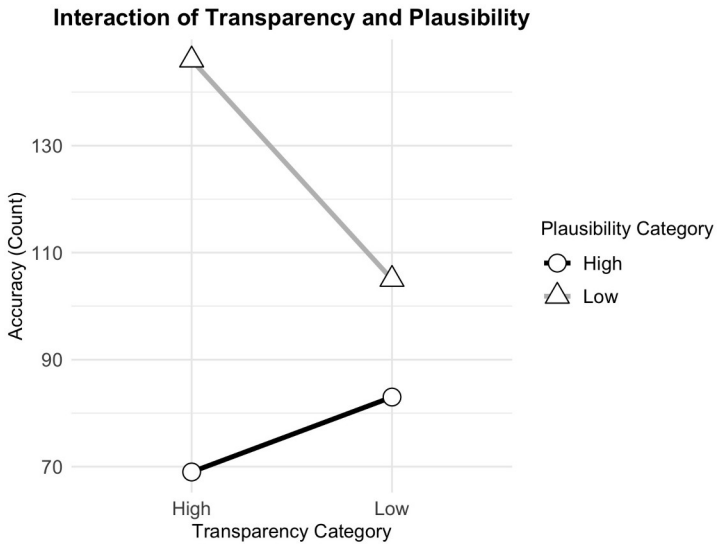


Figure 2

References

Bock, B. M., & Pappert, S. (2023). Leichte Sprache, Einfache Sprache, verständliche Sprache. narr/francke/attempto.

Bredel, U., & Maass, C. (2016). Leichte Sprache: Theoretische Grundlagen, Orientierung für die Praxis [OCLC:ocn949848478]. Dudenverlag.

Knudson, C. (2024). Glmm: Generalized linear mixed models via monte carlo likelihood approximation [R package version 1.4.5]. https://CRAN.R-project.org/package=glmm

Koppel, I., & Wolf, K. D. (2014). Otu.lea. eine niedrigschwellige online-diagnostik für funktionale analphabetinnen in der kursarbeit. Alfa-Forum, (86), 28–31.

L1 and L2 Speakers' Performance in Receptive Multilingualism

Wei Xue & Bernd Möbius (Saarland University)
weixue@lst.uni-saarland.de

Introduction: Receptive multilingualism allows speakers to comprehend utterances in a foreign language (e.g., Dutch) using a known language (e.g., English), facilitated by similarities in their vocabulary and pronunciation. However, the development of sound categories in the known language varies between L1 and L2 speakers (Lecumberri et al., 2010). L2 speakers may have less clearly defined or even absent sound categories (Scharenborg & van Os, 2019), leading to less accurate phoneme recognition or auditory word recognition (AWR), particularly in adverse contexts. Consequently, L1 and L2 speakers of the known language may exhibit different levels of comprehension when processing utterances in the foreign language. In this study, we investigate how L1 and L2 speakers process and adapt language to different contexts in receptive multilingualism. We focus on accuracy and reaction time in AWR tasks, exploring the effects of phonological and semantic similarities on the AWR of English words in Dutch-English prime-target pairs. In addition, we examine how the processes differ across various listening conditions. We specifically compare L1 English speakers and L2 English learners with L1 Chinese backgrounds.

Methods: To address the investigation, we conducted web-based experiments on lexical decision tasks in a priming paradigm following that of Kudera et al. (2021). Specifically, to examine priming effects, we introduced four types of word pairs (cognates, false friends, translation equivalents, and fillers) that differ in the degree of similarity in phonological forms and semantics as shown in Figure 1. The contrasts between them aim to determine how the lack of semantic similarity (i.e., false friends vs cognate) and phonological overlap (i.e., translation equivalents vs cognate) impacts AWR. To study the effect of listening conditions, we studied these pairs in five listening conditions including quiet, white noise and babble noise. We had two versions of the experiments differing in the order of conditions as shown in Figure 1. We used *glmer* and *lmer* models in *lme4* and *lmerTest* R packages to study the effects of the treatment-coded contrasts between different prime-target pairs, listening conditions, versions, L1/L2, and their interactions in predicting response correctness and reaction time, respectively as shown in Figure 1.

Results and Conclusion: We recruited 84 L1 speakers and 46 L2 speakers of self-reported intermediate or higher level via Prolific. Figure 2 shows the mean and error bars for accuracy (left panel) and for reaction time of correct responses (right panel). The prediction results showed a significantly lower accuracy of L2 than L1 (β = 0.779023, SE = 0.188443, z = 4.134, p < 0.0001) but a null effect of L1/L2 (L1_v_L2) on reaction time (β = -63.222, SE = 57.799, df = 126.805, t = -1.094, p = 0.37837). Also, as expected, L1/L2 showed significant interactions with listening conditions, such as with quiet vs. noise contrast (Q_v_N) for both accuracy and reaction time, suggesting their different performance when noise exists. We found a significant interaction between L1/L2 and cognate vs. false-friend (CG_v_FF) contrasts for accuracy, indicating different effects of lacking semantic similarity between L1 and L2. We also found a significant interaction between L1/L2 and cognate vs. filler (CG_v_FL) contrasts for reaction time, suggesting that L2 speakers appear to be confused with filler words compared to L1 speakers. Note that significance values were corrected based on Benjamini-Hochberg method. Overall, L2 speakers seem to suffer more in more adverse contexts either via listening conditions or linguistic context. Further analyses are necessary to reveal the differences between them.

- Experimental setting:

Listening condition:	1	2	3	4	5
Version 1:	White Noise SNR=0 dB	White Noise SNR=-6 dB	Quiet	Babble Noise SNR=0 dB	Babble Noise SNR=-6 dB
Version 2:	Babble Noise SNR=0 dB	Babble Noise SNR=-6 dB	Quiet	White Noise SNR=0 dB	White Noise SNR=-6 dB

- Stimuli examples:

Type of word pair	Phonological similarity (similar sound)	Semantic similarity (same meaning)	Example	
			Dutch	English
cognate	yes	yes	arm /ɑrm/	arm /ɑːm/
false friend	yes	no	wet /wɛt/ (means *law*)	wet /wɛt/
translation equivalent	no	yes	fiets /fits/	bike /baɪk/
filler	no	no	prent /prɛnt/ (means *print*)	liss /lɪs/

- Statistical model formular and coded contrasts:

CG_v_FF * Q_v_N + CG_v_FF * B_v_W + CG_v_FF * Bz_v_Bs + CG_v_FF * Wz_v_Ws + CG_v_TE * Q_v_N + CG_v_TE * B_v_W + CG_v_TE * Bz_v_Bs + CG_v_TE * Wz_v_Ws + CG_v_FL * Q_v_N + CG_v_FL * B_v_W + CG_v_FL * Bz_v_Bs + CG_v_FL * Wz_v_Ws + L1_v_L2 * Q_v_N + L1_v_L2 * B_v_W + L1_v_L2 * Bz_v_Bs + L1_v_L2 * Wz_v_Ws + V1_v_V2 * Q_v_N + V1_v_V2 * B_v_W + V1_v_V2 * Bz_v_Bs + V1_v_V2 * Wz_v_Ws + (1|Participant_id) + (1|audiofile_nr)

Word type	CG_v_FF	CG_v_TE	CG_v_FL
CG	0	0	0
FF	-1	0	0
TE	0	-1	0
FL	0	0	-1

Version	V1_v_V2
V1	0
V2	-1

Language	L1_v_L2
L1	0
L2	-1

Condition	Q_v_N	B_v_W	Bz_v_Bs	Wz_v_Ws
1	-1	-0.5	0	-0.5
2	-1	-0.5	0	0.5
3 (Quiet)	0	0	0	0
4	-1	0.5	-0.5	0
5	-1	0.5	0.5	0

Note. Taking Version 1 as an example.

Figure 1: Experimental setup, examples of stimuli for the four types of words, model formular and the contrasts. Note that the words for English in the fillers are not meaningful, existing words. The variables in the formular are as follows: CG, FF, TE, and FL refer to word types of cognate, false friend, translation equivalent, and filler; Q, N, B, W, Bz, Bs, Wz, and Ws refer to listening conditions of Quiet, Noise, Babble noise, White noise, Babble noise with SNR = 0 dB, Babble noise with SNR = -6 dB, White noise with SNR = 0 dB, and White noise with SNR = -6 dB; V1 and V2 refer to Versions 1 and 2. The contrasts were treatment-coded with cognate, quiet, v1, and L1 as the baselines.

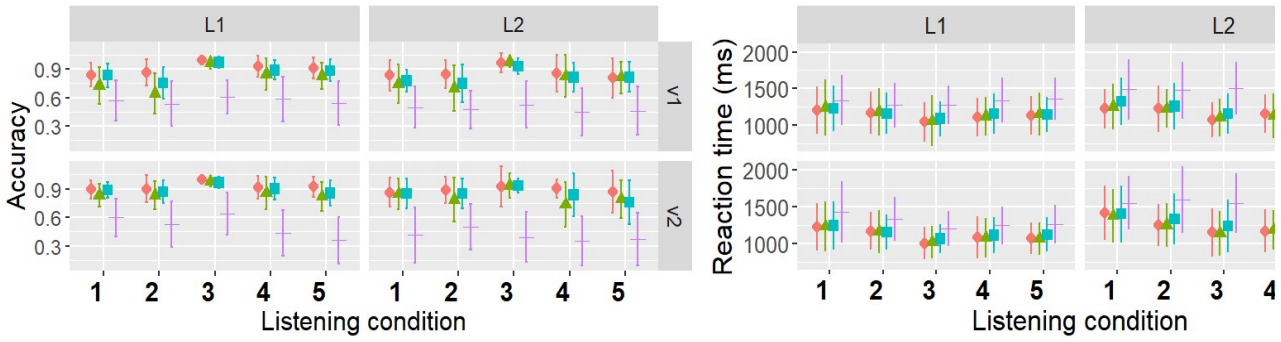


Figure 2: Accuracy (left panel) and reaction time (right panel) plots for the four types of word pairs in five listening conditions. Both L1 and L2 speakers are shown for the two versions of experiments (v1 and v2). The 1-5 on x-axes for v1 (Version 1) and v2 (Version 2) can be found in Figure 1 with 3 refering to the Quiet condition.

References:

Kudera, J., Georgis, P., Möbius, B., Avgustinova, T., & Klakow, D. (2021). Phonetic Distance and Surprisal in Multilingual Priming: Evidence from Slavic. In *Proc. INTERSPEECH 2021 – 22rd Annual Conference of the International Speech Communication Association*, 3944–3948.

Lecumberri, M. L. G., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech communication*, 52(11-12), 864–886.

Scharenborg, O., & van Os, M. (2019). Why listening in background noise is harder in a non-native language than in a native language: A review. *Speech Communication*,108(2), 53–64.

Is frequency effect on phonological and phonetic encoding word-based or syllable-based?

Yuen, I. (ivyuen@lst.uni-saarland.de), Andreeva, B., Ibrahim, O., Möbius, B
Saarland University

Linguistic predictability is pervasive and has been shown to influence acoustic realization (e.g., Arnon & Cohen-Priva, 2013; Aylett & Turk, 2004). Yet these effects have been mostly focused on a specific linguistic level, rather than across levels. One type of predictability is 'frequency of occurrence', which could occur at the level of word (word frequency) and syllable (syllable frequency). Word frequency effect is assumed to arise from the ease of retrieval, leading to fast response latency (RT). So is syllable frequency effect, arising from a postulated mental syllabary that mediates phonological and phonetic encoding (e.g., Cholin et al., 2006). Although frequency effects manifest in RT and acoustic realization, few studies examine both measures to get a better understanding of the processes that RT and acoustics reflect during phonological and phonetic encoding.

To investigate this issue, the current study examined the effect of high vs. low frequently occurring monosyllabic (e.g., *Kind* gloss: child vs. *Gift* gloss: poison) and disyllabic words (e.g., *Fehler* gloss: mistake vs. *Feder* gloss: feather) containing a short or long vowel in a stressed syllable in German. Word and syllable frequency were estimated from CELEX and SUBTLEXDE. High frequency monosyllabic words were manipulated to have low syllable frequency, whereas high frequency disyllabic words covary with stressed syllable frequency. Twenty monolingual German adults were instructed to formulate a verbal sentence in response to an auditory prompt question, incorporating the target stimulus label. Each target stimulus was elicited in 2 utterance positions: medial vs. final. We expect short acoustic vowel duration and fast RT in high frequently-occurring word and syllable. If word and syllable frequency conflict with one another, a serial encoding of phonological/phonetic processes will lead to longer RT for low frequency syllables. Two measures were taken and analyzed using lmer (Bates, 2015) in R (R Core, 2022): vowel duration and RT as measured from the beginning of the prompt question to the beginning of the verbal response.

Results of vowel duration revealed a significant 4-way interaction with Word frequency, Number of syllables, Utterance position and Vowel type ($F = 5.1$, $df = 1, 1579$, $p = .02$*). Separate analyses revealed Word frequency effect on short vowels in utterance-final monosyllabic words, but no overall effect in disyllabic words. Low frequency monosyllabic words, despite conflicting high syllable frequency, increase duration of short vowels (Fig. 1). Results of RT revealed significant main effects: Word frequency ($F = 4.5$, $df = 1, 22$, $p = .04$*), Sentence duration ($F = 19.6$, $df = 1, 1233$, $p < .0001$***), and a significant Number of syllable-by-Utterance position interaction ($F = 5$, $df = 1, 21$, $p = .04$*). The number of syllables affects RT as a function of word frequency, with the effect in monosyllabic words (Fig. 2). In light of the interaction, RT is faster for low frequency monosyllabic words which have conflicting high syllable frequency, suggesting RT to be primarily driven by syllable frequency during phonological/phonetic encoding. In sum, acoustic duration and RT reflect separate frequency effects at word and syllable level.

Figure 1. Vowel duration (ms) in disyllabic and monosyllabic words containing a long or short vowel with high vs. low word frequency in utterance-final vs. medial positions, with +/- 1 SD

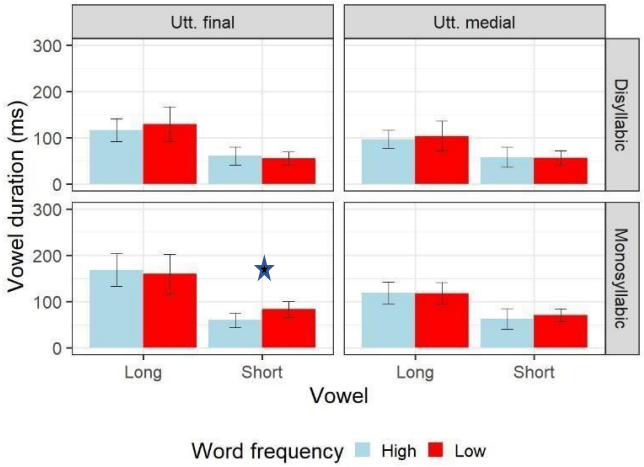
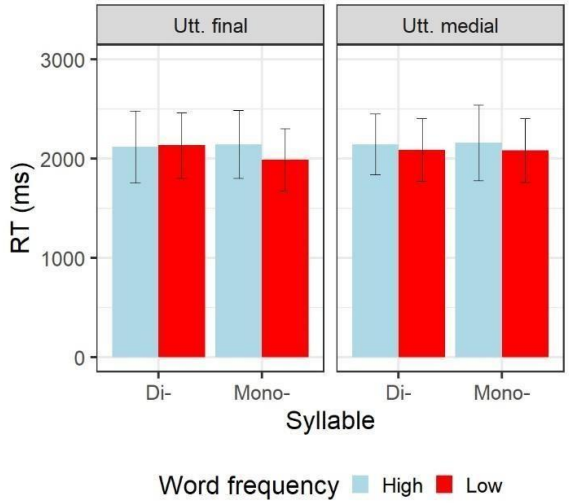


Figure 2. RT (ms) in disyllabic and monosyllabic words with high vs. low word frequency in utterance-final vs. medial positions, with +/- 1 SD



References:

Arnon, I., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, 56 (3), 349-371.

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence duration in spontaneous speech. *Language and Speech*, 47, 31-56.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Cholin, J., Levelt, W.J.M., & Schiller, N.O. (2006). Effects of syllable frequency in speech production. *Cognition*, 99, 205-235.

R Core Team (2022). R: A Language and Environment for Statistical Computing

From Moments to Memories: Unveiling the Role of Event Boundaries in Narratives

Doruntinë Zogaj, Regine Bader, Axel Mecklinger (Saarland University)
doruntine.zogaj@uni-saarland.de

Our daily lives unfold continuously, yet when we think about the past, we tend to organize our memories into distinct and cohesive events.
An influential framework that provides an explanation of how continuous daily live activity is segmented into meaningful subunits to guide attention and memory is Event Segmentation Theory (EST). According to EST, within a continuous stream of information, people can detect transitions between events, known as event boundaries, which naturally segment the stream into discrete and meaningful events (Kurby, & Zacks, 2008). This segmentation can have wide-ranging cognitive consequences, for instance support for the encoding and retrieval of episodic memories. Research has shown that items that belong to the same events are more likely to be recalled together (Shin and DuBrow, 2021), and that recency judgments are less accurate for items from different events (DuBrow and Davachi, 2013). In addition to the mnemonic effects for items within- and between-events, recent evidence suggests that the points in time constituting event boundaries are particularly well-represented in episodic memory. It is conceivable that increased attention at these points contributes to this memory advantage for event boundaries (Heusser et al., 2018).
In the present study, ERPs were employed to investigate the online processing of event boundaries during spoken language comprehension in narratives. We extended upon previous research by exploring whether the principles of predictive processing and its mnemonic consequences are applicable to larger and more naturalistic contexts. Participants listened to short stories, each consisting of five sentences describing a common activity (e.g., going to the supermarket). In the third sentence, a critical word was introduced, referring either to a predictable action (e.g., shopping) marking no boundary or to a less predictable action (e.g., reading) marking an event boundary. The fourth and fifth sentence reinforced the activity mentioned in the third sentence. EEG was recorded while participants listened to the sentences. In a subsequent memory test, conducted after every 17 to 18 stories across eight blocks, critical words from the stories both boundary and no-boundary words, were presented together with new words. Participants were asked to indicate which words were from the sentences they heard previously, using an old/new recognition memory task.
Although our results, suggest that there is no difference in memory performance between the boundary and no-boundary conditions, the ERP findings are intriguing. Consistent with Delogu et al. (2018), there was a larger N400 for the event boundary condition compared to the no-boundary condition, replicating their N400 effect using an ecologically more valid setting. Most importantly, ERPs recorded during the encoding of critical words were compared for critical words that were subsequently remembered versus those forgotten. Interestingly, critical words in the boundary condition elicited an increased N400 if they were subsequently remembered as compared to those that were forgotten. Notably, this effect was not observed in the no-boundary condition (see Figure 1). These results suggest that detecting a shift in the narrative structure at an event boundary initiates semantic processing that supports the formation of successful memories for upcoming events. Our findings provide new insights into how event boundaries during encoding segment a continuous experience into episodic events by shaping their subsequent representation in memory.

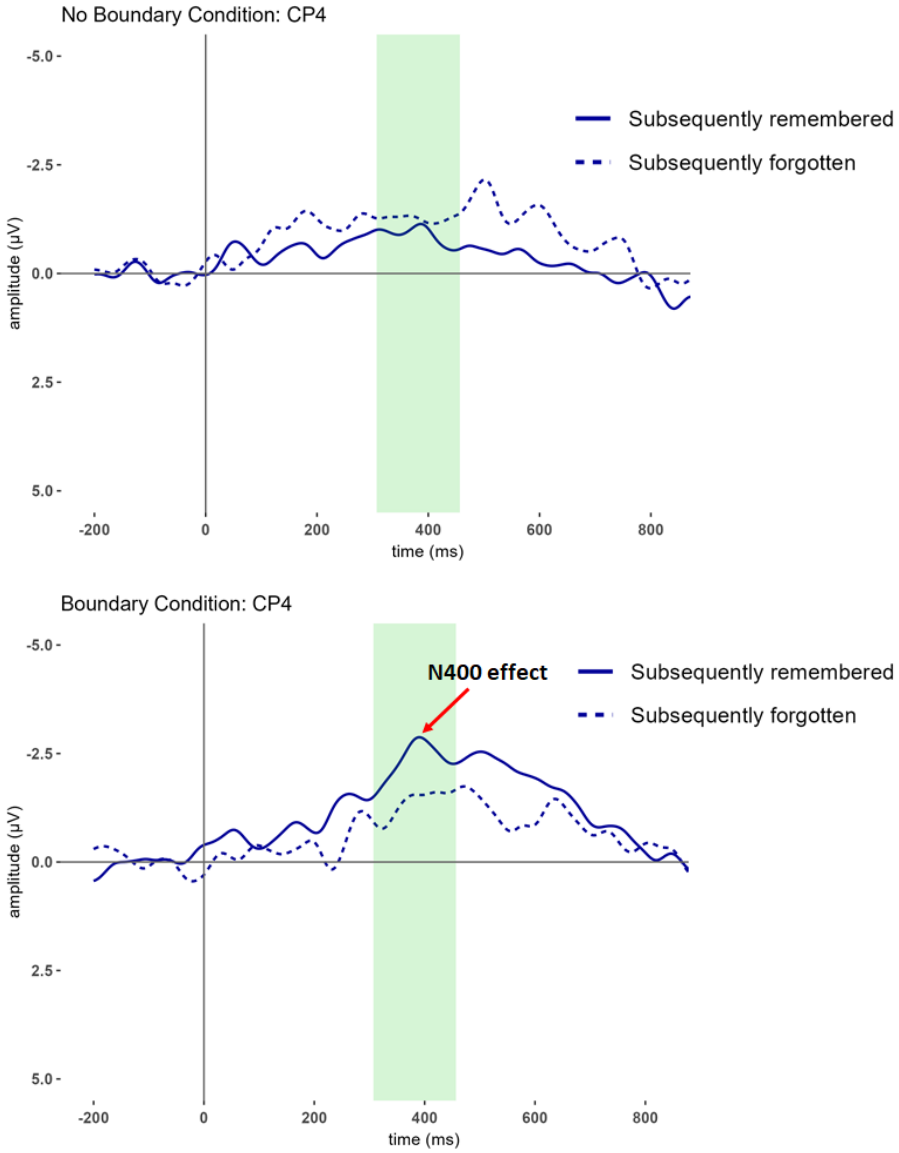


Fig 1. Grand-averaged ERP waveforms at electrode CP4 during the N400 time window, comparing the Subsequent Memory Effect (SME) between the Boundary and No-Boundary conditions. Time zero on the x-axis marks the onset of the critical words.

References

Delogu, F., Drenhaus, H., & Crocker, M. W. (2018). On the predictability of event boundaries in discourse: An ERP investigation. *Memory & Cognition, 46*(2), 315–325. https://doi.org/10.3758/s13421-017-0766-4

DuBrow, S., & Davachi, L. (2013). The influence of context boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General, 142*(4), 1277–1286. https://doi.org/10.1037/a0034024

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences, 12*(2), 72–79.

Heusser, A. C., Ezzyat, Y., Shiff, I., & Davachi, L. (2018). Perceptual boundaries cause mnemonic trade-offs between local boundary processing and across-trial associative binding. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(7), 1075–1090.

Shin, Y. S., & DuBrow, S. (2021). Structuring memory through inference-based event segmentation. *Topics in Cognitive Science, 13*(1), 106–127.

The Role of Linguistic and Conceptual Feature Properties in Native and Non-Native Processing and Memory

Bordag, D. (denisav@uni-leipzig.de)[1], Opitz, A.[1] [1]University of Leipzig

Previous research focusing on differences in memory of linguistic verbatim vs. content information suggests that memory of formal information (e.g., word order, grammatical voice) begins to decay rapidly shortly after processing (Garnham & Oakhill, 1996; Gurevich et al. ,2010), while propositional meaning is more likely to be retained in the long-term memory. In addition, non-native (L2) speakers manifested better memory for formal aspects of language than native (L1) speakers (Sampaio & Konopka, 2013; Bordag et al., 2021).

In the present study, *we explored and compared the processing and memory of features in the morphosyntactic domain.* We investigated how differently salient *formal and conceptual properties* contribute to retention of grammatical information during reading.

Empirically, we focused on the German number and tense features. Typically, differences in grammatical number also correspond to imageable conceptual differences in meaning. In generic contexts, however, generic singular marking refers to a whole class, and the morphosyntactic distinction between singular and plural does not correspond to meaning differences (see Table 1). In the German tense system (Table 2), similar (e.g., past in preterite and perfect) or different (e.g., present vs. preterite) temporal meanings can be expressed by similar or different formation types (analytically by affixation only vs. synthetically via auxiliaries).

Research Question: What is the role of conceptual and formal salience in processing and retention of morphosyntactic information (i.e., tense, number)? Is this different in native (L1) versus non-native (L2) processing?

Methods: 64 L1 German speakers and 64 L2 German learners (L1 Czech, B2-C1 proficiency) read single sentences that were repeated (12-16 intervening sentences) either identical or changed according to the experimental manipulations. Eye movements were tracked, and gaze durations measured and analyzed (mixed effects models).

Rationale: Participants' registering of grammatical changes (e.g., singular to plural, or preterite to perfect) should be reflected in longer fixation times at the changed regions compared to rereading an unaltered sentence, if they retained the grammatical/conceptual information from the first reading in memory.

Results (see also Figure 1): Non-native participants (L2) showed retention/registration effects (longer reading times in changed condition) only if the manipulation involved *salient formal changes* (affixation vs. analytic forms). No effects were observed for less prominent formal changes (i.e., affixation), irrespective of conceptual changes (e.g., number manipulation).

For native participants (L1), effects of retention were observed for grammatical changes that were related to imageable *conceptual differences* (e.g., specific readings in the number manipulation), or if both *prominent formal and conceptual changes* were involved at the same time (present↔perfect). Non-imageable alternations of tense that were not accompanied by salient formal changes were not registered (present↔preterite).

We conclude that formal aspects of grammatical features play a pivotal role for information processing and retention in L2. In L1, the impact of formal aspects on retention is less pronounced and comes into play only if it is accompanied by salient conceptual (functional) changes.

	Contrast in Meaning	Contrast in Formation	*Examples*
Specific	salient	similar	***Das Schwein wurde*** *beim Transport am Bauch verletzt.* ***Die Schweine wurden*** *beim Transport am Bauch verletzt.* 'The pig/pigs was/were injured in the abdomen during transport.'
Generic	similar	similar	***Der Elefant wird*** *in vielen Ländern immer noch illegal gejagt.* ***Die Elefanten werden*** *in vielen Ländern immer noch illegal gejagt.* 'The elephant/elephants is/are still hunted illegally in many countries.'

Table 1. Number Alternation: Examples

	Meaning	Formation	*Examples*
Present ↕ **Perfect**	different (present/ past)	different	*Der Chemiker* ***vermischt*** *vorsichtig die beiden Substanzen.* *Der Chemiker* ***hat*** *vorsichtig die beiden Substanzen* ***vermischt****.* 'The chemist carefully mixes / mixed the two substances.'
Present ↕ **Preterite**	different (present/ past)	similar	*Der Mechaniker* ***lagert*** *die Ersatzteile in der Garage.* *Der Mechaniker* ***lagerte*** *die Ersatzteile in der Garage.* 'The mechanic stores / stored the spare parts in the garage.'
Preterite ↕ **Perfect**	similar (past)	different	*Der Dirigent* ***eröffnete*** *das Dorffest mit einer Rede.* *Der Dirigent* ***hat*** *das Dorffest mit einer Rede* ***eröffnet****.* 'The conductor opened/opened the village festival with a speech.'

Table 2. Tense Alternation: Examples

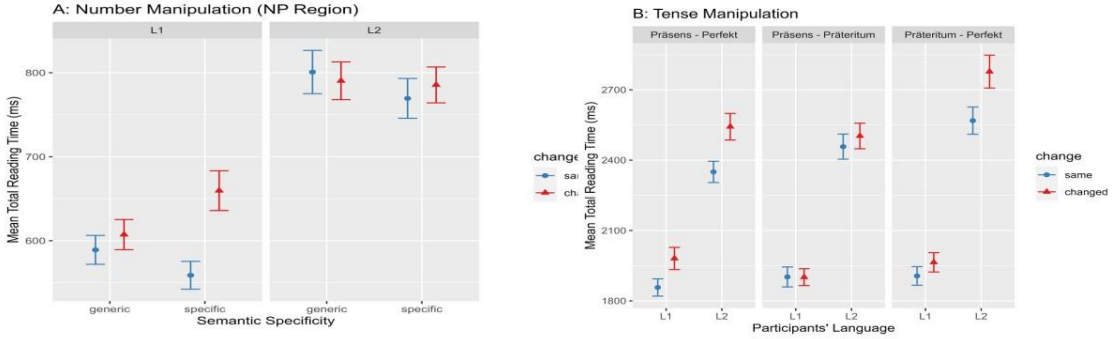


Figure 1. Results. Mean Reading Times for Critical Regions in the Number (A, left) and Tense (B. right) Manipulation

References

Garnham, A., & Oakhill, J. (1996). The mental models theory of language comprehension. In: *Models of understanding text* (pp. 313–339). Lawrence Erlbaum Associates, Inc.

Gurevich, O., Johnson, M. A., & Goldberg, A. E. (2010). Incidental verbatim memory for language. *Language and Cognition, 2*(1), 45–78.

Sampaio, C., & Konopka, A. E. (2013). Memory for non-native language: The role of lexical processing in the retention of surface form. *Memory, 21*(4), 537–544.

Bordag, D., Opitz, A., Polter, M., & Meng, M. (2021). Non-native Readers Are More Sensitive to Changes in Surface Linguistic Information than Native Readers. *Bilingualism: Language and Cognition, 24*(4), 599–611.

Reading Tiramisu in Czech and English: Robust Processing Speed Differences in Translation Equivalent Stimuli

Jan Chromý, Markéta Cehláková (Charles University), Michael Ramscar (University of Tübingen)
jan.chromy@ff.cuni.cz

Background: Crosslinguistic studies on sentence comprehension using matched translation-equivalent stimuli are rather rare. A recent study [1] revealed intriguing differences in agreement attraction effects between Czech and English. Upon closer examination, however, another interesting pattern emerged: native English readers processed individual words significantly faster in their language than Czech readers did in theirs. The present study aimed to further investigate this phenomenon.

Method: We created 70 pairs of translation-equivalent sentences (Table 1), matched in length in words between Czech and English. These sentences included words that have identical graphical forms in both languages (e.g., *tiramisu* or *Robert*). While some of these words were declinable in Czech, meaning they exhibited distinctive inflectional paradigms with varying endings for different cases, others were indeclinable, retaining the same form regardless of case. Native speakers of Czech (N=176) and English (N=176) read these sentences, along with 48 fillers, in a self-paced reading task using a moving-window presentation. After each sentence, participants answered a yes-no comprehension question. Reaction times (RTs) for individual words served as the dependent variable, while the independent variables were language, word length (in characters), and declinability in Czech.

Results: A linear mixed-effects model with language and word length as fixed effects, and participant and item as random effects, revealed significant effects for both factors and their interaction (Figure 1). Specifically, (i) English words were processed faster than Czech words, (ii) longer words took more time to process, and (iii) the effect of word length was more pronounced for Czech. In a model focusing on identical words across the two languages, we found a main effect of language (English was faster) and an interaction between language and declinability (declinable words were processed more slowly in Czech than indeclinable ones; Figure 2).

Discussion: The results show robust differences in RTs between English and Czech. We argue that these differences are driven by the greater morphological complexity (i.e. efficiency) of Czech compared to English [2], which increases the cognitive load (i.e., prior) on Czech speakers even when processing identical words.

Table 1: Item example.

Moje	teta	obvykle	pije	Bordeaux	a	další	francouzská	vína.
My	aunt	usually	drinks	Bordeaux	and	other	French	wines.

Figure 1: Predicted RTs from the linear mixed-effects model targeting the effects of language and word length.

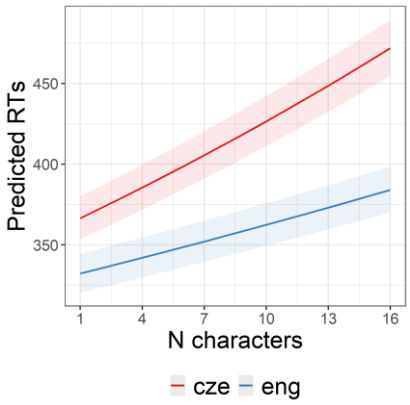
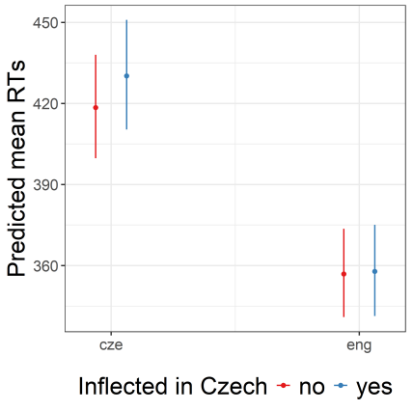


Figure 2: Predicted RTs from the linear mixed-effects model targeting the effects of language and declinability in Czech.



References

[1] Chromý, J., Brand, J. L., Laurinavichyute, A., & Lacina, R. (2023). Number agreement attraction in Czech and English comprehension: A direct experimental comparison. Glossa Psycholinguistics, 2(1).

[2] Sadeniemi, M., Kettunen, K., Lindh-Knuutila, T., & Honkela, T. (2008). Complexity of European Union Languages: A comparative approach. Journal of Quantitative Linguistics, 15(2), 185-211.

The Presentation Format in Cloze Tasks Influences Syntactic Predictability but It Does Not Influence Semantic Predictability

Bozhidara Hristova, Robin Lemke, Lisa Schäfer, Heiner Drenhaus & Ingo Reich
(Saarland University)
bozhidara.hristova@uni-saarland.de

In research on predictability effects in language comprehension, predictability is often operationalized in terms of cloze probability. Although predictability effects based on cloze probabilities have been consistently attested in research (e.g., Rayner & Well 1996, Federmeier & Kutas 1999, Smith & Levy 2011), the concern has emerged that the standard cloze task (Taylor 1953) does not capture the influence of memory decay on predictability (*lossy-context surprisal*, Futrell et al. 2020). For this reason, Apurva and Husain (2021) perform cloze norming with a new paradigm, in which the context is presented in a self-paced reading (SPR) format. However, it remains unclear whether a new paradigm is needed, i.e. whether the presentation format in cloze experiments affects the obtained probabilities.

We investigate this question on the case of a cloze task conducted to test for predictability effects on the usage of Gapping (Exp. 1, N = 160). Since the obtained cloze probabilities did not predict reading times in an SPR experiment in the expected direction, we hypothesized that they did not model the expectations of the participants in the SPR experiment accurately. We therefore conducted a written sentence completion study with a centered SPR presentation (Exp. 2, N = 48) and compared the responses to the ones from Exp. 1. We focused on the predictability of the verb in the second conjunct (C2) of parallel coordinations (i.e. coordinations with the same verb in both conjuncts) (1) and intended to manipulate the predictability of the C2 verb through the number of objects in the context sentence (OBJECT NUMBER, one/two) (2). We hypothesized that mentioning only one object in the context would increase the probability of the C2 verb. In Exp. 1, the context/target pairs were displayed together with a text box after the C2 subject. In Exp. 2, subjects read the items word by word by pressing the spacebar. After reading the C2 subject, a text box appeared. In both studies, participants were asked to type in the continuation they considered most likely. We aimed to test whether the SPR format would change the produced continuations due to participants accessing the preceding linguistic context only from memory. For the probability of a parallel (same-verb) response, we expected that the SPR presentation would make our manipulation less effective, i.e. that OBJECT NUMBER will have a weaker effect on the probability of a parallel continuation in Exp. 2 than in Exp. 1 (OBJECT NUMBER × PRESENTATION TYPE interaction). Regarding the syntactic realization of the parallel response (Gapping vs nonelliptical), we hypothesized that Gapping will be less frequent in Exp. 2 than in Exp. 1, since its licensing is conditioned on the content of the first conjunct (main effect of PRESENTATION TYPE).

With respect to the probability of a parallel response, we found no significant interaction between OBJECT NUMBER and PRESENTATION TYPE (z = -0.46, p = 0.65) and no main effect of PRESENTATION TYPE (z = -0.37, p = 0.71) (Fig. 1). The similarity between the two experiments was also evident in the overall uncertainty in the sample of responses, with no main effect of PRESENTATION TYPE (z = -0.17, p = 0.87) and no OBJECT NUMBER × PRESENTATION TYPE interaction effect (z = 0.15, p = 0.88) on the entropy of the produced C2 verb per item (Fig. 2). However, looking at the form of the parallel continuations, we observed a significantly lower probability of Gapping in Exp. 2 than in Exp. 1 (z = -3.34, p = < 0.001) (Fig. 3). The results suggest that syntactic predictability is more sensitive to the accuracy of memory representations than the predictability of semantic content. One reason could be that producing an unlicensed syntactic structure will lead to an ungrammatical sentence. No such risk is associated

with producing a semantically different verb. Thus, when subjects are uncertain about the preceding context, they opt for the safe (nonelliptical) syntactic option more often.

(1) *Anna is knitting a sweater, and Max* ⟨***is knitting***⟩ *a scarf.*

(2) *Die Anna und der Max haben im Bastelladen* (***Wolle*** | ***Wolle und Origamipapier***)
the Anna and the Max have in craft.store wool wool and origami.paper
gekauft. Die Anna strickt einen Pulli und der Max _____.
bought the Anna knits a.ACC sweater and the.NOM Max _____

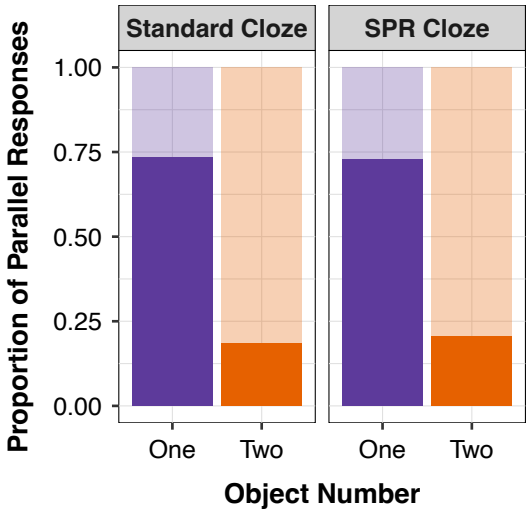


Fig. 1 Proportion of parallel (same-verb-as-in-first-conjunct) continuations

Fig. 3 Proportion of Gapping in the parallel responses

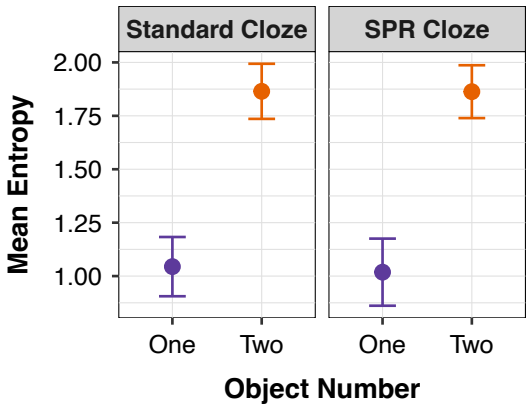


Fig. 2 Mean entropy and SE per item in the produced C2 verbs. For the calculations for the standard cloze task, we used the averaged entropy values from 4000 samples with the size of 48 subjects.

References

Apurva, & Husain, S. (2021). Revisiting anti-locality effects: Evidence against prediction-based accounts. *Journal of Memory and Language, 121*, 104280.

Federmeier, K. D., & Kutas, M. (1999). Right words and left words: Electrophysiological evidence for hemispheric differences in meaning processing. *Cognitive Brain Research, 8*(3), 373–392.

Futrell, R., Gibson, E., & Levy, R. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science, 44*, e12814.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review, 3*(4), 504–509.

Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the Cognitive Science Society* (pp. 1637–1642). Cognitive Science Society.

Taylor, W. L. (1953). ‘‘Cloze procedure’’: A new tool for measuring readability. *Journalism Quarterly, 30*, 415–433.

Analysis of simplification in coreference from two perspectives

Jablotschkin, S. (sarah.jablotschkin@uni-hamburg.de)[1], Lapshinova-Koltunski, E.,[2] Zinsmeister, H.[1]

In this paper, we analyse coreference features of the German language, focusing on the phenomenon of *simplification*, i.e. the tendency to use words and constructions that are assumed to be easier perceived, understood, or produced. Simplification is one of the means used by language users in order to optimise communication effectively. We are interested in how simplification is reflected in coreference in two different language products exposed to the phenomena of simplification: simultaneous interpreting and Easy German. As seen from example (1), the English source contains the chain *the practice of sandblasting – which – jeans sandblasted* with mentions filled with a relative pronoun and a full lexical phrase. At the same time, the interpreting into German contains a demonstrative pronoun (*das*) and an adverb (*so*) instead. From the lexical point of view, the means of referring are simpler in the interpreted output. In contrast, the coreference chain in the Easy German example in (2) contains no pro-forms, but lexical repetitions as a simplification strategy. In addition, the anaphors are highlighted by being positioned sentence-initially.

(1) **English original**: *In particular, I want to draw attention to the practice of sandblasting of jeans which happens more in Bangladesh than anywhere else in the world. Up to one hundred million pairs of jeans sandblasted a year being export from Bangladesh.* **German interpreting**: *Aber was dort in Bangladesch passiert, ist weiter eine Bedrohung für die Gesundheit der Arbeitnehmer, insbesondere die Sandstrahlmethode für Jeans. Das wird in Bangladesch vor allen Dingen durchgeführt. Einhundert Millionen Jeans werden so hergestellt und exportiert pro Jahr.*

(2) **Easy German**: *In Hamburg sind am Wochen·ende 2 große Veranstaltungen. Diese 2 großen Veranstaltungen sind: • Ein Musik·fest. • Und eine Sport·veranstaltung. Die 2 großen Veranstaltungen sind in St. Pauli. […] Und die 2 großen Veranstaltungen sind […] Zu diesen 2 großen Veranstaltungen kommen sehr viele Menschen.*

While both language products are known to be simplified, the driving forces of the optimisation process differ. Easy German is simplified to be better perceived and understood by the target audience, i.e. the receiver side. At the same time, simultaneous interpreting is simplified due to the production constraints on the producer side, i.e. the interpreter who optimises the output to reduce their own cognitive load.

We are interested in the differences and similarities of the simplified language products that are the results of these two varying optimisation reasons. For instance, shorter coreference chains, mentions used as subjects and fewer expression variants per chain indicate simplification, as well as the length of the expression measured in words: the shorter the mention expressions, the simpler the text. The formulated features are based on the studies in the area of automatic coreference resolution for German (see e.g. [5]) as well as accessibility analysis for German (see e.g. [2]).

We use two different sets of data. For the analysis of simultaneous interpreting, we use a sample of 137 texts of German interpreting from English extracted from EPIC-UdS ([3]), a multilingual parallel and comparable corpus of simultaneous interpreting of political speeches. For the analysis of Easy German, we use a sample of about 4,700 texts from DE-Lite v1 ([1]).To analyse coreference, we annotated the data with the state-of-the-art coreference resolver CorPipe ([4], [6]). In our presentation, we will compare the frequency distributions of the annotated coreference features across the texts in the two data sets and discuss them in relation to the different simplification strategies.

[1] Univeristy of Hamburg, [2] University of Hildesheim

References

1. Jablotschkin, Sarah, Teich, Elke, & Heike Zinsmeister (2024). DE-Lite - a New Corpus of Easy German: Compilation, Exploration, Analysis. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 106–117. Association for Computational Linguistics, St. Julian's, Malta.

2. Kunz, Kerstin (2010). *Variation in English and German nominal Coreference. A study of political essays*. Peter Lang, Frankfurt am Main, Germany.

3. Przybyl, Heike, Lapshinova-Koltunski, Ekaterina, Menzel, Katrin, Fischer, Stefan, & Teich, Elke (2022). EPIC UdS - Creation and Applications of a Simultaneous Interpreting Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1193–1200. European Language Resources Association, Marseille, France.

4. Straka, Milan (2023). ÚFAL CorPipe at CRAC 2023: Larger Context Improves Multilingual Coreference Resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 41–51. Association for Computational Linguistics, Singapore.

5. Strube, Michael, & Hahn, Udo (1999). Functional Centering: Grounding Referential Coherence in InformationStructure. *Computational Linguistics* 25(3). 309–344.

6. Žabokrtský, Zdeněk, & Ogrodniczuk, Maciej (2023). *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*. Association for Computational Linguistics, Singapore.

The Role of Impliedness in Investigating the Interplay of Word Order and Information Status

Torsten Kai Jachmann, Heiner Drenhaus, Francesca Delogu, Matthew W. Crocker
Language Science and Technology, Saarland University, Germany

In languages with free word order, non-canonical structures generally increase processing costs compared to canonical ones (e.g. [1], [2], [3]). Given information reduces processing costs when it appears earlier in a sentence, following the 'given-before-new' principle (e.g., [4], [5], [6]). When combined, the cost of object-first structures can be mitigated if the object is given rather than new (e.g. [2], [7]). However, sometimes new information is implied or context-related rather than entirely new, which may affect these cost modulations.

In our study in German, we investigated the interplay between Word Order (WO) preferences (SOV over OSV) and different Information Status (IS) (Given – explicitly mentioned vs. Implied – inferable from the context vs. New – unmentioned and unrelated to the context), as illustrated in Example 1, with the focus on investigating the behavior of Implied information as compared to the other two status. To examine the effect of IS on both NP1 and NP2, 12 conditions were created. The IS of NP2 was fully counter balanced according to that of NP1.

Analyses were conducted in Julia by fitting Linear Mixed-Effect Models using the MixedModels package. Models included a covariate of Word Length (centered and scaled to a range between -1/2 and 1/2) and fixed effects for WO (-1/2, 1/2) and IS. Contrasts for the factor IS were comparing Implied trials to Given trials (1/2, -1/2, 0) and New trials (0, -1/2, 1/2) respectively. The results (see Fig.1 and Table 1) showed effects of WO only on NP1, such that Object-first structures led to longer reading times (RTs) compared to Subject-first structures. IS showed effects for both comparisons such that Implied NPs were taking an intermediate position between Given (fastest) and New NPs (slowest). No interactions between WO and either IS comparison were observed. In the NP2 region, an interaction between WO and the contrast comparing Given-first with Implied-first structures was observed, indicating that an NP following a Given NP1 was read faster than such NPs following either an Implied or New NP1 only in Object-first structures.

Taken together, the results suggest that effects of WO and IS are of an additive nature on NP1, indicating that a violation of either Given-first or Subject-first expectations lead to increased processing difficulties. At the same time, implied information can still be processed faster than entirely new information, underlining that differentiation between new and implied entities is critical. The interaction observed on NP2 is in line with earlier findings that object-first difficulties are reduced when the object is given ([2], [7]).

Our results, thus, are in line with previous ERP studies (e.g., [8]) indicating that implied information is more accessible than entirely new information, leading to lower retrieval costs, but still requires integrating a new discourse referent.

These results will further be investigated by conducting an ERP study on the same stimuli utilized in this experiment. We expect modulations in the P600 time-window to pattern with our RT results presented here (as RTs have been shown to be sensitive to integration rather than retrieval costs [9]). We further expect N400 effects in line with gradually increasing retrieval costs as a function of accessibility of the presented entities.

References
[1] Bader & Meng (1999), J. Psycholinguist. Res.; [2] Kaiser & Trueswell (2004), Cognition; [3] Bornkessel & Schlesewsky (2006), J. Ger. Linguistics; [4] Arnold et al., (2013), Wiley Interdiscip. Rev. Cogn. Sci.; [5] Primus (2017), Syntax; [6] Krifka, & Musan (2012), De Gruyter Mouton; [7] Yano & Koizumi (2018), Lang. Cogn. Neurosci.; [8] Burkhardt (2006). Brain Lang; [9] Aurnhammer et al. (2021), PloS One.

Example 1:

Ein Bäcker ging auf ein Konzert.
(A baker went to a concert.)

Canonical word order (SOV):

a. Ich habe gesehen, dass der Bäcker [G] dort gestern den Musiker [I] / einen Piloten [N] …
(I saw that the baker [SUBJ / G] there yesterday the musician [OBJ / I] / a pilot [OBJ / N] ...)
b. Ich habe gesehen, dass der Musiker [I] dort gestern den Bäcker [G] / einen Piloten [N] ...
(I saw that the musician [SUBJ / I] there yesterday the baker [OBJ / G] / a pilot [OBJ / N] ...)
c. Ich habe gesehen, dass ein Pilot [N] dort gestern den Bäcker [G] / den Musiker [I] ...
(I saw that a pilot [SUBJ / N] there yesterday the baker [OBJ / G] / the musician [OBJ / I ...)

Non-Canonical word order (OSV):

d. Ich habe gesehen, dass den Bäcker [G] dort gestern der Musiker [I] / ein Pilot [N] …
(I saw that the baker [OBJ / G] there yesterday the musician [SUBJ / I] / a pilot [SUBJ / N] ...)
e. Ich habe gesehen, dass den Musiker [I] dort gestern der Bäcker [G] / ein Pilot [N] ...
(I saw that the musician [OBJ/Implied] there yesterday the baker [SUBJ / G] / a pilot [SUBJ / N] ...)
f. Ich habe gesehen, dass einen Piloten [N] dort gestern der Bäcker [G] / der Musiker [I] ...
(I saw that a pilot [OBJ/New] there yesterday the baker [SUBJ / G] / the musician [SUBJ / I ...)

G – Given ; I – Implied ; N – New ; WO – Word Order

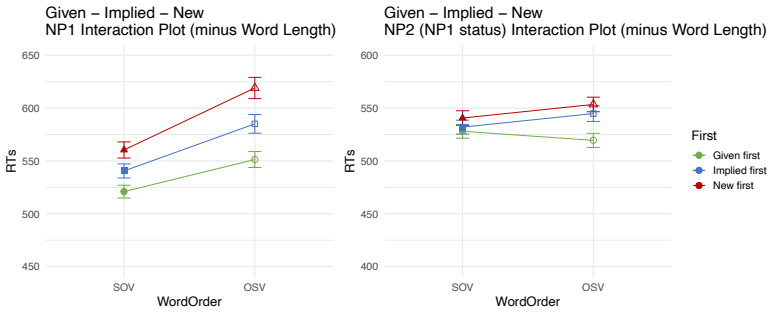


Fig.1 – Regression based Interaction plots of effects in NP1 and NP2 regions with Word Length effects removed.

	Word Length	Word Order	I vs. G	I vs. N	WO * IvG	WO * IvN
NP1	***	***	***	***	–	–
NP2	***	–	**	–	*	–

– - no effect , * - p < 0.05 , ** - p < 0.01 , *** - p < 0.001 ; G – Given ; I – Implied ; N – New ; WO – Word Order

Table 1 – Results.

Evaluating LLMs Through Self-Play: Testing Comprehension and Generation of Referring Expressions in Multi-Modal LLMs by Employing a Picture-Guessing Dialogue Game

Eileen Kammel, Anne Beyer, David Schlangen (University of Potsdam)
eileen.niedenfuehr@uni-potsdam.de

Referring expressions (REs) are integral to communication, as they allow speakers to identify and refer to specific entities within discourse. Human speakers tend to adapt and optimize their expressions based on context and the intended referent by adhering to conversational norms (Grice, 1975). This study examines how large language models (LLMs), particularly multi-modal models, approach this task, investigating their success and limitations in processing REs in different contexts. The research employs a version of the picture-guessing game *reference-game* as implemented in the clembench framework (Chalamalasetti et al., 2023). Using two image sets of varying levels of complexity, key questions explored in this work include (1) whether LLMs scores differ in RE production versus RE comprehension, (2) whether the complexity of images impacts performance, and (3) how the amount of information in LLM-generated REs compares to both theoretically optimal REs and those produced by human speakers. Challenging the latest SOTA open-source models and selected commercial LLMs, the study allows for evaluation of possible performance discrepancies between models of the open-source versus the commercial tier as well as a comparison between models of the same tier.

The game involves two players presented with the same images in different orders. One player describes a target image to differentiate it from the others, while the second player must select the correct image based on this description. Attributes may be shared between the images, making context important. Images from the TUNA corpus (Gatt et al., 2009) are chosen for simpler images, 3D shapes corpus (Burgess and Kim, 2018) provides more complex images. Following Glucksberg et al. (1966), who have shown that the development of understanding and production of REs can be attested to different phases during language acquisition, we also investigate the presence of these abilities in LLMs separately.

To assess comprehension, successful REs are collected from human trials and presented to an LLM acting as a guesser. Pairs of volunteers play the game via the chat-room like slurk framework (Götze et al., 2022). Expressions that led to successful target identification at least once are presented to the LLMs. To rule out potential location biasthe images are shown in all possible permutations, and the model's responses are compared across these variations.

For production assessment, the LLM generates the REs. To enable a more nuanced evaluation, context-dependent descriptions are compared to "ground truth" descriptions generated by the same models without context for the same target image. This comparison aims to determine whether contextual information influences the LLM's descriptions. Additionally, successful game episodes are compared across both comprehension and production tasks to uncover potential differences in the challenges posed to LLMs by these two settings. Building on the findings of Hakimov et al. (2024), who noted that LLM-generated descriptions often resemble image captions with excessive detail, prompts will be adapted to elicit more concise, RE-like expressions, adhering to communicative maxims described by Grice. In the qualitative analysis, expressions generated by both humans and LLMs will be compared to optimal expressions calculated using methods such as the brevity algorithm and the incremental algorithm, as described by Dale and Reiter (1995).

References

Burgess, C. and Kim, H. (2018). 3d shapes dataset. https://github.com/google-deepmind/3d-shapes.

Chalamalasetti, K., Götze, J., Hakimov, S., Madureira, B., Sadler, P., & Schlangen, D. (2023). clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents. In H. Bouamor, J. Pino, & K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 11174–11219). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.689

Dale, R., & Reiter, E. (1995). Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. In Cognitive Science (Vol. 19, Issue 2, pp. 233–263). Wiley. https://doi.org/10.1207/s15516709cog1902_3

Gatt, A., Belz, A., & Kow, É. (2009, March 1). The TUNA-REG Challenge 2009: Overview and Evaluation results. ACL Anthology. https://aclanthology.org/W09-0629

Glucksberg, S., Krauss, R. M., & Weisberg, R. (1966). Referential communication in nursery school children: Method and some preliminary findings. Journal of Experimental Child Psychology, 3(4), 333–342. https://doi.org/10.1016/0022-0965(66)90077-4

Götze, J., Paetzel-Prüsmann, M., Liermann, W., Diekmann, T., & Schlangen, D. (2022). The slurk Interaction Server Framework: Better Data for Better Dialog Models. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 4069–4078). European Language Resources Association. https://aclanthology.org/2022.lrec-1.433

Grice, H. P. (1975). Logic and Conversation. In Speech Acts (pp. 41–58). BRILL. https://doi.org/10.1163/9789004368811_003

Hakimov, S., Abdullayeva, Y., Koshti, K., Schmidt, A., Weiser, Y., Beyer, A., & Schlangen, D. (2024). Using Game Play to Investigate Multimodal and Conversational Grounding in Large Multimodal Models (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2406.14035

On the Limits of LLM Surprisal as a functional Explanation of ERPs

Benedict Krieger (Saarland University), Harm Brouwer (Tilburg University),
Christoph Aurnhammer (Saarland University), Matthew W. Crocker (Saarland University)
bkrieger@lst.uni-saarland.de

The impressive comprehension-like behavior of LLMs trained on next word prediction has led researchers to suggest that these models are to some extent accurate models of human comprehension (e.g., Goldstein et al., 2022; Schrimpf et al., 2021). Studies correlating LLM-derived surprisal and neural correlates have focused predominantly on the N400 - an event-related potential (ERP) component sensitive to the expectancy of a word in context - in naturalistic comprehension (De Varda et al., 2023; Michaelov et al., 2024). Experiments, however, show that beyond expectancy, the N400 is also sensitive to semantic association, defined as the extent to which word meaning is primed by its prior context (see Kutas & Federmeier, 2011). While LLMs are also sensitive to association (Michaelov & Bergen, 2022), the influence of expectancy on the N400 can be overridden entirely when target word meaning is contextually primed, such that semantically unexpected words do not increase N400 amplitude (e.g., Nieuwland & Van Berkum, 2005 and Delogu et al., 2019., shown in Fig. 1). While these words were clearly surprising to humans, as reflected in increased P600 amplitude, it is unclear how LLMs perform in these cases. Moreover, the P600 - which has received little attention in this line of research - has been found to be graded for plausibility and insensitive to association (Aurnhammer et al., 2023; Aurnhammer et al., 2021; Brouwer et al., 2021). We examine the ability of LLM surprisal to model three German ERP-studies that specifically sought to disentangle the influence of expectancy, plausibility, and association on both the N400 and P600 (Aurnhammer et al., 2023; Aurnhammer et al., 2021; Delogu et al., 2019). Using two transformer models, a smaller model (GPT-2) and a larger state-of-the-art model (LeoLM), we replicated the sensitivity of LLMs to both expectancy and association. However, results from an rERP analysis (Smith & Kutas, 2015) using LLM-derived surprisal to re-estimate ERPs led to mixed results: Surprisal collected with the larger LLM predicted an N400 difference that was unobserved (in Delogu et al. 2019, see Fig. 1, right panel), while surprisal collected with the smaller LLM did not predict such a difference – in line with the observed ERP profile, but revealing sensitivity of surprisal towards association. Furthermore, the magnitude of effects was underestimated. For the P600, LLMs were able to capture violations of selectional restrictions, but failed to account for the graded sensitivity of the P600 to plausibility (Aurnhammer et al., 2023). If LLMs are indeed an accurate characterisation of (aspects) of human comprehension mechanisms, they should account for N400 and P600 effects and their differential sensitivity to association, expectancy and plausibility. Our findings suggest that LLM surprisal may not offer an accurate characterisation of the underlying functional generators of either the N400 or P600, and motivate exploring alternative LLM-derived linking hypotheses to the N400 and P600 informed by mechanistic accounts of the processes associated with these components (Brouwer et al., 2017; Fitz & Chang, 2019; Li & Futrell, 2023; Li & Ettinger, 2023). We argue that until LLMs are shown to account for critical data points through such linking hypotheses, strong conclusions about their validity as models of the human comprehension system (e.g., Goldstein et al., 2022; Schrimpf et al, 2021) are too premature.

References

Aurnhammer, C., Delogu, F., Brouwer, H., & Crocker, M. W. (2023). The P600 as a continuous index of integration effort. Psychophysiology, 60(9), e14302.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. PLOS ONE,16(9), e025743.

Brouwer, H., Crocker, M., Venhuizen, N., & Hoeks, J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. Cognitive Science, 41(S6),1318-1352.

Brouwer, H., Delogu, F., Venhuizen, N., & Crocker, M. W. (2021). Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. Frontiers in Psychology, 12, 615538.

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Eventrelated potentials index lexical retrieval (N400) and integration (P600) during language comprehension. Brain and Cognition, 135, 103569.

De Varda, A. G., Marelli, M., & Amenta, S. (2023). Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. Behavior Research Methods.

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. Cognitive Psychology, 111, 15–52.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., . . . Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. Nature Neuroscience, 25(3), 369–380.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual review of psychology, 62, 621-47.

Li, J., & Ettinger, A. (2023). Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. Cognition, 233, 105359.

Li, J., & Futrell, R. (2023). A decomposition of surprisal tracks the N400 and P600 brain potentials. In Proceedings of the 45th Annual Meeting of the Cognitive Science Society (p. 587-594).

Michaelov, J., Bardolph, M., Van Petten, C., Bergen, B., & Coulson, S. (2024). Strong prediction: Language model surprisal explains multiple N400 effects. Neurobiology of Language, 1-29.

Michaelov, J., & Bergen, B. (2022). Collateral facilitation in humans and language models. In Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL) (pp. 13–26).

Nieuwland, M. S., & Van Berkum, J. J. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. Cognitive Brain Research, 24(3), 691-701.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., . . . Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences, 118(45), e2105646118.

Smith, N., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: I. The rERP framework. Psychophysiology, 52(2), 157-168.

Figures

Condition	Assoc.	Plaus.	Cloze
A Baseline	6.32	6.28	0.38
B Event-related	6.32	2.42	0.13
C Event-unrelated	1.56	1.93	0.008

Mean association ratings between prime noun and target, mean plausibility ratings for stimuli up until including target, cloze probabilities for target.

Example Item

John entered the restaurant.
Before long, he opened the menu...

John left the restaurant.
Before long, he opened the menu...

John entered the apartment
Before long, he opened the menu...

English translation.

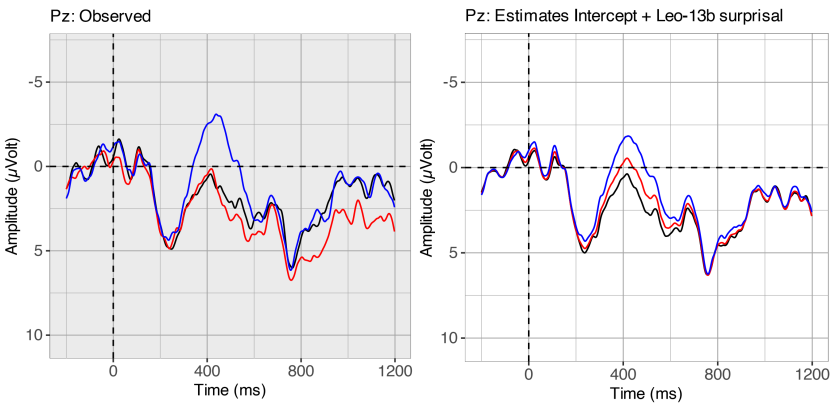


Figure 1: Left: experimental conditions from Delogu et al. (2019), middle: observed ERP profile, right: rERP forward estimates with LeoLM surprisal

Intercomprehension of Slavic Functional Multiwords: Translation Experiment Results

Maria Kunilovskaya, Iulia Zaitova, Wei Xue, Ira Strenger (University of Saarland)
maria.kunilovskaya@uni-saarland.de

This study reports the results of a free translation experiment as a probe for Slavic intercomprehension between Russian and Czech, Polish, Bulgarian, Belarusian, Ukrainian. The experiment focuses on non-compositional functional multiword expressions – or microsyntactic units (MSUs) (Avgustinova & Iomdin, 2019) from five word classes: prepositions, adverbial predicatives, conjunctions, particles, parentheses. MSUs require additional cognitive effort in cross-lingual comprehension because their meaning cannot be inferred from the components. The lack of transparency and discourse-organising functions make MSU a good case for intercomprehension studies. They reflect the ability of the participants to understand the message in a foreign language.

The data comes from an experiment, where native speakers of Russian translated contextualised MSUs from five Slavic languages into Russian. The study engaged 126 users without formal knowledge of the source languages (SLs). The translation tasks were based on at least 50 sentences containing a unique MSU stimulus, and each stimulus has generated at least 20 responses. The Slavic MSUs (and their contexts) were extracted as correspondences for Russian MSUs using bidirectional parallel corpora and lexicographic resources of the Russian and Czech National Corpora[1]. The participants' responses were annotated for seven types of translation solutions (paraphrase, correct, fluent literal, awkward literal, fantasy, noise, and empty), designed to capture the level of the cross-linguistic intelligibility of the stimuli. The annotation was used to calculate items' intelligibility scores.

The study aims to reveal factors that favour intercomprehension across Slavic languages based on a range of computational representations and modelling approaches. In particular, regression and correlation analysis are used to identify the most important intercomprehension predictors. The stimuli are represented by features that reflect the properties of the translation tasks and their outcomes, including a pronunciation-based variant of the point-wise Phonologically Weighted Levenshtein Distance (PWLD) motivated in our previous work (Zaitova et al., 2024), cosine similarities, surprisals, translation quality scores and translation solution entropy indices. Cosine similarities and surprisals are calculated based on ruRoBERTa-large model (Zmitrovich et al., 2024), a dedicated Russian language Transformer, which can process input in Latin script. Automatic translation quality scores were calculated using COMET models (Rei et al., 2022). These approaches utilise contexts for the targeted MSUs available in our dataset.

The experimental results from both annotation and computational models confirm the expected gradual increase of mutual intelligibility from West-Slavic to East-Slavic languages. We show that intelligibility is highly contingent on the ability of speakers to recognise and interpret formal similarities between languages as well as on the size of these similarities. For several Slavic languages, the context sentence complexity was a significant predictor of intelligibility.

[1] https://ruscorpora.ru/en/ and https://www.korpus.cz/

References

Avgustinova, T., & Iomdin, L. (2019). Towards a typology of microsyntactic constructions. In *Computational and Corpus-Based Phraseology: Third International Conference, Europhras 2019, Malaga, Spain, September 25–27, 2019, Proceedings 3* (pp. 15-30). Springer International Publishing.

Rei, R., C. de Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L., & Martins, A. F. T. (2022, December). COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In P.Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, . . . M. Zampieri (Eds.). In *Proceedings of the seventh conference on machine translation (WMT)* (pp. 578–585). Association for Computational Linguistics.

RNC. (2003–2023). Russian National Corpus [Retrieved September 28, 2023]. *Russian National Corpus Project*.

Zaitova, I., Stenger, I., Butt, M. U., & Avgustinova, T. (2024, May). Cross-Linguistic Processing of Non-Compositional Expressions in Slavic Languages. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon@ LREC-COLING 2024* (pp. 86-97).

Zmitrovich, D., Abramov, A., Kalmykov, A., Kadulin, V., Tikhonova, M., Taktasheva, E., ... & Fenogenova, A. (2024, May). A Family of Pretrained Transformer Language Models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 507-524).

Can Translation Task Difficulty Predict the Properties of Translation?

Maria Kunilovskaya, Elke Teich (University of Saarland),
Ekaterina Lapshinova-Koltunski (University of Hildesheim)
maria.kunilovskaya@uni-saarland.de

Translations are known to have lexical, morphosyntactic and semantic deviations from comparable originally-authored target language (TL). These deviations are known as translationese. Translation studies accumulated extensive evidence of translationese to raise concerns in the related research fields, such as machine translation (Artetxe et al., 2020) and contrastive studies (De Baets et al., 2020). Pinpointing the factors that contribute to translationese deviations remains a challenging task. The explanatory efforts link translationese to trends in translational behaviour (simplification, explicitation, etc), to socio-cultural factors (expertise, registers) or to the cross-linguistic nature of the translation process. It has been shown that more challenging cognitive conditions may trigger a simpler, more conventionalised (Kruger & De Sutter, 2018), more explicit (Olohan & Baker, 2000) or more implicit output (Lapshinova-Koltunski et al., 2022).

The current project aims to explore the impact of the source document complexity on the properties of translations. We rely on a range of complexity measures, including information-theoretical indices.

Previous work used average sentence surprisal to show that the amount of information in the source is positively correlated with the amount of information in the target regardless of the translation mode (Kunilovskaya et al., 2023; Przybyl et al., 2022). Other computational studies have compelling evidence that translationese deviations can be explained by the source language (SL) influence (Rabinovich, Ordan & Wintner, 2017; Evert & Neumann, 2017; Kunilovskaya & Lapshinova-Koltunski, 2020). The novelty of our approach consists (I) in the truly cross-lingual nature of the experiments where the values for the source document are used to predict the properties of the translation, (ii) in the focus on the sub-sentential source and target items (content tokens and syntactic subtrees) to represent document pairs, and (iii) in indexing the two subprocesses involved in translation: source text comprehension and SL-TL transfer. The comprehension difficulty is captured by measures of syntactic complexity such as dependency distance, hierarchical distance, tree depth, and branching factor as well as by the surprisal of the content items from GPT-2. The transfer difficulty is operationalised (a) as the entropy of translation variants for a SL item registered in a large parallel corpus and (b) as the semantic alignment score (cosine similarity between contextualised embeddings of content items). The response variable – translationese properties of target – is a document-level probability of being a translation from a classifier that can reliably distinguish translations and comparable non-translations in the TL (F1-score 80-90%) using hand-engineered delexicalised translationese predictors.

Theoretically, the more demanding SL documents can be expected to generate more deviant translations. If this is the case, translationese can be explained as a rational response to increased cognitive pressure on the assumption that producing deviant translations requires less production effort.

The preliminary regression results on a large bi-directional Europarl corpus (English-German) show that transfer difficulty indicators are more relevant to the task than syntactic complexity measures or lexical comprehension difficulty measured as surprisal of the source content items, although the correlation is very weak.

References

Artetxe, M., Labaka, G., Agirre, E., & Center, H. (2020). Translation Artifacts in Cross-lingual Transfer Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, *(EMNLP)* (pp. 7674–7684). https://github.com/pytorch/fairseq

De Baets, P., Vandevoorde, L., & De Sutter, G. (2020). On the usefulness of comparable and parallel corpora for contrastive linguistics. testing the semantic stability hypothesis. *New Approaches to Contrastive Linguistics. Empirical and Methodological Challenges*. Berlin/Boston: De Gruyter, 85–126.

Evert, S., & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. *Empirical translation studies: New methodological and theoretical traditions*, 300, 47–80.

Kruger, H., & De Sutter, G. (2018). Alternations in contact and non-contact varieties: Reconceptualising that-omission in translated and non-translated English using the MuPDAR approach. *Translation, Cognition & Behavior*, 1(2), 251–290.

Kunilovskaya, M., & Lapshinova-Koltunski, E. (2020, May). Lexicogrammatic translationese across two targets and competence levels. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 4102-4112).

Kunilovskaya, M., Przybyl, H., Teich, E., & Lapshinova-Koltunski, E. (2023, April). Simultaneous Interpreting as a Noisy Channel: How Much Information Gets Through. In *Proceedings of the international conference on recent advances in natural language processing* (pp. 608–618). INCOMA Ltd. https://doi.org/10.26615/978-954-452-092-2\ 066

Lapshinova-Koltunski, E., Pollkläsener, C., & Przybyl, H. (2022). Exploring explicitation and implicitation in parallel interpreting and translation corpora. *The Prague Bulletin of Mathematical Linguistics*, 119, 5–22. https://ufal.mff.cuni.cz/pbml/119/art-lapshinova-koltunski-pollklaesener-przybyl.pdf

Olohan, M., & Baker, M. (2000). Reporting that in translated english. Evidence for subconscious processes of explicitation? *Across languages and cultures*, 1(2), 141–158.

Przybyl, H., Karakanta, A., Menzel, K., & Teich, E. (2022). Exploring linguistic variation in mediated discourse: Translation vs. interpreting. In *Mediated discourse at the european parliament: Empirical investigations* (pp. 191–218). Language Science Press. https://doi.org/10.5281/zenodo.6977050

Rabinovich, E., Ordan, N., & Wintner, S. (2017, July). Found in Translation: Reconstructing Phylogenetic Language Trees from Translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 530-540).

The Effect of Noun Phrase Complexity in Scientific Texts on Reading Times of Experts and Novices

Isabell Landwehr (Saarland University), Marie-Pauline Krielke (Saarland University), Stefania Degaetano-Ortlieb (Saarland University)
isabell.landwehr@uni-saarland.de

We investigate how different types of noun phrase (NP) complexity in scientific texts affect the reading times of experts and novices, for both in-domain and out-of-domain texts. The use of complex NPs is a key feature of scientific writing (Biber & Gray, 2011). For sentence processing, NP complexity can pose various challenges: More complex structures often include longer dependencies between head and dependent, increasing the integration cost of syntactic elements (cf. Dependency Locality Theory, Gibson, 1998). Moreover, complex NPs allow for information to be transmitted in a more compressed way increasing implicitness (Biber & Gray, 2010): Logical relations between the constituents of a compound remain implicit. Previous eye-tracking experiments show that increased complexity correlates with increased reading times (e.g. Just & Carpenter, 1980, for scientific texts). Individual reader characteristics, e.g. background knowledge and experience, also influence reading comprehension (Kendeou & Van den Broek, 2007). This is particularly the case for scientific texts, typically targeted at an expert audience (Halliday, 1988). Previous studies have considered word frequency or novelty (Just & Carpenter, 1980), dependency locality (Demberg & Keller, 2008) or terminology (Škrjanec et al., 2023) as complexity features.

We consider grammatical complexity by looking at structural compression (Biber & Gray 2016, p. 207). In particular, we analyze (a) different types of NP modification, i.e. different degrees of compression (see Table 1), and (b) differing internal structure (see Table 2). We use PoTeC (Jakobi et al., 2024), a German naturalistic eye-tracking-while-reading corpus of university students (novices: BA, experts: MA, PhD) of biology or physics reading in/out-of-domain textbooks. We foresee an effect of expertise, given higher domain knowledge for experts vs. novices: increased processing difficulty for novices for NPs with higher degree of compression and more complex internal structure, such as compounds, compared to e.g. nouns modified by a genitive construction. Additionally, experts are likely to outperform novices when reading texts from other scientific fields as their general scientific reading competence provides an extra advantage.

We fit linear mixed effects regression models using the *lme4* (Bates et al., 2015) package in R (R Core Team, 2023). Our dependent variables are first-pass reading time, total fixation time and total no. of incoming regressions. Our predictors are NP modification type or internal structure, and reader expertise, allowing us to model the effect of NP complexity considering reader's level of expertise and domain familiarity. As in previous work, we control for word length, type frequency, technicality of a term and surprisal. We also include an interaction of complexity and expertise, technicality and expertise as well as by-subject and by-word random effects. As a result, we aim to highlight the role of NP complexity on processing difficulty, and its interaction with readers' domain expertise.

Table 1: Compounds with different types of modification

Modification type (degree of compression)	Example
Nominal compound (higher compression) vs. modification by adjective (lower compression)	Wildtypprotein vs. endogenes zelluläres Protein
Nominal compound (higher compression) vs. modification by genitive construction (lower compression)	Phosphatverarmungszonen vs. Wellenvektor des Elektrons

Table 2: Compounds with differing internal structure

Internal structure	Example
Compound with one vs. two dependents	Energieminimum vs. Phosphatverarmungszonen

References

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Biber, D. & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes 9*, 2-20. https://doi.org/10.1016/j.jeap.2010.01.001

Biber, D. & Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. *English Language and Linguistics 15* (2), 223-250. https://doi.org/10.1017/S1360674311000025

Biber, D. & Gray, B. (2016). *Grammatical Complexity in Academic English.* Cambridge University Press. https://doi.org/10.1017/CBO9780511920776

Demberg, V. & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition 109*, 193-210. https://doi.org/10.1016/j.cognition.2008.07.008

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition 68*, 1-76.

Halliday, M. A. K. (1988). On the language of physical science. In M. Ghadessy (Ed.), *Registers of Written English: Situational Factors and Linguistic Features* (pp. 162-178). Pinter Publishers.

Jakobi, D. N., Kern, T., Reich, D. R., Haller, P. & Jäger, L. A. (2024). *PoTeC: A German naturalistic eye-tracking-while-reading corpus*. https://doi.org/10.17605/OSF.IO/DN5HP

Just, M. A. & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review 87*(4), 329-354. https://doi.org/10.1037/0033-295X.87.4.329

Kendeou, P. & Van Den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & Cognition 35*(7), 1567-1577. https://doi.org/10.3758/BF03193491

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Škrjanec, I., Broy, F. Y. & Demberg, V. (2023). Expert-adapted language models improve the fit to reading times. *Procedia Computer Science 225*, 3488-3497. https://doi.org/10.1016/j.procs.2023.10.344

Informativity Modulates Linearization Preferences in Referential Interaction

Muqing Li (muqingli@coli.uni-saarland.de)[1]; Noortje J. Venhuizen[2]; Torsten Kai Jachmann[1]; Heiner Drenhaus[1]; Matthew W Crocker[1]
[1] Department of Language Science and Technology, Saarland University, 66123 Saarbrücken, Germany
[2] Department of Cognitive Science and Artificial Intelligence, Tilburg University, 5037 AB Tilburg, Netherlands

In referential communication, speakers have been shown to strategically overspecify informative pre-nominal adjectives in their expressions [1] and order them to position the most informative property early in the sequence [2], a strategy we refer to as the "informative-first linearization preference", which can facilitate target identification for the listener [3]. Less is known, however, regarding whether informativity can influence linearization at the syntactic level, e.g., in pre- or post-nominal modifications, especially in interactive communication environments where the collaborative speaker may seek to be especially informative for the listener [4;5].

To quantify informativity of referential expressions in a visual scene, we use Referential Entropy Reduction (RER), which measures how much uncertainty about the target is reduced by each property word in an utterance [1]. Words have higher RER when they reduce uncertainty to a greater extent, by narrowing a greater referential scope in a shared visual scene. We compared Animal-Informative and Action-Informative conditions using stimuli depicting animals performing actions, which in German can be encoded flexibly using pre- and post-nominal structures (e.g., in Figure1, *der weinende Hase* vs. *der Hase, der weint*). In both conditions, the informative property (Animal or Action) yielded a higher RER than the uninformative one.

Across three experiments we investigated whether speakers prefer the informative-first linearization preference, above and beyond the overarching syntactic preference for pre-nominal modifications: In Experiment 1, participants acted solely as speakers, collaborating with a virtual partner online to complete a maze-based sentence task for target descriptions (Figure 1). In Experiment 2, also conducted online using the maze task, increased interaction, such that participants alternated between speaker and listener roles trial by trial. In Experiment 3, participants alternated roles face-to-face in a lab setting, communicating orally with a confederate about the target figures.

When in the speaker role (Figure 2), a significantly higher proportion of participants (Group Varied) exhibited syntactic variations in Exp2 (64.56%) and Exp3 (88.89%) compared to Exp1 (44.30%). The remaining participants consistently used a single syntactic structure, predominantly the pre-nominal structure. In Group Varied, the informative-first linearization preference was observed across the three experiments, especially for the Animal property that was more likely to be encoded first in the Animal-informative Condition than in the Action-informative Condition, forming the less preferred post-nominal structure more frequently (Exp1: $\beta = 1.08, SE = 0.19, z = 5.76, p < .01$; Exp2: $\beta = 0.56, SE = 0.21, z = 2.68, p < .05$; Exp3: $\beta = 0.28, SE = 0.12, z = 2.40, p < .05$, using logistic mixed model regression).

Our experiments provide support for the informative-first linearization preference, based on RER, in a subset of participants (Group Varied). Further, this preference is enhanced in more engaging and interactive communication settings. We reason that this may be due to the trial-by-trial alternation between the speaker and listener roles, which required participants to change perspectives more frequently [6;7], resulting in more informative encoding of the utterances for efficient communication.

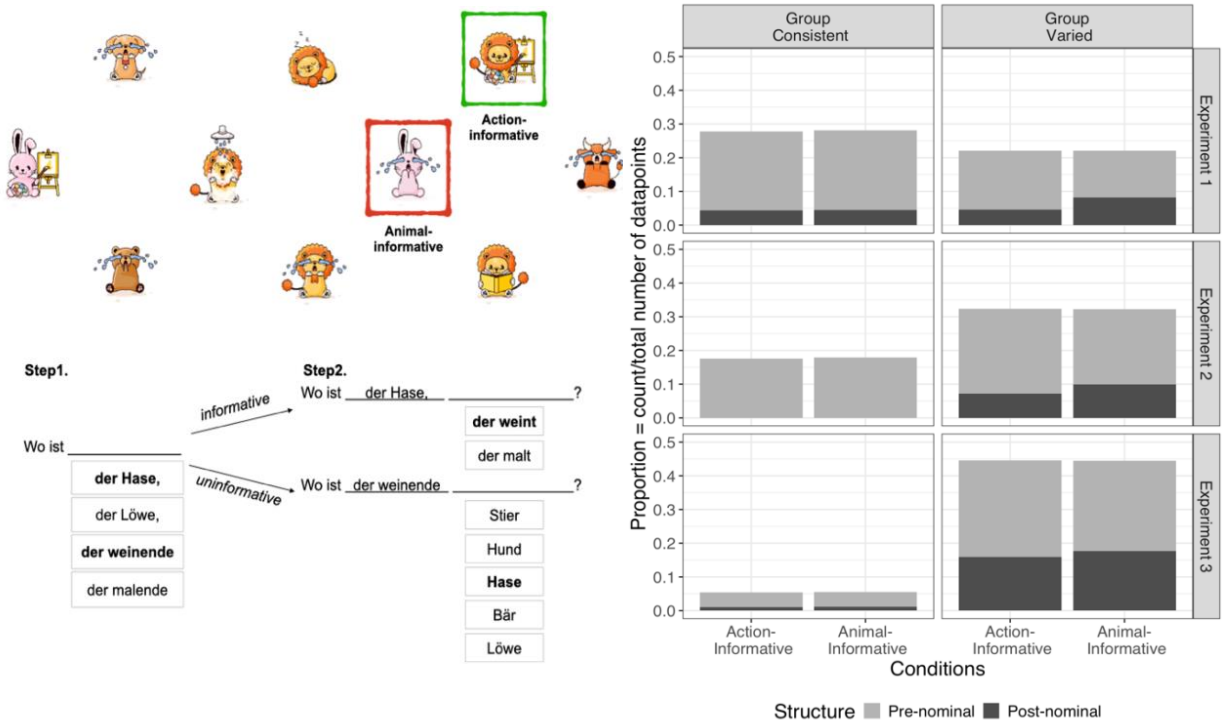


Figure 1 (left). Example visual stimulus and maze-based sentence completion task. Targets as in the Action-informative and Animal-informative Conditions. The informative property narrows down the selection scope from 10 to 2 figures, while the uninformative one narrows from 10 to 5. The two maze steps were presented sequentially. Only one target was highlighted in each trial for the subjects. Only one stem of Step2 was shown, depending on subjects' decisions at Step1.

Figure 2 (right). Proportions of the two modification structures used in each condition in the three experiments. A pre-nominal expression starts with the action property, while a post-nominal modification starts with the animal property.

Reference

[1] Tourtouri, E., Delogu, F., Sikos, L., & Crocker, M. W. (2019). Rational over-Specification in Visually-Situated Comprehension and Production. *Journal of Cultural Cognitive Science* 3: 175–202.

[2] Fukumura, K. (2018). Ordering Adjectives in Referential Communication. *Journal of Memory and Language* 101: 37–50.

[3] Rubio-Fernandez, P., Mollica, F., & Jara-Ettinger, J. (2021). Speakers and Listeners Exploit Word Order for Communicative Efficiency: A Cross-Linguistic Investigation. *Journal of Experimental Psychology: General* 150 (3): 583–94.

[4] Grice, H. P. (1975). Logic and Conversation. In *Syntax and Semantics 3: Speech Acts*. Leiden, The Netherlands: Brill.

[5] Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science* 336 (6084): 998–98.

[6] Sikos, L., Venhuizen, N., Drenhaus, H., & Crocker, M. W. (2021). Speak Before You Listen: Pragmatic Reasoning in Multi-Trial Language Games. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, 1465–71. Vienna,Austria: Cognitive Science Society.

[7] Vogels, J., Howcroft, D. M., Tourtouri, E., & Demberg, V. (2020). How Speakers Adapt Object Descriptions to Listeners Under Load. *Language, Cognition and Neuroscience* 35 (1): 78–92.

Searching for the Neurocognitive Mechanisms Underlying Gist and Verbatim Encoding

Julia Meßmer & Axel Mecklinger (Experimental Neuropsychology Unit, Saarland University)
Julia.messmer@uni-saarland.de

Many higher cognitive functions, such as language processing, benefit from our brains' ability to organize episodic memories within a network of accumulated knowledge, which can be drawn from when interpreting input or generating output. Creating such a network requires organizing the plethora of single detailed (henceforth 'verbatim') memories formed every day. This is achieved by extracting commonalities from specific memories (*gist extraction*; Gilboa & Marlatte, 2017), which can follow or run in parallel to the creation of verbatim memory traces (e.g., Gilboa & Moscovitch, 2021). The goal of the current study (planned work) is to shed light on the neurocognitive mechanisms underlying the successful encoding of verbatim as compared to less-detailed gist traces by using online EEG measures of successful memory formation.

We use a modified version of an associative memory paradigm introduced by Cheng and Rugg (2004; 2010), in which participants study lists of semantically overlapping arbitrary word pairs. In a subsequent memory test, participants must classify originally studied pairs ('Old pairs') as 'old' and reject distractors as 'new' (a schema of the experimental design is depicted in Figure 1). The strength of the paradigm lies in its ability to disentangle the neurocognitive mechanisms underlying verbatim and gist encoding. This is achieved by two characteristics of the design:

First, there are different types of distractors in the memory test that vary in their degree of semantic overlap with the studied word pairs and thereby in whether their rejection can be based on gist or verbatim memory traces: 'Intra pairs' are recombinations within a list. By this, they maintain full semantic overlap to the original word pairs and thus, their successful rejection requires the formation of verbatim memory traces. In contrast, 'Inter pairs' are across-lists combinations of original word pairs, which are not fully semantically overlapping with the original word pairs anymore. By this, they can be rejected on the basis of gist traces, only. By contrasting encoding ERPs on correctly rejected 'Intra pairs' versus false alarms on 'Intra pairs', neural activity related to verbatim trace formation can be isolated.

Second, to isolate neural activity related to gist trace formation, we additionally manipulated the degree of semantic overlap during encoding in order to directly vary demands on gist extraction and verbatim trace formation: Higher semantic similarity in the high gist (HG; for example 'Pfeil-Forelle', 'Bogen-Karpfen') versus low gist (LG; for example 'Theorie-Rochen', 'Praxis-Koi') condition should impose higher processing demands on the formation of verbatim traces and potentially on gist extraction. Semantic similarity is operationalized as the relative frequency of a pair's second word (e.g., 'Koi'), being reported as a member of the shared category ('Fish'; Glauer et al., 2007). German stimuli are newly created based on category exemplar words from Glauer et al. (2007).

To the extent to which a parietal event-related potential (ERP) positive slow wave reflects the successful formation of verbatim, item-specific memory traces (Mecklinger & Kamp, 2023), we expect a larger early parietal ERP effect on subsequent 'Intra pair' correct rejections in the HG versus LG condition. Second, the ERP difference between subsequent 'Intra pair' false alarms in the HG versus LG condition should resemble the early frontal subsequent memory effect, which is indicative of semantic processing in the service of memory encoding (see Mecklinger & Kamp, 2023), reflecting neural activity related to successful gist trace formation. Based on a power analysis, we plan to sample at least $N = 24$ participants.

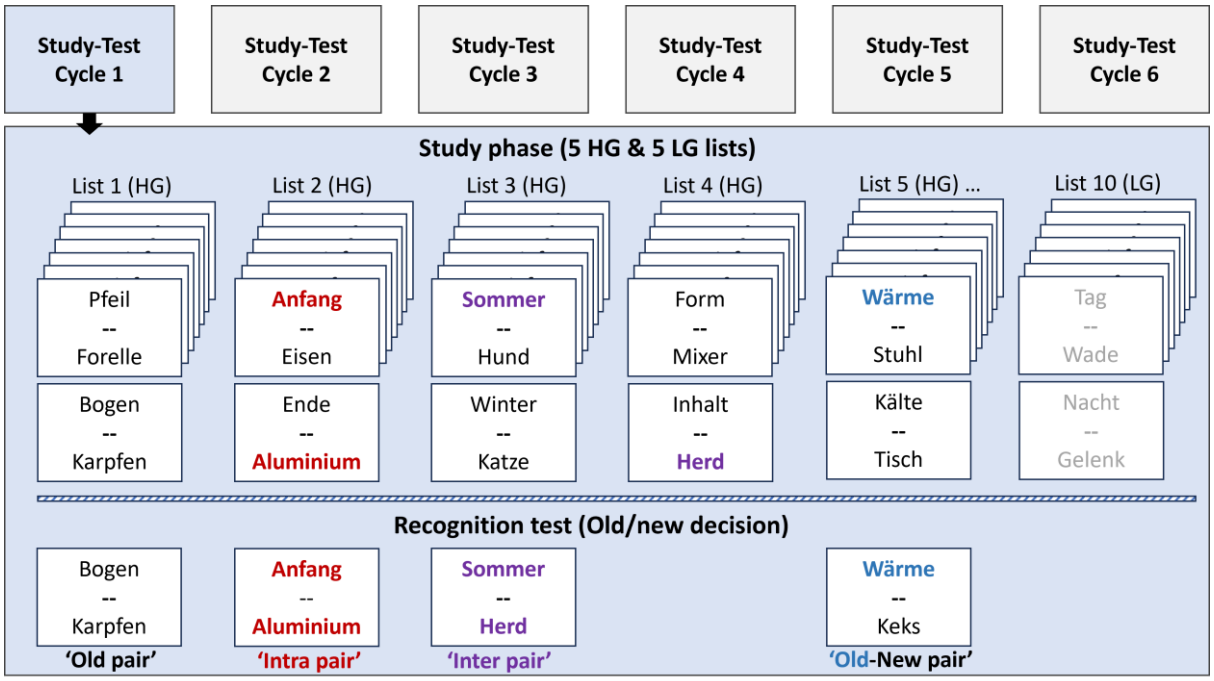


Figure 1: Experimental design (adapted from Cheng & Rugg, 2004; 2010). Participants will complete six study-test cycles. In each study phase, 10 lists are presented, each consisting of eight highly (high gist, HG) or moderately (low gist, LG) semantically overlapping arbitrary word pairs. After each study phase, a recognition test is performed in which original 'Old pairs' (black) must be discriminated from three types of distractors pairs: 'Intra pairs' (red) 'Inter pairs' (purple) and 'Old-New pairs'(blue-black). Note that 'Inter pairs' are separately created for the HG and LG condition.

References

Cheng, S., & Rugg, M. D. (2004). An event-related potential study of two kinds of source judgment errors. *Cognitive Brain Research, 22*(1), 113–127. https://doi.org/10.1016/j.cogbrainres.2004.08.003

Cheng, S., & Rugg, M. D. (2010). Event-related potential correlates of gist and verbatim encoding. *International Journal of Psychophysiology, 77*(2), 95–105. https://doi.org/10.1016/j.ijpsycho.2010.04.010

Gilboa, A., & Marlatte, H. (2017). Neurobiology of Schemas and Schema-Mediated Memory. *Trends in Cognitive Sciences, 21*(8), 618–631. https://doi.org/10.1016/j.tics.2017.04.013

Gilboa, A., & Moscovitch, M. (2021). No consolidation without representation: Correspondence between neural and psychological representations in recent and remote memory. *Neuron 109*(14), 2239–2255. https://doi.org/10.1016/j.neuron.2021.04.025.

Glauer, M., Häusig, S., Krüger, M., Betsch, T., Renkewitz, F., Sedlmeier, P., & Winkler, I. (2007). Typizitätsnormen für Vertreter von 30 Kategorien. *Neurolinguistik, 21*, 21-31.

Mecklinger, A., & Kamp, S.-M. (2023). Observing memory encoding while it unfolds: Functional interpretation and current debates regarding ERP subsequent memory effects. *Neuroscience & Biobehavioral Reviews, 153*, 105347. https://doi.org/10.1016/j.neubiorev.2023.105347

Listeners Adapt to Speakers' Pragmatic Competence

Kata Naszádi[1], Alexandra Mayn[2], John Duff[2], Vera Demberg[2]
[1]University of Amsterdam, [2]Saarland University
k.naszadi@uva.nl

A message that appears ambiguous under literal interpretation might be successfully resolved using pragmatic reasoning about the speaker's intentions and the communicative context. The Rational Speech Act framework (Goodman & Frank, 2016) formalizes this as Bayesian reasoning over the behavior of a cooperative partner. Such reasoning on the listener's side is the right strategy if the speaker also engaged in pragmatic reasoning during production. On the other hand, the listener's pragmatic effort to resolve ambiguities may lead to the wrong interpretation if the speaker did not select her utterance cooperatively. Previously, Mayn et al. (2024) showed that people adjust their interpretations based on information about the speaker: participants were less likely to interpret a child speaker pragmatically than an adult speaker. In our study, instead of revealing the speaker's pragmatic profile, we expect the listeners to adjust their application of reasoning based on **task success** during repeated interaction with the same partner. Bottom-up adjustments like this have been noted in work on contrastive inferences from scalar adjectives (Ryskin et al., 2019).

Method We situate our participants in a collaborative reference game (Frank & Goodman, 2012), where they play the role of listeners. Participants are paired with two partners, one of which follows a **pragmatic** (S_1) and the other a **literal** (S_0) production strategy. Each participant is exposed to both speakers across two blocks in a randomized order.

Candidate images have one of three possible shapes and three possible colors. A trial consists of three candidate images, a set of four available shape and color messages, and the message sent by the speaker. On critical trials, the speaker's message is **ambiguous** and can be literally true of two possible referents. With a pragmatic partner, applying reasoning about the alternative messages will always yield the correct referent. With a literal speaker, however, the ambiguous message may apply to any literally valid candidate, hence applying pragmatic reasoning will sometimes result in choosing an incorrect object. After each trial, we reveal the speaker's intended referent. Figure 1 shows an example critical trial from the literal speaker block.

Note that in critical trials, the pragmatically plausible target only has one available message, while there are always two messages for the competitor. Thus, with a literal speaker choosing from the valid messages by chance, the ambiguous message will have the pragmatically plausible candidate as the target in ⅔ of trials, and the competitor as the target in the remaining ⅓.

We also record participants' confidence about their selections on a 4-point scale throughout the experiment, and examine how it changes as they gain experience with each speaker's behavior.

Each block contains 24 critical and 8 filler items. In the filler items, the message is unambiguous.

Hypothesis If participants adapt to their partner's behavior, their confidence ratings on critical trials in the literal speaker block should decrease through exposure. Responses on filler items should not change.

Results 96 participants were recruited on Prolific. Their responses (Figure 2) were analyzed by fitting ordinal mixed effect regressions to a combination of referent selection and confidence rating (Table 1). On critical trials, participants widely preferred the pragmatically-correct referent in both blocks, but expressed more confidence in their interpretations when interacting with the pragmatic speaker. Confidence developed through repeated interactions, differently for the two speaker types, growing with experience in pragmatic speaker blocks, and falling in literal speaker blocks. Filler trials showed none of these effects.

Discussion When interacting with a literal speaker, participants tended to select the pragmatic target, but with a lower confidence than when interacting with a pragmatic speaker. We take this to provide further evidence that comprehenders can quickly adjust pragmatic interpretation to the demonstrated competence of a partner. While this adjustment is only evident here in metacognitive responses, in a natural interaction lower confidence could drive comprehenders to request clarification. In a follow-up study we will test this prediction with a more naturalistic paradigm.

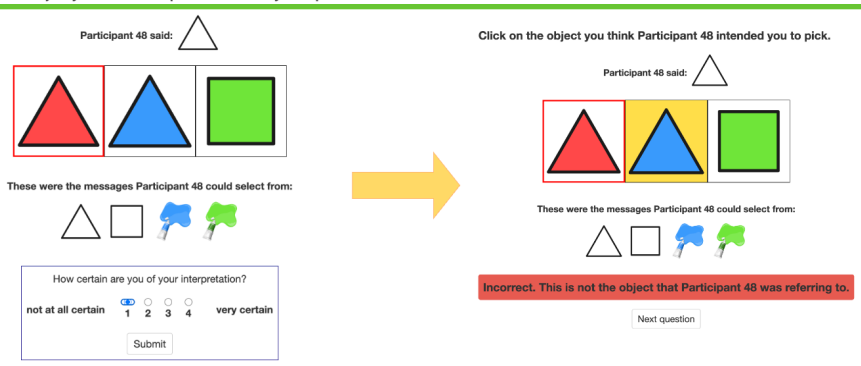


Figure 1: Example trial item from the literal speaker block. The participant chooses the pragmatically plausible interpretation which ends up being wrong. This is the learning signal for the participant that the speaker is picking messages without reasoning about alternative messages.

Parameter	$\hat{\beta}$	95% HDPI
Speaker Type (S_1)	0.72	(0.54, 0.90)
Block Number (Block 2)	-0.02	(-0.19, 0.15)
Quarter of Block	0.13	(0.05, 0.22)
Speaker Type $\times$ Block Number	-0.05	(-0.34, 0.23)
Speaker Type $\times$ Quarter of Block	0.22	(0.14, 0.31)
Block Number $\times$ Quarter of Block	-0.08	(-0.16, -0.01)
Speaker Type $\times$ Block Number $\times$ Quarter of Block	0.00	(-0.08, 0.07)

Table 1: Excerpted parameters from Bayesian ordinal regression fit in `brms` to responses in critical items. Binary factors were sum-coded (-1, 1), with positive levels indicated in parentheses. Quarter of Block was centered. Effects are taken as noteworthy if the 95% highest density posterior interval excludes 0.

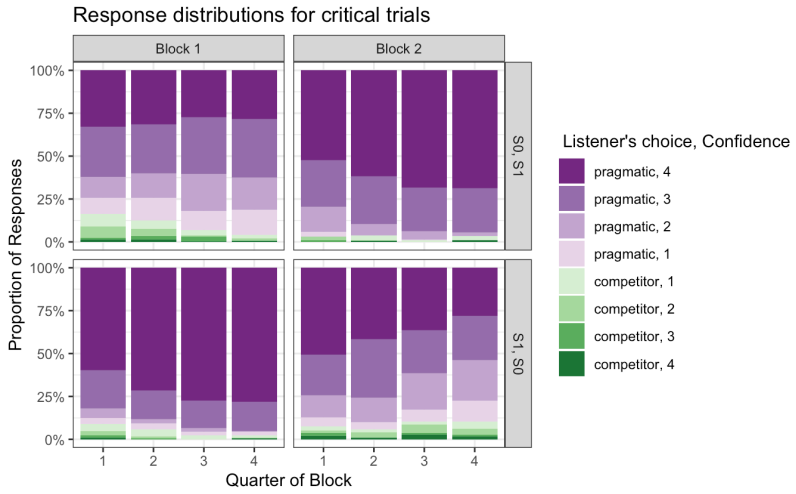


Figure 2: Listeners' responses over the course of the interaction. In the first row we see participants who encountered the literal S_0 first, then pragmatic S_1, in the second row the speaker order is reversed. Responses are shown on the combined choice/confidence scale used for analysis.

Preregistration

https://osf.io/w7yqg?view_only=fa7f9034042946f0928d2c772e0a23ad.

References

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998–998.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Mayn, A., Loy, J. E., & Demberg, V. (2024). Beliefs about the speaker's reasoning ability influence pragmatic interpretation: Children and adults as speakers [Manuscript hosted on PsyArXiv, to appear in *Open Mind*].

Ryskin, R., Kurumada, C., & Brown-Schmidt, S. (2019). Information integration in modulation of pragmatic inferences during online language comprehension. *Cognitive Science*, *43*(8), e12769.

Classroom Dialogic Interaction: Contextual Variability of Allo-repetitions[1]

Prokaeva Valeriya **(**valerie.prokaeva21@gmail.com**)** Saint Petersburg State University

In the research on effective classroom interaction, the two-way nature of communication [Frelin, Grannäs 2010] is of particular interest. By examining classroom dialogue, we aim to identify linguistic features of classroom communication, and estimate the impact of various situational contexts on it. According to the Interactive alignment theory, aligned mental representations in speech manifest themselves linguistically through repetitions [Pickering, Garrod 2004; Branigan et al. 2014]. Based on the theoretical statements of the model, we selected discursive allorepetitions [Tannen 1987; Dumitrescu 1996] as a measurable feature of communicative interaction.

We examined grades 5 to 8 as a key contextual factor influencing classroom interaction, hypothesizing that the students' age might affect the linguistic characteristics of interaction in the classroom in different ways. The structural aspects of allo-repetitions included 1) echo and modified repetitions (expanded, reduced or reformulated), 2) distant/contact repetitions, the latter marked by one speaker repeating the previous speaker's statement upon taking their turn. We also considered repetition spontaneity (forced/unforced) and functions: accepting and recontextualizing. In total, we analyzed 24 lessons (about 40 minutes each) from 12 teachers, including literature and native language lessons, suggesting that the structure of repetition in teachers' speech would vary by grade: echo and reduced repetitions would decrease in higher grades, while expanded and reformulated ones increase. Additionally, we expected contact repetitions to dominate in younger grades and distant – to be more common in higher grades, reflecting qualitative shifts in alignment driven by the need to address more complex ideas [Girolametto, Weitzman 2002]. Similarly, we anticipated a decline in forced repetitions in students' speech in lower grades, corresponding to an increase in learner initiative. The functions of repetitions in teachers' speech likely reflect individual strategic choices.

The analysis identified a total of 3320 instances of repetition, 2415 in teachers' speech, 905 in students' speech. We developed models to examine whether grade influences the frequency of different repetition types, with a teacher and a specific lesson treated as random factors. We used Generalized Linear and Mixed Models (GLM, GLMM), implemented via the glm and glmer functions from the lme4 package. For each of the dependent variables, we constructed a set of models: a simple one with the grade factor, and two mixed ones with random factors including the teacher – lesson interaction. Model fitting was performed using the MLE, and model quality was assessed using the AIC. For the structural variables, the mixed model with fixed effect of a grade and a random effect of a lesson performed best, for the spontaneity and function the best was the model with a single fixed effect of grade**.**

Structurally, the probability of using echo repetitions decreased significantly for grades 7 ($\beta = -0.681$, $p = 0.030$) and 8 ($\beta = -0.758$, $p = 0.028$) compared to grade 5. In contrast, reformulation was more frequently used in grade 7 compared to grade 5 ($\beta = 0.724$, $p = 0.009$). We also found that contact repetitions tend to become more common in higher grades, 7 ($\beta = -0.4888$, $p = 0.026$) and 8 ($\beta = -0.8195$, $p < 0.001$). We found no differences for expansion and reduction, as well as for spontaneity, and any of the functional aspects. Our results suggest that grade may influence certain structural aspects of allo-repetition, particularly, the use of echo-repetition and reformulation, as well as the frequency of contact repetitions in higher grades. The absence of differences in functional aspects might indicate the consistent individual strategies of allo-repetitions regardless of grade.

[1]The research project is supported by Saint Petersburg State University, code 103923108

References

Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*(9), 2355–2368. https://doi.org/10.1016/j.pragma.2009.12.012

Dumitrescu, D. (1996). Rhetorical vs. nonrhetorical allo-repetition: The case of Romanian interrogatives. *Journal of Pragmatics*, *26*(3), 321–354. https://doi.org/10.1016/0378-2166(95)00052-6

Frelin, A., & Grannäs, J. (2010). Negotiations left behind: In-between spaces of teacher–student negotiation and their significance for Education. *Journal of Curriculum Studies*, *42*(3), 353–369. https://doi.org/10.1080/00220272.2010.485650

Girolametto, L., & Weitzman, E. (2002). Responsiveness of child care providers in interactions with toddlers and preschoolers. *Language, Speech, and Hearing Services in Schools*, *33*(4), 268–281. https://doi.org/10.1044/01611461(2002/022)

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(02). https://doi.org/10.1017/s0140525x04000056

Tannen, D. (1987). Repetition in conversation: Toward a poetics of talk. *Language*, *63*(3), 574. https://doi.org/10.2307/415006

Downstream Effects of Prediction on Word Recognition — The Influence of Working Memory Load and Capacity

Linda Sommerfeld[a], Katja Haeuser[a], Arielle Borovsky[b], Jutta Kray[a]
([a]Saarland University, [a]Purdue University)

Listeners use prior sentence contexts to predict upcoming words which can facilitate processing of these words.[1/2] Prediction not only has immediate effects on word processing, but also downstream effects on memory encoding of words. We aimed to replicate that prediction can initiate the formation of memory representations.[3/4/5] We also sought to examine how working memory load modulates prediction-driven forming of memory representations. Specifically, we tested two views against each other: Forming memory representations could benefit from short linguistic contexts where less information has to be kept in working memory during prediction, creating smaller working memory load.[6] Otherwise, longer sentence contexts may allow predictions to linger in working memory for a longer time, causing stronger representations.[7/8]

In three self-paced reading studies German adults read predictable sentences ending with plausible target words of *low* predictability (e.g., *To open the door Jens looks for the handle*). We manipulated working memory load: Study 1 presented short and long sentences. The distance (i.e., the number of words) between the predictive context and the target word consisted of four additional words in the long sentences. In study 2, the distance manipulation consisted of up to nine additional words in the long vs. the short sentences. Here, we also showed long sentences with an additional semantic cue prior to the target word (e.g., *below the doormat*) that should support lingering of predictions in working memory. In study 3, the target word was shown either in the mid or end of a sentence to control whether words in the end position, i.e. words that do not need to be kept in working memory across the whole sentence, allow stronger representations.

In all studies, we tested readers' (n = 80) memory for presented target words (e.g., *handle*), predicted but not presented lure words (e.g., *key*), and unrelated new words (e.g., *message*). In study 1, memory was also tested for unpredictable but semantically related context lures (e.g., *entry*). Table 1 shows an example item. All studies included two working memory span tasks.

GLMMs on the proportion of "old" ratings for the recognition words with the factors word type and sentence type revealed for each study that readers successfully discriminated old target words from new words while showing more false alarms to predicted lure vs. new words. Thus, predictable words lingered in memory even when predictions were disconfirmed, meaning that prediction has long-term effects on cognition. In study 1, memory did not differ for new words vs. context lures, showing that the effect did not derive from semantic association but from prediction. Study 2 found no evidence that additional cues supporting lingering of predictions in working memory affect recognition. In sum, we found no effect of the working memory manipulation (short vs. long distance; mid vs. end position). However, individual differences in working memory skill affected false memory. In study 1, higher working memory skill was related to fewer false alarms for lure words, showing that working memory plays a role for the prediction-driven forming of memory representations. In sum, we show that prediction affects memory, but future studies may test working memory manipulations with more complex linguistic structures to ascertain the impact of working memory load on memory representations.

Table 1

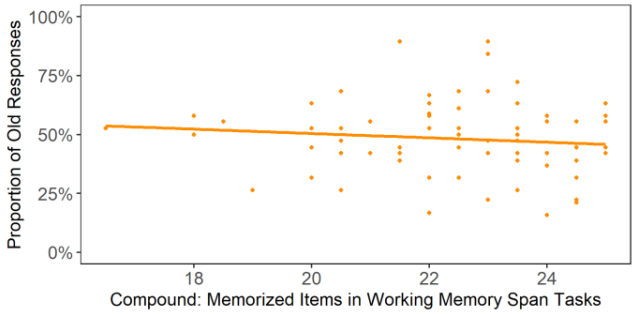
Example Item

Study	Condition	Sentence	Dist.
1	Short	**Weil Jens die Haustür öffnen möchte**, sucht er den eisernen **Griff** unter dem Stein.	4
1	Long	**Weil Jens die Haustür öffnen möchte**, sucht er den von einem Handwerker gefertigten eisernen **Griff** unter dem Stein.	8
2	Short	**Weil Jens die Haustür öffnen möchte**, sucht er den eisernen **Griff**.	4
2	Long	**Weil Jens die Haustür öffnen möchte**, sucht er unter dem Stein den von einem Handwerker gefertigten, eisernen **Griff**.	11
2	Long, additional cue	**Weil Jens die Haustür öffnen möchte**, sucht er unter der Fußmatte den von einem Handwerker gefertigten, eisernen **Griff**.	11
3	End position	**Weil Jens die Haustür öffnen möchte**, sucht er den von einem Handwerker gefertigten, eisernen **Griff** unter dem Stein.	8
3	Mid position, short	Jens sucht, **weil er die Haustür öffnen möchte**, den eisernen **Griff** unter dem Stein, den ein Handwerker gefertigt hatte.	2
3	Mid position, long	**Weil Jens die Haustür öffnen möchte**, sucht er den eisernen **Griff** unter dem Stein, den ein Handwerker gefertigt hatte.	4

Note. An item in its conditions across the studies with the predictive context and target word in bold. For study 2, the additional semantic cue is underlined. In all studies, memory was tested for the target word Griff, the lure word Schlüssel, and the new word Nachricht. In study 1, it was also tested for the context lure Eingang. Dist. (distance) is the number of words between the context and target word.

Figure 1

Recognition of Lure Words Depending on Working Memory Capacity (Study 1)



Note. Increases in working memory skill (indicated by the compound score of the number of memorized items in two working memory span tasks) was associated with fewer old ratings for lure words.

References

[1] Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044.

[2] Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135.

[3] Haeuser, K. I., & Kray, J. (in press.) Age differences in context use during reading and downstream effects on recognition memory.

[4] Hubbard, R. J., Rommers, J., Jacobs, C. L., & Federmeier, K. D. (2019). Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. *Frontiers in Human Neuroscience, 13,* Article 291.

[5] Haeuser, K., & Kray, J. (2022). Uninvited and unwanted: False memories for words predicted but not seen. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society* (pp. 2401–2408).

[6] Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, & Brain*, *2000*, 95–126.

[7] Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

[8] Levy, R. P. and Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*, *68*(2), 199–222.

Information structure and cognitive states in the structure of German sentences

Luigi Talamo, Andrew Dyer, Annemarie Verkerk (Saarland University)
luigi.talamo@uni-saarland.de

The word order of German and English is usually described as quite rigid, however with substantial differences. Thanks to case marking, German can easily flip its unmarked SVO order and bring non-subject elements to the topic position, while English has to use specific constructions (Durrell 2017: 935-936). By contrast, the structure of German sentences is organized around the predicate into three topological fields: the prefield, the midfield and the postfield, with specific restrictions for certain constituents; for instance, adverbial, predicate and arguments are forbidden in the postfield. Accordingly, German can easily modify the order of verbal arguments for information structure. But what about the order of elements in the sentence? And how does this interact with the cognitive accessibility of these elements, which Gundel, Hedberg and Zacharski (1993) has described as the Givenness hierarchy?

We conduct a quantitative analysis on 2,500 German sentences from miniCIEP+ (Verkerk and Talamo 2024), a parallel corpus parsed according to the Universal Dependency (UD) framework and annotated for information structure using the schema described in Anonymous (2024). The annotation provides referents with labels describing the information status (given vs. new) as well as mention type (anaphor, cataphor, predicate, apposition, discourse deixis and lexical coreference). The annotation is not currently available for German, but we plan to project it from English to German sentences using AWESOME, a word-by-word aligner (Dou & Neubig 2021). For each annotated mention of a referent, we extract features building up the German forms of the six types of cognitive states described by the givenness hierarchy (Gundel, Hedberg and Zacharski 1993): uniquely identifiable (indefinite article + N), referential (indefinite pronoun/referential expression + N), uniquely identifiable (definite article + N), familiar (distal demonstrative + N), activated (distal/proximate demonstrative, proximate demonstrative + N) and in focus (personal pronoun). We then extract the information structure (status and coreference) and compute the relative position of the referent in the sentence, operationalized as a scale ranging from 0 (at the very beginning of the sentence) to 1 (at the very end of the sentence). For instance, take the German sentence *Im letzen Jahr war er zu dem ersten Mal dort gewesen* ‘‘Last year he has been there for the first time’, whose annotation is shown in Figure 1; the three annotated mentions are: *Im letzen Jahr*, which is in the referential state, *er*, which is the focus state, and *dort*, which is in the activated state.

We fit a non-linear regression models with the information status as the response variable and relative position, coreference and type of cognitive state as the independent variables. We expect that given referents show lower values for the relative position i.e., when they appear at the beginning of the sentence and that different degrees of coreference play a meaningful role in the sentence structure and processing, similarly to what is discussed for English in Ye, Tu, and Pustejovsky, 2023.

```
1-2  Im  _  _  _  _  _  _  _  _
1  In  in  ADP  APPR  _  4  case  _  _
2  dem  der  DET  ART  Case=Dat|Definite=Def|Gender=Neut|Number=Sing|PronType=Art  4  det  _  Entity=(e20-Time-1-CorefType:coref,InfStat:new
3  letzten letzt  ADJ  ADJA  Case=Dat|Degree=Pos|Gender=Neut|Number=Sing  4  amod  _  _
4  Jahr  Jahr  NOUN  NN  Case=Dat|Gender=Neut|Number=Sing  12  obl  _  Entity=e20)
5  war  sein  AUX  VAFIN  Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin  12  aux  _  _
6  er  er  PRON  PPER  Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs  12  nsubj  _  Entity=(e1-Person-1-CorefType:ana,InfStat:given)
7-8  zum  _  _  _  _  _  _  _  _
7  zu  zu  ADP  APPR  _  10  case  _  _
8  dem  der  DET  ART  Case=Dat|Definite=Def|Gender=Neut|Number=Sing|PronType=Art  10  det  _  _
9  ersten  erst  ADJ  ADJA  Case=Dat|Degree=Pos|Gender=Neut|Number=Sing|NumType=Ord  10  amod  _  _
10  Mal  Mal  NOUN  NN  Case=Dat|Gender=Neut|Number=Sing  12  obl  _  _
11  dort  dort  ADV  ADV  _  12  advmod  _  Entity=(e18-Place-1-CorefType:ana,InfStat:given)
12  gewesen  sein  AUX  VAPP  VerbForm=Part  0  root  _  SpaceAfter=No
13  .  .  PUNCT  $.  _  12  punct  _  _
```

Figure 1. The sentence *Im letzen Jahr war er zu dem ersten Mal dort gewesen* 'Last year he has been there for the first time' annotated for Universal Dependencies (column 1-9) and for information structure (column 10).

References

Anonymous (2024). A Multilingual Parallel Corpus for Coreference Resolution and Information Status in the Literary Domain.

Dou, Zi-Yi and Graham Neubig. (2021). Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2112–2128, Online. Association for Computational Linguistics.

Durrell, Martin (2017). Hammer's German Grammar and Usage (Routledge Reference Grammars). Sixth Edition. Abingdon, Oxon ; New York, NY : Routledge.

Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski (1993). "Cognitive Status and the Form of Referring Expressions in Discourse". In: Language 69.2, pp. 274–307.

Verkerk, Annemarie and Luigi Talamo (2024). "mini-CIEP+ : A Shareable Parallel Corpus of Prose". In: Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024. Ed. by Pierre Zweigenbaum, Reinhard Rapp, and Serge Sharoff. Torino, Italia: ELRA and ICCL, pp. 135–143.

Ye, Bingyang, Jingxuan Tu, and James Pustejovsky (2023). "Scalar Anaphora: Annotating Degrees of Coreference in Text". In: Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023). Ed. by Maciej Ogrodniczuk et al. Singapore: Association for Computational Linguistics, pp. 28–38. doi: 10.18653/v1/ 2023.crac-main.4.

Modeling Decision Problems for Relevant Answers to Polar Questions

Polina Tsvilodub, Michael Franke, Robert Hawkins
polina.tsvilodub@uni-tuebingen.de

Imagine you are working as a barista at a coffeeshop. A customer asks "Do you have iced tea?" but you've run out. They have asked a yes-no (or *polar*) question, so you should respond "no", as suggested by classic accounts of questions in linguistics (Hamblin, 1973). However, this minimal answer is intuitively unsatisfying. Instead, you may prefer to say something like "No, I'm afraid we're out of iced tea *but we do have iced coffee*", mentioning a *relevant alternative* (Clark, 1979).

In previous work, we proposed a novel cognitive model of pragmatic overinformative question answering (the PRIOR-PQ model) and empirically evaluated some of its key predictions (anonymous, n.d.). We formulated our model in the tradition of the Rational Speech Act (RSA) framework (Frank & Goodman, 2012), couching it in an *action-oriented* definition of relevance hinging on the questioner's *decision problem* (DP) (van Rooy, 2003). However, one limitation of many cognitive models like RSA is the necessity to elicit auxiliary intuitive world knowledge in costly human experiments. One potential alternative to human data are predictions of SOTA large language models (LLMs). Yet to maintain the quality of the cognitive model, careful testing of LLM-supplied data in the context of the model is needed.

In this work, we test predictions of `gpt-4o-mini` for intuitive information about the DP in PRIOR-PQ. PRIOR-PQ captures a cooperative answerer that chooses an answer increasing the expected utility of the questioner's future actions under their DP. The DP is a tuple consisting of a set of world states, a set of actions, a *utility function*, and a probability distribution capturing the questioner's prior beliefs about the world states (see Fig. 1). The model is presented in formal detail in Fig. 1. We compared the predictions of PRIOR-PQ to human data in two experiments (case study 2, 3). In both, a polar question about a target appeared in a context presenting available options (but not the target) which varied in terms of their practical utility for the questioner (example vignettes are below). We elicited free production responses from humans ($N = 162$ and $N = 130$). To model the inference about the likely questioner DP in PRIOR-PQ and predict the optimal answer, we modeled four or five types of DPs, one corresponding to each of the available options (see example). Each DP was associated with different utilities, or, payoffs for each other option, given a target option. Supplied with utility ratings elicited in human experiments (slider ratings, $N = 453$ and $N = 130$), the model's predictions aligned well with human data, particularly capturing the preference for overinformative *competitor* responses mentioning only a relevant option (Fig. 2A).

Here, we explore whether DP utilities sampled from `gpt-4o-mini` given the prompt from human experiments align with human ratings. The utilities were sampled with temperature $\tau = 0.1$, given the additional instruction to produce ratings from 0 to 100 instead of a slider, for ten iterations, for each option pair. The text predictions were cast to numbers. Figure 2(B) shows predicted utilities for each item, averaged over runs, against results from the human experiments for both case studies, indicating high correlation ($R^2 = 0.92$ and 0.87). The order of preferences for different alternatives (e.g., "iced coffee" vs. "Chardonnay"), given a target ("iced tea"), corresponds to intuitions for our vignettes for both human and LLM results. These results provide a promising avenue for ongoing work in which we integrate LLM utility predictions into PRIOR-PQ simulations. Including LLMs in PRIOR-PQ, given careful comparison of LLM and human results, provides a promising avenue towards scaling up rational cognitive models.

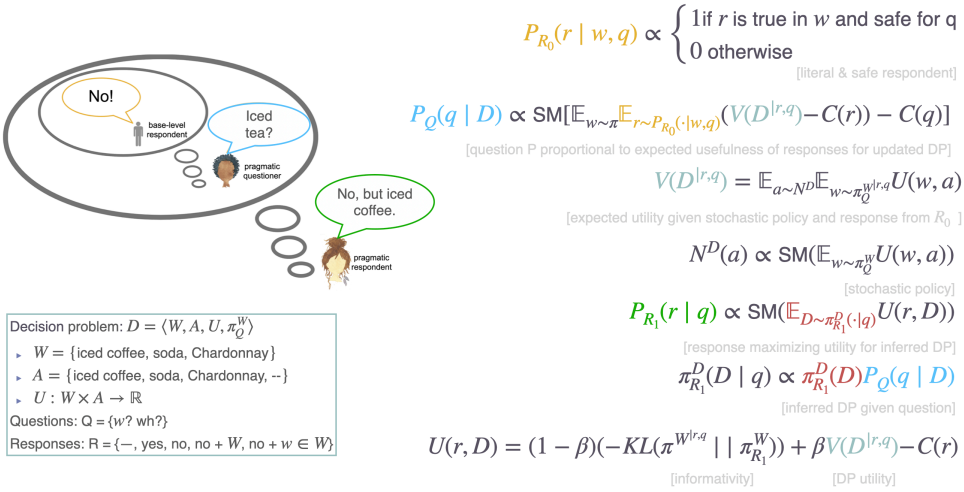


Figure 1: PRIOR-PQ model overview. The pragmatic answerer R_1 reasons about a questioner Q who selects a question according to the utility of the information for their DP that it is likely to elicit from a safe and true base respondent R_0.

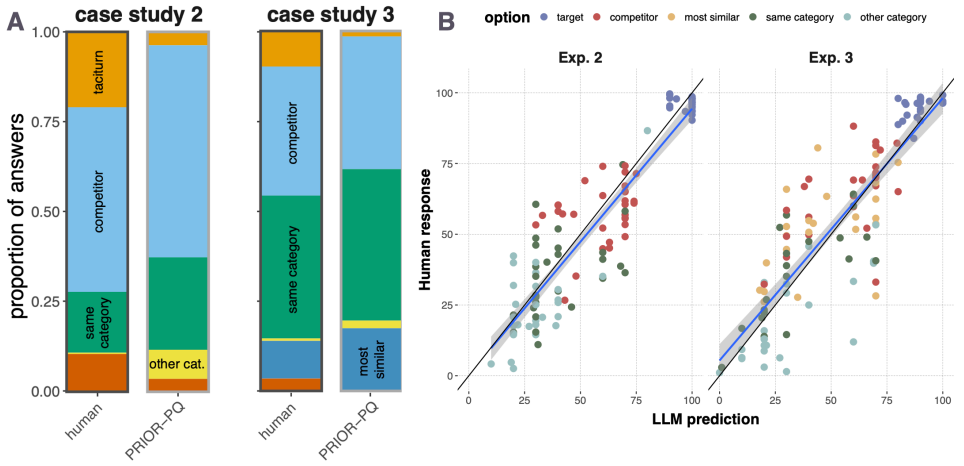


Figure 2: **A**: Proportions of responses mentioning different alternatives (color) in the two experiments, produced by humans and predicted in simulations by PRIOR-PQ. **B**: Mean by-item utilities of different options (color) when the target option (e.g., iced tea) is the goal, predicted by `gpt-4o-mini` against human ratings.

Example vignette from Exp. 2: You are a bartender in a hotel bar. The bar serves only soda (same category), iced coffee (competitor) and Chardonnay (other category). A woman walks in. She says: "Do you have iced tea?" (target) **Example vignette from Exp. 3:** Context 1: Your friend is having a sleepover with some friends on the weekend. [...] Context 2: Your roommate [...] has a large mirror that she needs to pack for transportation. Shared options and question: You have the following items at home that you could spare for some time: some bubble wrap (competitor 2), a pillow (most similar), a sleeping bag (competitor 1) and a carpet (other category). Your friend asks: "Do you have a blanket?" (competitor i was "same category" in other context)

References: Clark, H.H. (1979) Responding to indirect speech acts. *Cognitive Psychology*. Frank, M.C. & Goodman, N.D. (2012) Predicting Pragmatic Reasoning in Language Games. *Science*. Hamblin, C. (1973) Questions in Montague English. *Foundations of Language*. [Redacted for anonymity] (under review) Relevant answers to polar questions. van Rooy, R. (2003) Questioning to resolve decision problems. *Language & Philosophy*

The Role of Surprisal in Perceptual Chunking of Spontaneous Speech

Svetlana Vetchinnikova (University of Helsinki)
svetlana.vetchinnikova@helsinki.fi

Recent studies in cognitive neuroscience suggest that when processing continuous stimuli such as speech, humans rely on periodic neural oscillations across different frequency bands (Giraud & Poeppel, 2012). If this hypothesis holds, the rhythmicity of oscillations imposes temporal constraints on the structure of speech. Specifically, theta-band oscillations are thought to align with the duration of syllables, which tend to be relatively stable both within and across languages (Ding et al., 2017; Varnet et al., 2017). Meanwhile, delta-band oscillations appear to correspond to syntactic phrases (Ding et al., 2016; Kaufeld et al., 2020) and/or intonation units (Inbar et al., 2020). Given the primacy of cognitive constraints, it is plausible that both syntax and prosody have evolved as adaptive mechanisms, facilitating the segmentation of speech into perceptually manageable units for both speakers and listeners. However, what role does statistical information play in this process, given its recognized importance in language processing?

In an earlier study (Vetchinnikova et al. 2023), we selected 97 short extracts from spoken corpora and re-recorded them with a trained speaker to achieve uniform audio quality. We then asked 50 experiment participants to listen to the extracts and intuitively mark chunk boundaries in the accompanying transcripts through a custom-built tablet application. Next, we annotated all spaces between every two words for pause duration, prosodic and syntactic boundary strength, chunk duration, and bigram surprisal. Prosodic boundary strength was estimated automatically using continuous wavelet analysis of fundamental frequency, energy and word duration (Suni et al. 2017). To measure syntactic boundary strength, we marked the start and end of each clause with a bracket and counted the total number of brackets for each space assigning a value of 0.5 to an opening bracket and 1 to a closing bracket. Since pause duration, prosodic and syntactic boundary strength as well as chunk duration were collinear, we built a separate logistic regression model predicting chunk boundary perception for each predictor. All models included random effects for listeners and extracts.

We found that pause duration, prosodic boundary strength, syntactic boundary strength, and temporal duration significantly predicted chunk boundary perception, supporting the influence of the temporal constraint and the role of prosody and syntax in perceptual chunking. In contrast, the effect of bigram surprisal contradicted the hypothesis that perceptual chunks were multi-word units: chunk-final words tended to be less predictable while chunk-initial words tended to be more predictable. We interpreted this finding as evidence of a dissociation between perceptual chunking, or the segmentation of incoming speech into temporal units, and usage-based chunking, or the extraction of statistical regularities from input.

Given the limitations of using bigram surprisal to capture statistical information, in this paper, I use surprisal values derived from the GPT2 model that incorporate the full preceding context of each extract. Preliminary results suggest that full context surprisal does not predict chunk boundary perception. I discuss these results from the perspective of the interplay between statistical and structural information in speech processing.

References

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Inbar, M., Grossman, E., & Landau, A. N. (2020). Sequences of Intonation Units form a ~ 1 Hz rhythm. *Scientific Reports*, *10*(1), 15846. https://doi.org/10.1038/s41598-020-72739-4

Kaufeld, G., Bosker, H. R., & Martin, A. E. (2020). Linguistic Structure and Meaning Organize Neural Oscillations into a Content-Specific Hierarchy. *The Journal of Neuroscience*, *40*(49), 9467–9475.

Suni, A., Šimko, J., Aalto, D., & Vainio, M. (2017). Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, *45*, 123–136. https://doi.org/10.1016/j.csl.2016.11.001

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, *142*(4), 1976–1989. https://doi.org/10.1121/1.5006179

Vetchinnikova, S., Konina, A., Williams, N., Mikušová, N., & Mauranen, A. (2023). Chunking up speech in real time: Linguistic predictors and cognitive constraints. *Language and Cognition*, *15*(3), 453–479. https://doi.org/10.1017/langcog.2023.8

The Object Order in the German Middle Field through the Lens of Information Theory: A Diachronic Study

Sophia Voigtmann (University of Kassel)
sophia.voigtmann@uni-kassel.de

The relative order of dative (Dat) and accusative (Acc) objects in German is variable. Dat>Acc (1a) is the canonical order, but Acc>Dat (1b) can be found when the Acc is given (Lenerz, 1977; Rauth, 2020; Speyer, 2011, 2015, 2016).

1. a) Ich gebe [einem Athleten]$_{Dat}$ [den Ball]$_{Acc}$.
   I give [an athlete] [the ball].
   'I gave an athlete the ball.'
   b) Ich gebe [den Ball]$_{Acc}$ [einem Athleten]$_{Dat}$.
   I give [the ball] [an athlete].
   'I give the ball to an athlete.'

This study proposes two previously unconsidered factors to influence the object order, i.e. prediction of the constituents in a sentence based on the position of the full verb (FVP) and the clauses' information profile. Thus, we test these two hypotheses:

1) Acc>Dat is more likely when the lexical verb precedes the objects.
2) Dat>Acc is more likely when the clause's lexical information profile is uneven.

When the lexical verb containing the valency information follows the objects (FV-VL), Dat>Acc is preferable because recipients discard sentence continuations with a transitive verb earlier and rank those with a di- or intransitive verb higher (Levy, 2008). If the full verb is presented first (FV-V2), Acc>Dat is possible as the necessity of *both* objects is *known*. A certain object order is less crucial to reduce uncertainties.

The second hypothesis refers to the Uniform Information Density (UID) (Levy & Jaeger, 2007): In lexically uneven clauses, it is better to use the more common Dat>Acc as familiarity with a certain construction can facilitate processing even under disadvantageous conditions, i.e.an uneven information profile (e.g. Futrell et al., 2021).

We also want to test the stability of these assumptions over time. Thus, we conducted a corpus study using the Anselm (~16th century), RIDGES (16th/17th century), GerManC corpus (17th/18th century), the Tiger and TüBa-D/Z corpus for modern German. The objects were found automatically. We analyze 1733 clauses here, 76% of them are from modern data. Acc>Dat occurs in 8% of the modern data, 10% of the 17th and 18th century and to 15% in the 16th century. Each clause was annotated for the FVP, DORM (Cuskley et al., 2021), a measure for UID based on unigram-lemma-surprisal[1] of each word in each corpus, the object's length ratio and their givenness status as well as the publication century of each text. A general logistic regression analysis (glm, (R Core Team, 2023))[2] with backward model selection was conducted. We found (among others, Table 1) that the FV-V2 is connected to the Acc>Dat-order in sentences with an uneven information profile in historical but not in modern data where Dat>Acc is generally more frequent (Figure 1). This result is interpreted as an interplay between grammatical and lexical processing difficulties which is used to keep the general processing effort constant. Lexically harder to process sentences have the more common order and vice versa. As more variation was possible in the past, speakers were more sensitive to means of facilitating processing.

Given accusative objects are linked to the Acc>Dat-order, but the likelihood of Dat>Acc increases when the full verb is in the RSB even for given accusative objects

[1] Using lemma unigram surprisal neutralizes any grammatical information and is preferable for DORM.
[2] glm(Object order ~(DORM+Length Ratio+Acc givenness (sum-coded) + Dat givenness (sum-coded) +FVP (sum-coded) + period)^3, data=data, family = "binomial").

(Figure 2). Thus, we find further (and period stable) evidence for the influence of the position of the full verb in line with our prediction.

	Est.	Std. Error	z-value	p-value			Est.	Std. Error	z-value	p-value	
Intercept	1.26	0.25	4.95	<0.001	***	**DORM: Period**	0.07	0.04	1.73	0.08	.
DORM	-0.06	0.07	-0.86	0.39		**Length ratio: FVP**	0.18	0.09	2.03	<0.05	*
Length ratio	-0.83	0.04	-1.88	0.06	.	**$Acc_{Info\text{-}Status}$: FVP**	1.18	0.55	2.15	<0.05	*
$Acc_{Info\text{-}Status}$	-2.11	0.44	-4.78	<0.001	***	**$Acc_{Info\text{-}Status}$: Period**	-0.47	0.25	-1.87	0.06	.
$Dat_{Info\text{-}Status}$	0.08	0.41	0.19	0.84		**$Dat_{Info\text{-}Status}$: Period**	0.65	0.24	2.66	<0.01	**
FVP	0.96	0.49	1.98	0.047	*	**FVP: Period**	-0.63	0.29	-2.19	<0.05	*
Period	0.38	0.16	2.42	<0.05	*	**DORM: FVP**	0.35	0.13	2.67	<0.01	**
DORM: Length ratio	0.02	0.01	2.09	<0.05	*	**DORM: FVP: Period**	-0.16	0.08	-2.05	<0.05	*

Table 1 Results of the regression analysis.

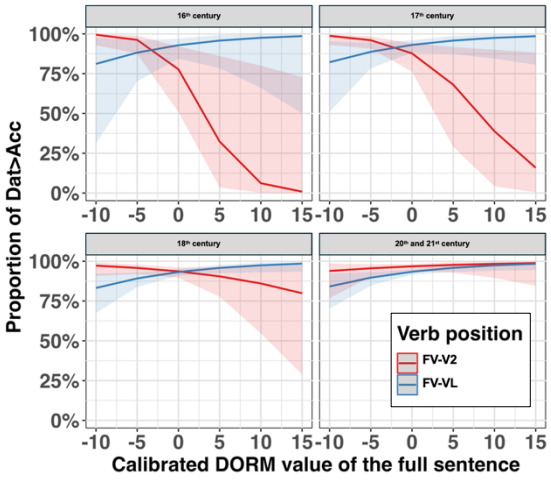


Figure 1 Interaction plot of the variables DORM, position of the full verb and period. A higher DORM indicates less uniformity.

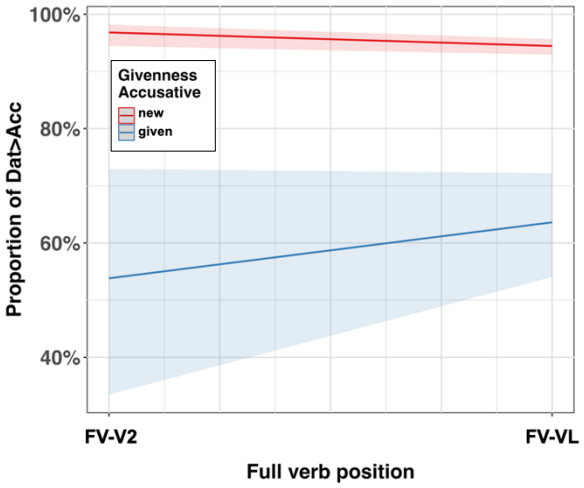


Figure 2 Interaction plot of the position of the full verb and the information status of the accusative object.

References

Cuskley, C., Bailes, R., & Wallenberg, J. (2021). Noise resistance in communication: Quantifying uniformity and optimality. *Cognition*, *214*, 104754. https://doi.org/10.1016/j.cognition.2021.104754

Futrell, R., Gibson, E., & Levy, R. (2021). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, *44*(3), e12814. https://doi.org/0.1111/cogs.12814

Lenerz, J. (1977). *Zur Abfolge nominaler Satzglieder im Deutschen*. Narr.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference* (pp. 849–856). The MIT Press. https://doi.org/10.7551/mitpress/7503.003.0111

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rauth, P. (2020). *Ditransitive Konstruktionen im Deutschen. Geschichte und Steuerung der Objektabfolge im Mittelfeld*. Stauffenburg.

Speyer, A. (2011). Die Freiheit der Mittelfeldabfolge im Deutschen – ein modernes Phänomen. *Beiträge Zur Geschichte Der Deutschen Sprache Und Literatur*, *133*, 14–31.

Speyer, A. (2015). Object order and the Thematic Hierarchy in Older German. In J. Gippert & R. Gehrke (Eds.), *Historical Corpora: Challenges and Perspectives* (pp. 101–124). Narr.

Speyer, A. (2016). The relative object order in High and Low German. In S. Featherston & Y. Vearsley (Eds.), *Quantitative Approaches to Grammar and Grammatical Change. Perspectives from Germanic.* (Vol. 290, pp. 143–164). De Gruyter.

Predictive Potential of Linguistic Distances and Surprisal in Multilingual Intercomprehension Experiments

Iuliia Zaitova, Wei Xue, Irina Stenger, Tania Avgustinova
(Department of Language Science and Technology, Saarland University, Germany)
izaitova@lsv.uni-saarland.de

This study explores the predictive potential of linguistic distances and surprisal in multilingual intercomprehension experiments. Linguistic distances refer to the measurable differences between languages (Wichmann et al., 2010). They can be quantified in various domains, such as phonology and orthography (Gooskens and van Bezooijen, 2013), with each domain contributing differently to the overall distance between languages. Previous research showed that higher linguistic distances were associated with decreased intercomprehension (Gooskens and Swarte, 2017; Moller and Zee-¨ vaert, 2015; Vanhove and Berthele, 2015).

The difficulty in processing a linguistic unit is proportional to the metric of surprisal, as estimated by language models (Hale, 2001; Levy, 2008). Surprisal is defined as the negative log-likelihood of encountering a unit given its preceding context derived from language models (surprisal = $-\log P(w_i \mid context)$ for a given unit w_i in a sequence), and it effectively measures the unpredictability of that unit (Crocker et al., 2016).

Given the above background, we conducted two web-based experiments to examine the intercomprehension of microsyntactic units (specific constructions between the lexicon and the grammar, idiomatic properties of which are closely tied to syntax, see Avgustinova and Iomdin, 2019) in context under different input conditions: (1) spoken and (2) written. Each experiment included two tasks: free translation and multiple choice. Native Russian speakers participated in the experiments covering five closely related Slavic languages (Belarusian, Bulgarian, Czech, Polish, and Ukrainian). We examined the participants' intercomprehension performance through accuracy. We calculated Pearson correlations of the accuracy values with phonologically weighted Levenshtein distance (PWLD), orthography-based Jaccard similarity, and surprisal estimates from Automatic Speech Recognition models, namely Wav2Vec2-Large-Ru-Golos-With-LM (Bondarenko, 2022) and Whisper Medium Russian and language models, namely ruBERTa-large and ruGPT3large (Zmitrovich et al., 2023).

Figure 1 shows the accuracy results for both experiments. In general, we found that spoken input led to higher accuracy values in both tasks except those in the multiple choice task for Ukrainian and Bulgarian, suggesting that the written modality might introduce a confounding factor. As surprisal from ruBERTa-large and PWLD showed stronger correlations in both tasks, we only present those factors in relation to the free translation and the multiple choice tasks, as shown in Table 1. We observed significant correlation of free translation accuracy for all languages together and for Ukrainian individually. As for multiple choice accuracy, significant correlations with PWLD were observed when pooling all languages together, as well as for all languages individually except Belarusian. We also observed stronger correlations in the experiment with written input, especially for the multiple choice task. Overall, this study underscores the predictive potential of surprisal and linguistic distances in multilingual intercomprehension experiments, providing valuable insights for the field of computational linguistics. Future research should expand to diverse language groups to validate these findings and explore their broader applicability.

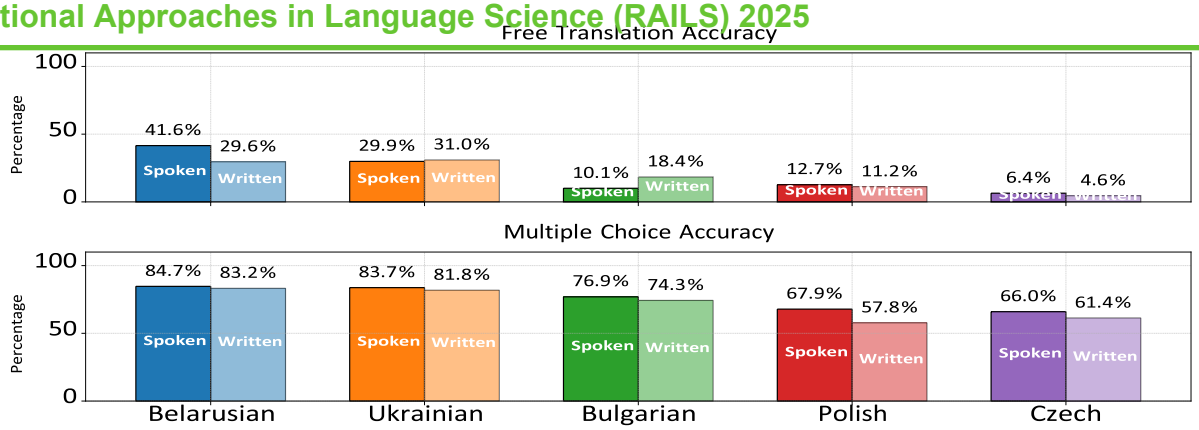


Figure 1: Experimental results for both tasks.

	Free Trans with ruBERTa-large surprisal		Multiple Choice with PWLD	
Language	**Written**	**Spoken**	**Written**	**Spoken**
Belarusian	-0.06	-0.03 (NS)	-0.2 (NS)	-0.15 (NS)
Bulgarian	-0.17 (NS)	0.00 (NS)	-0.42**	-0.33*
Czech	-0.23 (NS)	-0.06 (NS)	-0.28*	-0.25 (NS)
Polish	-0.21 (NS)	-0.16 (NS)	-0.38**	-0.45***
Ukrainian	-0.25*	-0.06 (NS)	-0.50***	-0.43***
All	-0.38***	-0.38***	-0.42***	-0.39***

Note: * = p < .05, ** = p < .01, *** = p < .001, NS = Non-significant

Table 1: Pearson correlation of predictors with accuracy of participants' responses

References

Avgustinova, T., & Iomdin, L. (2019, September). Towards a typology of microsyntactic constructions. https://doi.org/10.1007/978-3-030-30135-4 2

Bondarenko, I. (2022). Xlsr wav2vec2 russian with 2-gram language model by ivan bondarenko.

Crocker, M., Demberg, V., & Teich, E. (2016). Information density and linguistic encoding (ideal). *Kunstliche Intelli-¨ genz*, *30*, 77–81.

Gooskens, C., & Swarte, F. (2017). Linguistic and extra-linguistic predictors of mutual intelligibility between germanic languages. *Nordic Journal of Linguistics*, *40*, 123–147.

Gooskens, C., & van Bezooijen, R. (2013). Lexical and orthographic distances between germanic, romance and slavic languages and their relationship to geographic distance (wilbert heeringa, jelena golubovic, charlotte gooskens, anja schuppert, femke swarte & stefanie voigt). https://api.semanticscholar.org/CorpusID:¨ 6289144

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second meeting of the north american chapter of the association for computational linguistics*.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Moller, R., & Zeevaert, L. (2015). Investigating word recognition in intercomprehension: Methods and findings.¨ *Linguistics*, *53*.

Vanhove, J., & Berthele, R. (2015). Item-related determinants of cognate guessing in multilinguals. *Crosslinguistic Influence and Crosslinguistic Interaction in Multilingual Language Learning*, *95*, 118.

Wichmann, S., Holman, E., Bakker, D., & Brown, C. (2010). Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, *389*, 3632–3639. https://doi.org/10.1016/j.physa.2010.05.011

Zmitrovich, D., Abramov, A., Kalmykov, A., Tikhonova, M., Taktasheva, E., Astafurov, D., Baushenko, M., Snegirev, A., Shavrina, T., Markov, S., Mikhailov, V., & Fenogenova, A. (2023). A family of pretrained transformer language models for russian.

Getting around

The conference takes place in building A2.1 (Innovation Center) on the main campus of Saarland University. See the back cover of this booklet for a campus map. The university campus is located about 5 kilometers outside the city center. The bus ride from the center ("Rathaus" or "Johanneskirche" stops) takes about 12 minutes. The closest bus stop to the venue is "Universität Campus". Note that only a reduced number of buses operate on Saturdays. For up-to-date information on bus routes, we recommend checking either the website `https://saarfahrplan.de/` or the free Saarfahrplan app.

Bus service from the city center to campus (bus stop: Universität Campus)

Line	Destination	Service days
101	Dudweiler Dudoplatz	Thu, Fri, Sat
102	Dudweiler Dudoplatz	Thu, Fri, Sat
109	Universität Busterminal	Thu, Fri
111	Universität Busterminal	Thu, Fri

Bus service from campus to the city center (bus stop: Johanneskirche)

Line	Destination	Service days
101	Füllengarten Siedlung	Thu, Fri, Sat
102	Altenkessel Talstraße	Thu, Fri, Sat
109	Goldene Bremm	Thu, Fri
111	Rabbiner-Rülf-Platz	Thu, Fri

Bus service from campus to the train station (bus stop: Hauptbahnhof)

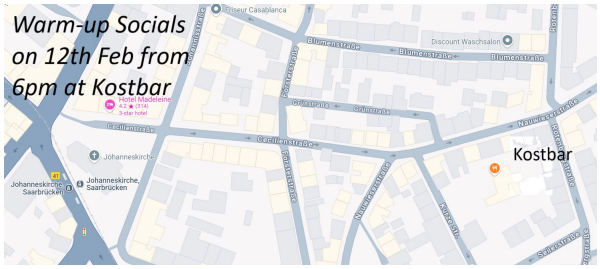
Line	Destination	Service days
102	Altenkessel Talstraße	Thu, Fri, Sat
112	Hauptbahnhof	Thu, Fri
124	Betriebshof	Thu, Fri

The University campus is serviced by taxis and buses alike. Should you need a taxi, you can contact one of the following companies:

Taxi Schneider	+49 (0) 681 71111
Taxi Zentrale e.G.	+49 (0) 681 55000
Taxi Saarbrücken e.G.	+49 (0) 681 33033

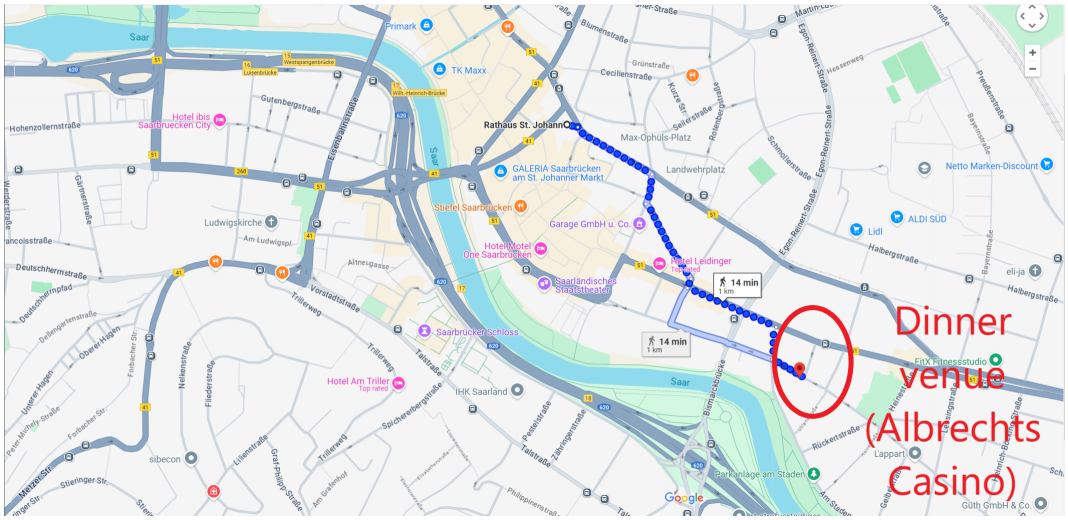
Pre-conference socials (warm-up)

The pre-conference event will take place at Restaurant Cafe Kostbar, which is located at Nauwieserstr. 19, Innenhof (inner courtyard). The venue is within walking distance from Rathaus in the city center.



Conference dinner

The conference dinner will take place at the restaurant “Albrechts Casino”, which is located on Bismarckstraße 47. The venue is within walking distance from Rathaus in the city center (about 15 minutes walk). See page 128 and the Saarfahrplan app for bus connections from campus.



Internet and WiFi

Guests will be able to access the Internet through the wireless network. There are two ways to connect to the network:

- Academics: Guests from academic institutions can use the Eduroam network with their institution’s credentials. No extra configuration is required.
- Industry and other guests: We have provided individual guest accounts for the university Wifi HIZ-GUEST for the entire duration of the conference. Please contact the registration desk to receive your ID and Password for Internet access. By connecting

to a wireless network of the University, you agree to the Terms of Use of the Hochschul-IT-Zentrum (HIZ) of Saarland University[1], along with the terms of the National Telecommunications Act.

[1] http://hiz-saarland.de

How to get to the campus from the city center?

- Bus stop at the university: ***"Universität Campus"***
- Bus stop (near the train station) ***"Saarbrücken Hbf"***
  Buses that run: **124, 112, 102**
- Bus stop ***"Rathaus"***
  Buses that run: **101,102, 111, 150**