Chapter 3

An information-theoretic account of constituent order in the German middle field

Katrin Ortmann[a], Sophia Voigtmann[b], Stefanie Dipper[a] & Augustin Speyer[b]

[a]Ruhr-Universität Bochum [b]Universität des Saarlandes

This paper proposes a novel approach to explain object order in German. Although the order of constituents is relatively free in modern German, there are clear preferences for the order dative before accusative (nominal) objects and for the order given before new objects. A range of influential factors have been described in the literature, most prominently givenness and length. We assume processing-related reasons and use information-theoretic measures, in particular surprisal and DORM (Cuskley et al. 2021), to explore the interplay of information structure and information density as factors for object order. We propose a measure called $\text{DORM}_{\text{diff}}$ and the *corpus of variants* method for comparing information profiles between different plausible constituent orders. Our investigations show that language users follow information-theoretic principles (UID, Levy & Jaeger 2007) in choosing the object order that leads to a more uniform distribution of information. We argue that this preference also explains deviations from the unmarked object order (i.e., accusative preceding dative and new preceding given) if it is associated with smoother information profiles.

1 Introduction

In contrast to languages with a fixed word order, the order of constituents in a language like German is relatively free. Nevertheless, there still exist clear preferences for certain word and constituent orders in German. One such preference

Katrin Ortmann, Sophia Voigtmann, Stefanie Dipper & Augustin Speyer. 2024. An information-theoretic account of constituent order in the German middle field. In Robin Lemke, Lisa Schäfer & Ingo Reich (eds.), *Information structure and information theory*, 55–86. Berlin: Language Science Press. DOI: 10.5281/zenodo.13383787

concerns the relative order of nominal dative and accusative object. For example, sentence (1a) is generally preferred over sentence (1b), even though both constituent orders are possible and occur in natural data.

(1) a. *Ich werde* [*einem Jungen*]$_{\text{DAT}}$ [*ein Buch*]$_{\text{ACC}}$ *geben.*
I will [a boy]$_{\text{DAT}}$ [a book]$_{\text{ACC}}$ give
'I will give a boy a book.'

b. *Ich werde* [*ein Buch*]$_{\text{ACC}}$ [*einem Jungen*]$_{\text{DAT}}$ *geben.*
I will [a book]$_{\text{ACC}}$ [a boy]$_{\text{DAT}}$ give
'I will give a book to a boy.'

There are numerous works on this phenomenon which try to capture the observed preferences. Among the known influential factors are animacy, familiarity, givenness, salience and length (cf., e.g., Lenerz 1977, Speyer 2011, Behagel 1932, and for English, Bresnan 2007). However, these factors cannot *explain* the preferences but only describe them. In this paper, we try to go beyond a mere description and attempt to explain this phenomenon based on the cognitive processing effort of the constructions (cf., e.g., Fenk-Oczlon 1983).

Some of the factors mentioned above certainly have an influence on processing effort, e.g., givenness as illustrated in (2). These sentences all have the marked case order accusative before dative, but differ with respect to givenness. Regarding givenness, the order given before new represents the common order (Section 5), so (2a) should be easier to process than the other examples since it is the most common object order, and familiarity can facilitate processing (cf., e.g. Futrell et al. 2021). However, such factors, and in fact all of the factors mentioned above except length, are difficult to quantify and thus hard to operationalize.

(2) a. *Ich werde* [*das Buch*]$_{\text{ACC, given}}$ [*einem Jungen*]$_{\text{DAT, new}}$ *geben.*
I will [the book]$_{\text{ACC, given}}$ [a boy]$_{\text{DAT, new}}$ give
'I will give the book to a boy.'

b. *Ich werde* [*ein Buch*]$_{\text{ACC, new}}$ [*dem Jungen*]$_{\text{DAT, given}}$ *geben.*
I will [a book]$_{\text{ACC, new}}$ [the boy]$_{\text{DAT, given}}$ give
'I will give a book to the boy.'

c. *Ich werde* [*das Buch*]$_{\text{ACC, given}}$ [*dem Jungen*]$_{\text{DAT, given}}$ *geben.*
I will [the book]$_{\text{ACC, given}}$ [the boy]$_{\text{DAT, given}}$ give
'I will give the book to the boy.'

In the present study, we explore the application of information-theoretic concepts to objectively quantify and approximate the effects of processing effort on object order in the middle field of the German sentence. We expect that a certain constituent order is used to assure an optimal information flow and to avoid processing difficulties. As a measure of processing difficulties, we use information density (Shannon 1948). In this framework, information is derived from the probability of a word in context. Information theory has been widely used to relate the probability of linguistic material occurring in an utterance (measured as surprisal: $S(\mathit{unit}) = -\log_2 P(\mathit{unit}|\mathit{context})$, Hale (2001)) to the effort required to process that utterance. Lower predictability (probability) correlates with higher processing effort (e.g., Hale 2001). Also, very high surprisal values or an uneven information profile are correlated with information loss, as (Cuskley et al. 2021) argue. Therefore, speakers aim to keep the information flow as uniform as possible to ensure optimal communication ("Uniform Information Density Hypothesis", UID, Levy & Jaeger 2007, Aylett & Turk 2004).

Since the predictability of a word depends strongly on its context, the order of words and constituents has a high impact on the uniformity of the utterance (Cuskley et al. 2021). Changing the order can thus lead to more successful communication and, based on this assumption, we propose that changes in object order in the German middle field can be described and even explained by information density. We test our hypothesis in a pilot study based on a large corpus of modern German.

The remainder of this paper is structured as follows: Section 2 gives an introduction of the theoretic background and explains the different factors that are known to influence constituent order in the German middle field. Section 3 describes the data selection for this study, and Section 4 details the methods used for analysis, including the calculation of constituent surprisal and information profiles. In Section 5, the results are presented and the effects of information-theoretic principles on constituent order are evaluated. Possible problems and enhancements of the methodology are discussed in Section 6. The paper concludes with a summary of the findings in Section 7.[1]

2 Constituent order in the German middle field

As already mentioned, German is a language with a relatively free constituent order. This means that constituent order is not exclusively governed by structural factors such as grammatical function (subject, direct object, etc.) as is the

[1] The statistical data and the R script used in this study as well as the list of light verb constructions applied in data preparation are available at https://gitlab.ruhr-uni-bochum.de/comphist/c6dormdiff.

case, e.g., in English. Instead, constituent order in German is influenced by several factors, many of which are non-syntactic factors but rather of a semantic or pragmatic nature (see, e.g., Lenerz 1977, Rauth 2020). This goes for historical stages of German as well (Speyer 2011, 2013, Rauth 2020).

The point of interest for our study is the so-called middle field in the German clause. The term *middle field* has its origin in the topological field model of the German clause (for a recent overview, see, e.g., Wöllstein 2010, 2014). We introduce the model using the terminology of Telljohann et al. (2017).

Word order in German sentences is best described not by notions such as SVO (subject > verb > object[2]) or the like, but rather by relating the constituents relative to the verb positions. Verb forms tend to be distributed over the German (matrix) clause in such a way that the finite part stands relatively early in the clause (linke (Satz-)Klammer (‘left sentence bracket’), abbreviated LK) and the remainder of the verb form at the end or close to the end of the clause, in a position often referred to as the *right sentence bracket* (’rechte Satzklammer’). In the scheme of Telljohann et al. (2017), this position is called *VC* (for verb complex). The positions of the nonverbal constituents of the clause can be described relative to these verbal positions. Nonverbal constituents can be located:

- either before the LK, i.e., in the *Vorfeld* (VF, ‘initial field’); this position is normally restricted to one constituent;

- or after the VC position, i.e., in the *Nachfeld* (NF, ‘final field’); this position is often not filled;

- or between the two brackets LK and VC, i.e., in the *Mittelfeld* (MF, ‘middle field’); it is this field that is in the focus of this paper.

A sample German declarative main clause with its topological structure is given in Table 1.

The middle field is the relevant area for our investigations because most constituents of the clause cluster in this field. For example, the example given in Table 1 shows four basic constituents: the subject *Uller*, the temporal adverbial *heute*, the indirect object *einem Freund* (in German usually in the dative case) and the direct object *ein Buch* (in German usually in the accusative case). Three of these constituents are located in the middle field.

[2] We use the notation *a* > *b* for denoting the order *a before b*.

Table 1: Example for the topological structure of a German declarative main clause

VF	LK	MF					VC	NF
Heute	hat	Uller	einem	Freund	ein	Buch	empfohlen	
today	has	Uller	a	friend	a	book	recommended	
‘Today, Uller recommended a book to a friend.’								

As already mentioned, the relative order of the constituents in the middle field is subject to different syntactic, semantic and pragmatic factors. In short, syntactic factors, such as grammatical function (subject > objects) or case (dative object > accusative object, in the following DAT > ACC) and the like are at play, but they can be easily overridden by non-syntactic factors (cf. the seminal study by Lenerz 1977). In this paper, we focus on the relative order of nominal objects in the German middle field. The unmarked order is DAT > ACC (Lenerz 1977).[3]

Semantic factors that have proven to be quite prominent are definiteness and animacy. The effect of definiteness is such that definite referents tend to precede indefinite referents (Lenerz 1977). It is questionable whether this definite > indefinite constraint is an effect of definiteness by itself or whether this is an epiphenomenon of other constraints. We will touch on this question later in this section.

Animacy has been identified as an important factor for the ordering of constituents in the German middle field by, e.g., Hoberg (1981). Here, the unmarked order is animated referent > unanimated referent, see (3). In the prehistory of German, this ordering principle might have been quite prominent and in the end might have led to the development of DAT > ACC as the unmarked order (see Speyer 2015) because the dative is correlated with the semantic role of recipient in the classical case of verbs with three arguments that instantiate the agent–patient–recipient scheme, such as *geben* (‘give’), *übermitteln* (‘convey’), or *anbieten* (‘offer’). The recipient is usually animated whereas the patient is normally not.

(3) *Heute hat* [*die Lehrerin*]$_{\text{NOM, anim}}$ [*der Schülerin*]$_{\text{DAT, anim}}$ [*das*
today has the teacher the student the
Buch]$_{\text{ACC, inanim}}$ *gegeben.*
book given
‘The teacher gave the book to the student today.’

[3] Interestingly, the unmarked order of *pronominal* objects is ACC > DAT. In this study, we focus on nominal objects, excluding pronominal objects.

We concentrate here on pragmatic factors, especially those that have traditionally been described in terms of *information structure* (Féry & Krifka 2008). Information-structural notions that have been found to play a role are, for example, the given > new constraint (Lenerz 1977) and the topic > comment constraint (Frey 2004). In our investigation, we focus on the given > new constraint. Basically, this ordering constraint says that knowledge that is assumedly familiar to the hearer is positioned before material that is new to the hearer. These constraints are not to be read as *given information always stands before new information* but rather as constraints that can override the unmarked constituent order DAT > ACC in certain cases, as in (4). In this example, the accusative object represents given information, whereas the dative object refers to a person that has not yet been introduced to the discourse.

(4) [Context: discussion about a certain mystery novel]
Und dann hat sie [*den Krimi*]$_{\text{ACC, given}}$ [*einer Freundin*]$_{\text{DAT, new}}$ *geschenkt.*
and then has she the novel a friend presented
'And then she gave the novel to a friend of hers as a present.'

We see in (4) that the objects bear different articles. A constraint that is correlated with given > new is the constraint that definite noun phrases precede indefinite noun phrases (Lenerz 1977, Rauth 2020). The correlation is as follows: Definite reference normally implies that the entity referred to is known to the speaker and hearer (hence given information). Using a definite determiner is felicitous only if the hearer can uniquely identify the referent, and this is only possible if it is known to the hearer or can be inferred by them (Prince 1981). In contrast, in conveying new information, speakers tend to refer via indefinite noun phrases, indicating that the referent is not yet part of the discourse universe. This comes in handy, as it allows us to use definiteness as a proxy for givenness and indefiniteness as a proxy for newness in our pilot study, when dealing with data that is not annotated for givenness or *information status.*

German is not the only language that allows for variable orders of the direct and indirect objects. In other closely related languages such as Dutch and English, the relative linearization of the direct object (DO) and the indirect object (IO) are subject to variation as well. An example is the phenomenon of dative alternation in English: The indirect object can be realized as a noun phrase preceding the direct object (5a), or as a prepositional phrase following the direct object (5b). The phenomenon of Heavy NP shift provides another example: long (i.e., heavy) direct objects can be put after the prepositional indirect object (5c).

(5) a. *Then she gave* [*her friend*]$_{\text{IO, NP}}$ [*the new mystery novel*]$_{\text{DO, NP}}$.

b. *Then she gave* [*the new mystery novel*]$_{\text{DO, NP}}$ [*to her friend*]$_{\text{IO, PP}}$.

c. *Then she gave* [*to her friend*]$_{\text{IO, PP}}$ [*the new mystery novel about the murderer from Dartmoor*]$_{\text{DO, NP}}$.

The factors governing these variations are partly of a different nature. While the length of the respective objects seems to be a governing factor, given-/newness does not seem to play a primary role here. Engel et al. (2022) found evidence that definiteness is a good predictor also for the English dative alternation (if the indirect object is indefinite, it is more often realized as prepositional phrase, but this effect is strongest in spoken informal texts). So it looks as though something similar to the German definite > indefinite constraint is at play in English as well, and the fact that the effect is strongest in orally produced texts indicates that it is a matter of constraints on language processing.

In our investigations, we focus on sentences with ditransitive verbs whose objects are located in the middle field. In our study, we compare the two objects in their original order with a generated, reversed order (see Section 4). In this direct comparison, we want to investigate whether the role of givenness for word order can be quantified with the help of information-theoretic measures such as surprisal. Hence, as described in Section 3, we exclude all cases where the objects are either both definite or both indefinite (i.e., where givenness does not play a role) and keep the mixed cases only so that the two variants differ with regard to definiteness, our proxy for givenness. Moreover, other factors that could influence the order of constituents should be excluded when comparing the two variants. Hence, we control for object length because variations in length are known to have an impact on the order of constituents in the sentence ("Gesetz der wachsenden Glieder", or *law of increasing constituents*, Behagel 1932).

3 Data

We use the SdeWaC corpus (Faaß & Eckart 2013)[4] as the source of data for our analysis. The corpus consists of 44M sentences with more than 845M tokens from German webpages. It has been automatically tokenized, tagged, lemmatized, and parsed with Bohnet (2010)'s dependency parser.[5] Using the dependency annotation, we select all sentences from the corpus that contain at least one ditransitive

[4] https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/sdewac, accessed 2022/12/01.

[5] Bohnet (2010)'s dependency parser was trained on the TIGER corpus (Brants et al. 2004: release August 2007) which had been converted to dependency structures by Wolfgang Seeker.

verb with a dative and an accusative object, labeled DA (= DAT) and OA (= ACC), respectively, in the dependency annotation.[6] In addition, the objects must meet the following criteria:

(i) Both objects have a nominal head. This means that the word forms labeled with the dependency relation OA and DA must be tagged with the STTS tag NN for "normal noun" (Schiller et al. 1999). For example, in (6), the accusative object *ein Buch* ('a book') and the dative object *dem Jungen* ('the boy') in (6a) are recognized as having a nominal head. In contrast, the pronominal dative object *ihm* ('him') in (6b) is tagged as PPER for personal pronoun and the sentence would be excluded from the sample.

(6) a. *Ich werde* [*dem Jungen*/NN/DA]$_{\text{DAT}}$ [*ein Buch*/NN/OA]$_{\text{ACC}}$ *geben.*
I will the boy a book give
'I will give a book to the boy.'

b. *Ich werde* [*ihm*/PPER/DA]$_{\text{DAT}}$ [*ein Buch*/NN/OA]$_{\text{ACC}}$ *geben.*
I will him a book give
'I will give him a book.'

(ii) To draw conclusions about the givenness of the objects, the object noun phrases must differ with regard to definiteness, one being definite, the other being indefinite. That is, the head nouns of one of the objects must directly dominate a definite article (def) and the head noun of the other object must directly dominate an indefinite article (indef). Definite articles are word forms that are tagged with the STTS tag ART and are lemmatized as *der* ('the'). Indefinite articles are word forms tagged as ART with the lemma *ein* ('a'). Examples (2a) and (2b) from the introduction would thus be included, while (1a), (1b), and (2c) with two given or two new objects would be excluded. This criterion also entails that sentences with an indefinite plural object, like *Bücher* ('books') in (7), are rejected because they do not have a determiner in German.

(7) *Ich werde* [*dem Jungen*/DA]$_{\text{DAT}}$ [*Bücher*/OA]$_{\text{ACC}}$ *geben.*
I will the boy books give
'I will give books to the boy.'

[6] The label DA is also used for free datives, see Brants et al. (2004). However, free datives occur mainly in pronominal form, which are excluded from the present study.

(iii) To control for effects of length, the objects must contain the same number of words (ignoring punctuation). Example (8a) with two objects of length two would be accepted, but not (8b) with objects of different lengths (two vs. three words).

(8) a. *Ich werde* [*dem Jungen*]$_{\text{DAT}}$ [*ein Buch*]$_{\text{ACC}}$ *geben.*
I will [the boy]$_{\text{DAT}}$ [a book]$_{\text{ACC}}$ give
'I will give a book to the boy.'

b. *Ich werde* [*dem Jungen*]$_{\text{DAT}}$ [*ein gutes Buch*]$_{\text{ACC}}$ *geben.*
I will [the boy]$_{\text{DAT}}$ [a good book]$_{\text{ACC}}$ give
'I will give a good book to the boy.'

(iv) Both objects must be located within the same middle field (`MF`).[7] We only keep sentences in which the same `MF` node dominates both objects, as in (9a). If one object is located in another field, for example, in another `MF` or in the initial field `VF` as in (9b), the sentence is excluded.

(9) a. *Ich werde* ⟦[*das Buch*]$_{\text{ACC}}$ [*einem Jungen*]$_{\text{DAT}}$⟧$_{\text{MF}}$ *geben.*
I will the book a boy give
'I will give the book to a boy.'

b. ⟦[*Das Buch*]$_{\text{ACC}}$⟧$_{\text{VF}}$ *werde* ⟦*ich* [*einem Jungen*]$_{\text{DAT}}$⟧$_{\text{MF}}$ *geben.*
the book will I a boy give
'I will give the book to a boy.'

(v) Finally, we exclude light verb constructions, in which a semantically faded ("light") verb establishes one fused meaning with its object. For instance, the phrase *einer Prüfung unterziehen* ('submit a check') in (10) is an example of such a construction: (*to) submit a check* corresponds to *(to) check*. In these constructions, there is a clear bias for the order in which the fused object is directly adjacent to the light verb. This even holds for cases where the fused object is the dative object, resulting in the fixed (otherwise marked) object order ACC > DAT, as in (10).

[7] For determining the topological structure, we parse the sentences with the Berkeley parser (Petrov et al. 2006) and a constituency model from Ortmann (2021) trained on the TüBa-D/Z treebank, a corpus that has been annotated with syntactic and topological categories (Telljohann et al. 2017). We use the `News1` model from https://github.com/rubcompling/konvens2021, which was trained on 80% of the TüBa-D/Z corpus. The model annotates constituents and topological fields at the same time.

(10) *Wir werden* [*die neuen Daten*]$_{\text{ACC}}$ [*einer genauen Prüfung*]$_{\text{DAT}}$
we will the new data a thorough check
unterziehen.
give
'We will submit the new data to a thorough check.'

We compiled a list of 120 light verb constructions from Eisenberg (2020) and ProGram2.0 (2018).[8] If the lemmas of the verb and of the head nouns of the objects are included in the list, the object pair is removed.

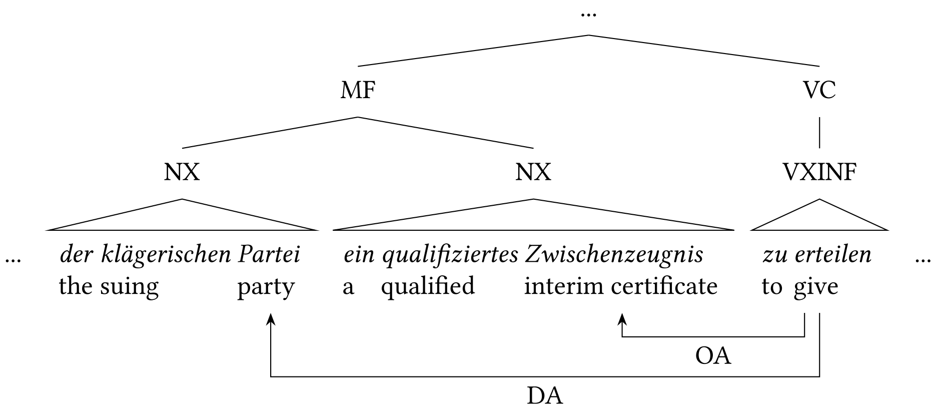
...
MF
VC
NX
NX
VXINF
... der klägerischen Partei ein qualifiziertes Zwischenzeugnis zu erteilen ...
the suing party a qualified interim certificate to give
OA
DA

Figure 1: Excerpt from an example sentence (engl. 'to give the suing party a qualified interim certificate') with a ditransitive verb and its two objects, along with a constituency (top) and dependency (bottom) analysis

Figure 1 shows an example object pair with the corresponding dependency and constituency analysis. On top of the text, the constituency tree is displayed, consisting of noun phrases (labeled as `NX`, following the TüBa-D/Z annotation scheme, Telljohann et al. 2017), an infinitive (`VXINF`), and nodes representing topological fields (`MF`, `VC`). Below the text, the relevant dependency relations are shown. As required, the verb dominates a nominal dative (`DA`) and accusative (`OA`) object pair within the same middle field (`MF`) and with the same number of words. The dative object has a definite article (*der* ('the')) and the accusative object an indefinite one (*ein* ('a')).

[8] The list is available at https://gitlab.ruhr-uni-bochum.de/comphist/c6dormdiff.

For our analysis, the selected sentences are split into constituents based on their constituency parse. For each terminal token (ignoring punctuation), we choose as the constituent node the highest dominating phrasal node below the next topological field node. (11) shows an example constituency analysis from the data set.

(11) [*Sie*]$_{\text{NX}}$ [*sind*]$_{\text{VXFIN}}$ [*zudem*]$_{\text{PX}}$ [*ein wichtiges Stilmittel*]$_{\text{NX}}$, [*um*]$_{\text{C}}$ [*dem*
they are moreover an important stylistic.device to the
Film]$_{\text{NX}}$ [*eine Struktur*]$_{\text{NX}}$ [*zu verleihen*]$_{\text{VXINF}}$
film a structure to give
'Moreover, they are an important stylistic device to give the film a structure.'

The SdeWaC corpus contains approximately 1.8M ditransitive verbs. Among those, 13,472 object pairs in 13,458 sentences meet the aforementioned criteria. Table 2 gives a summary of the data. It shows that in 95.87% of the cases, the dative object precedes the accusative object and 87.61% of the definite objects precede an indefinite object. Only 5.32% of the objects in the original data are longer than three words, so we decided to only include objects of length two and three in our final data set.[9]

The above constraints concerning case and definiteness result in a total of four possible combinations of object pairs:

(i) DAT.DEF > ACC.INDEF (i.e., the definitive dative object precedes the indefinite accusative object)

(ii) DAT.INDEF > ACC.DEF

(iii) ACC.DEF > DAT.INDEF

(iv) ACC.INDEF > DAT.DEF

Examples (12–15) show one sentence per group from the sample.

[9] This decision was also made because data processing proved to be error-prone for objects with more than three words. This could be solved by filtering as described above.

Table 2: Summary of the selected sentences and object pairs from the SdeWaC corpus, for the original complete data and the final data set with objects of length two and three only

	Original data		Final data set	
	n	%	*n*	%
Sentences	13,458		12,742	
Object pairs	13,472		12,756	
Sentences with >1 pair	14	0.10	14	0.11
Dative before accusative (DAT>ACC)	12,916	95.87	12,253	96.06
Definite before indefinite (def>indef)	11,803	87.61	11,171	87.57
(i) DAT.DEF>ACC.INDEF	11,601	86.11	10,999	86.23
(ii) DAT.INDEF>ACC.DEF	1,315	9.76	1,254	9.83
(iii) ACC.DEF>DAT.INDEF	354	2.63	331	2.59
(iv) ACC.INDEF>DAT.DEF	202	1.50	172	1.35
Min. object length (in words)	2		2	
Max. object length (in words)	13		3	
Avg. words per object	2.35		2.22	
Avg. constituents per sentence	12.23		12.28	

(12) Group (i): DAT.DEF > ACC.INDEF
Beim Zeichnen des eigenen Gesichts kann man [*dem Schüler*]$_{\text{DAT, def}}$
when drawing the own face can one the student
[*einen Spiegel*]$_{\text{ACC, indef}}$ *geben, aber man kann die Unterrichtseinheit auch*
a mirror give but one can the lesson also
mit der Fotografie beginnen.
with the photography start
‘When drawing your own face, you can give the student a mirror, but you can also start the lesson with photography.’

(13) Group (ii): DAT.INDEF > ACC.DEF
Ich fühle mich jetzt viel sicherer, schlafe nachts ruhig, weil
I feel myself now much safer sleep at.night peacefully because
ich mir keine Sorgen darüber machen muß, wie ich [*einem*
I me no worries about make must how I a

Geldverleiher]$_{\text{DAT, indef}}$ [*das Geld*]$_{\text{ACC, def}}$ *zurückzahlen soll.*
money.lender the money pay.back shall
'I feel much safer now, sleep peacefully at night because I don't have to worry about paying back a money lender.'

(14) Group (iii): acc.def > dat.indef
Ein paar Tage später zeigte ich [*den Film*]$_{\text{ACC, def}}$ [*einem Freund*]$_{\text{DAT, indef}}$
a few days later showed I the film a friend
und sah ihn noch einmal mit der gleichen Begeisterung.
and watched it once more with the same enthusiasm
'A few days later, I showed the film to a friend and watched it again with the same enthusiasm.'

(15) Group (iv): acc.indef > dat.def
Wegen der geänderten Zuständigkeiten im Grundgesetz
because.of the changed responsibilities in.the constitution
müsse der Bund [*eine Neukonzeption*]$_{\text{ACC, indef}}$ [*den*
would.have.to the federal.government a redesign the
Ländern]$_{\text{DAT, def}}$ *überlassen.*
states leave
'Because of the changed responsibilities in the constitution, the federal government would have to leave a redesign to the states.'

The vast majority follows the unmarked order of definite dative before indefinite accusative (group (i)), cf. Figure 2.[10] The example in Figure 1 is also an instance of the unmarked order dat.def > acc.indef.

4 Methods

We propose information density and, more specifically, the uniform distribution of information in the sentence as an explanation of object order. In the information-theoretic framework, information can be derived from the predictability of a word in context (Shannon 1948), with lower predictability causing higher processing effort (Hale 2001, Levy 2008).

We use language models to estimate the probability $p(w)$ of individual tokens w from bigram lemma frequencies in the SdeWaC corpus. To keep the

[10] The plots have been created with the R package ggplot2, https://github.com/tidyverse/ggplot2.

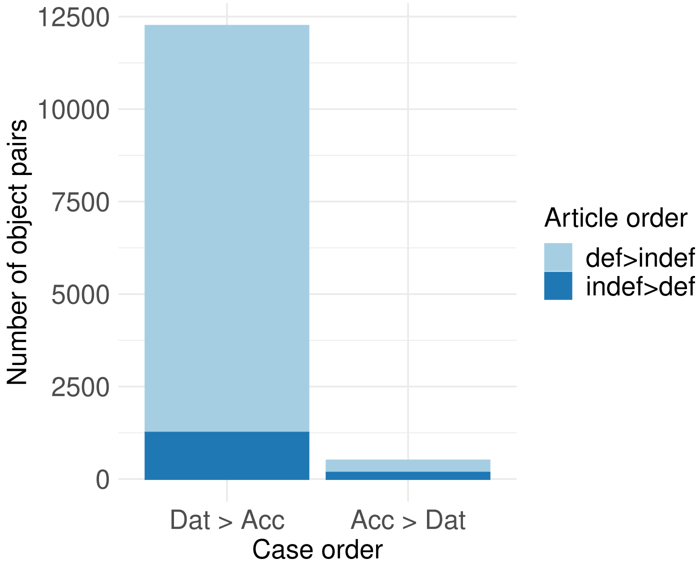

Figure 2: Frequencies of case and article order in the final data set. The majority of object pairs follow the unmarked order of definite dative before indefinite accusative (group (i); upper part of the left bar).

data size manageable, we include only bigrams with ≥50 occurrences and apply Jeffreys-Perks smoothing with $\lambda = 0.5$ (Jeffreys 1946), yielding a total amount of approximately 1M bigrams with 100K distinct lemma types. Punctuation is ignored as we assume that it does not provide any additional information about processing efforts in the German middle field.

As we are interested in the order of constituents, we measure predictability not at the word level but at the level of whole constituents. We calculate the mean surprisal $Surpr_{\text{mean}}$ of a constituent $c = w_1, \dots, w_n$ by adding up the individual surprisal values of all the words in the constituent and averaging them, see equation (16).

(16)
$$\text{Surpr}_{\text{mean}}(c) = \frac{1}{n} \sum_{i=1}^{n} -\log_2(p(w_i))$$

The information profile of a sentence, which indicates whether information is distributed uniformly and smoothly across the sentence, is composed of the surprisal values of all the constituents in the sentence, which are simply concatenated. Figure 3 shows an example: The fragment marked as *original* consists of the constituents [*um*] ('in order'), [*dem Film*] ('the film'), [*eine Struktur*] ('a structure'), [*zu verleihen*] ('to give') (see example (11) for the complete sentence).[11] The corresponding information profile is displayed in the second row ($\text{Surpr}_{\text{mean}}(\text{c})$): For instance, the lemma-based mean bigram surprisal of the dative object (*dem Film*) is 5.921 bits, and the surprisal of the accusative object (*eine Struktur*) is 9.879 bits. The resulting profile of this fragment is the sequence [16.701, 5.921, 9.879, 8.348].

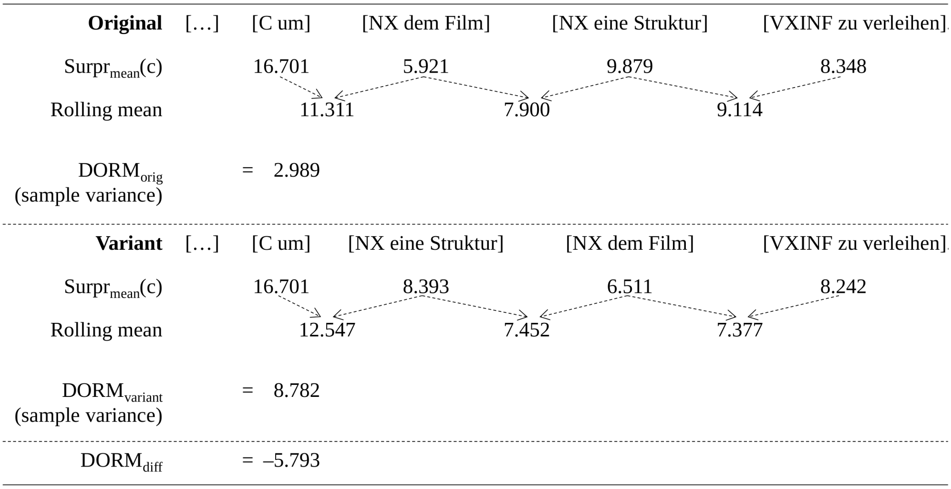
Original […] [C um] [NX dem Film] [NX eine Struktur] [VXINF zu verleihen]
Surpr_mean(c) 16.701 5.921 9.879 8.348
Rolling mean 11.311 7.900 9.114
DORM_orig (sample variance) = 2.989
Variant […] [C um] [NX eine Struktur] [NX dem Film] [VXINF zu verleihen]
Surpr_mean(c) 16.701 8.393 6.511 8.242
Rolling mean 12.547 7.452 7.377
DORM_variant (sample variance) = 8.782
DORM_diff = –5.793

Figure 3: Example calculation of rolling means and DORM values for a part of sentence (11)

We then compare this information profile with the profile of a competing variant, i.e., a generated alternative sentence that looks like the original sentence, except that the two objects are swapped. In Figure 3, the variant sentence with the two swapped objects is displayed below the original sentence. The upper part of Figure 3 shows the original constituent order, the variant is displayed in the lower part. Note how the surprisal values change because of the swapped objects. As the original order has a lower DORM value (i.e., a smoother profile) than the generated variant, $\text{DORM}_{\text{diff}}$ is negative for this fragment.

We call this approach the *corpus of variants* method because it allows us to inspect the differences between the observed word order and a plausible alternative order, while keeping other factors constant. The variant generation causes a change of bigram surprisals at the edges of the swapped objects, so we recalculate the surprisal values on the basis of the language model that was also

[11] One could argue that the phrase *eine Struktur verleihen* is a light verb construction because it can be replaced by *strukturieren* ('(to) structure'). However, it is not part of our list of light verb constructions (see Section 3) and is therefore not excluded from the data.

used for the original sentence and the information profile for the generated variant sentence, see Figure 3: The dative object now has a mean surprisal value of 6.511 bits and the accusative object a surprisal of 8.393 bits.

For comparing the information profiles of the original sentence and the generated sentence, we use measures called DORM and $\mathrm{DORM}_{\mathrm{diff}}$, as explained in the next sections.

4.1 DORM

DORM (Deviation of the Rolling Mean), which has been proposed by Cuskley et al. (2021), is a measure that allows us to quantify the uniformity of a sentence's information profile. Cuskley et al. (2021: 9) describe DORM as an "easily interpretable summary of how uniform or clumpy a particular utterance is". DORM is calculated as follows: Given the sequence of surprisal scores of all constituents in a sentence, we first compute the rolling means RM_i of each adjacent pair of surprisal scores s_i, s_{i+1} as in equation (17).

(17)
$$\text{for } i \text{ in } (1 \ldots n-1) : RM_i = \frac{s_i + s_{i+1}}{2}$$

For instance, the first mean RM_1 in Figure 3 (original sentence) is the mean of 16.701 (= [*um*]'s surprisal) and 5.921 (= [*dem Film*]'s surprisal):

(18)
$$\frac{(16.701 + 5.921)}{2} = 11.311$$

We next compute DORM, which corresponds to the sample variance of the rolling means and serves us as a measure of the overall smoothness, as shown in equation (19).

(19)
$$\mathrm{DORM} = s^2 = \frac{\sum_{i=1}^{n} (RM_i - \bar{x})^2}{n-1}$$

A lower DORM value indicates less variance, i.e., a smoother information signal, while a higher DORM value points at a less uniform information profile. This is usually achieved by placing linguistic units, in our case constituents, with similar surprisal values next to each other since extreme differences would no longer result in a low DORM value (Cuskley et al. 2021). Extreme surprisal values should, thus, be spread evenly across a sentence.

In Figure 3, the original sentence has a DORM value of 2.989, and the variant sentence has a DORM value of 8.782. This means that the original object order results in a smoother profile.

As we show in the next section, we use the DORM values for pairwise comparing information profiles of original sentences and their variants and introduce a new measure, $\text{DORM}_{\text{diff}}$, for measuring the difference between the original and the variant sentence.

4.2 $\text{DORM}_{\text{diff}}$

DORM values are directly comparable only for sequences that contain the same (number of) elements. Hence, the absolute DORM values can only be compared between the original constituent order ($\text{DORM}_{\text{orig}}$) and the swapped variant ($\text{DORM}_{\text{variant}}$) of the same sentence.

In order to compare values from different sentences, we use the difference between DORM value pairs, as defined in equation (20). That is, we collect the individual differences between all original and variant pairs of the sample and use these scores in our investigations.

(20) $$\text{DORM}_{\text{diff}} = \text{DORM}_{\text{orig}} - \text{DORM}_{\text{variant}}$$

$\text{DORM}_{\text{diff}}$ allows us to investigate the difference between the observed information profile and the profile of the variant constituent order. If there was no connection between object order and information profile, $\text{DORM}_{\text{diff}}$ should be zero. In contrast, if speakers aimed at a smooth information profile in accordance with the UID hypothesis (Levy & Jaeger 2007), DORM should be lower for original sentences than for the variants. If the information profile of the variant sentences was more uniform, there would have to be other explanations for the observed object order.

Our hypothesis is therefore that, in general, $\text{DORM}_{\text{diff}}$ should be negative (as in the example in Figure 3) – because this would mean that the original sentence has a smoother profile than its variant and, hence, that constituent order can be traced back to information-theoretic principles.

4.3 $\text{DORM}_{\text{case}}$ and DORM_{giv}: Case and givenness order

We use logistic regressions to investigate the effects of information profile, case, and givenness on object order. If any of these factors significantly influenced the order of dative and accusative or given and new object, they should help to predict which order will occur in the sentence.

However, we cannot simply use $\text{DORM}_{\text{diff}}$ as defined in equation (20) to predict case and givenness order because the order is encoded in the score. If the original sentence order is DAT > ACC, $\text{DORM}_{\text{diff}}$ is calculated as $\text{DORM}_{\text{DAT > ACC}}$ –

$\text{DORM}_{\text{ACC > DAT}}$. And if the original sentence order is ACC > DAT, $\text{DORM}_{\text{diff}}$ is calculated as $\text{DORM}_{\text{ACC > DAT}} - \text{DORM}_{\text{DAT > ACC}}$. The same applies analogously to def > indef.

Hence, $\text{DORM}_{\text{diff}}$ as a predicting factor must not be calculated with reference to *orig* and *variant*. Instead, it must abstract away from the actually occurring order and always use the same order of minuend and subtrahend, as shown in the equations (21) and (22).[12]

(21) $$\text{DORM}_{\text{case}} = \text{DORM}_{\text{DAT > ACC}} - \text{DORM}_{\text{ACC > DAT}}$$

(22) $$\text{DORM}_{\text{giv}} = \text{DORM}_{\text{DEF > INDEF}} - \text{DORM}_{\text{INDEF > DEF}}$$

Based on equations (21) and (22), we can predict if and how the order of the two objects will be influenced by a change in the uniformity of the information profile resulting from a change in case order ($\text{DORM}_{\text{case}}$) or givenness order ($\text{DORM}_{\text{giv}}$).

$\text{DORM}_{\text{case}}$ is smaller than zero if the order of DAT > ACC has a more uniform information profile than ACC > DAT, and greater than zero otherwise. Similarly, a negative DORM_{giv} indicates a more uniform information profile for DEF > INDEF, while a positive value shows a more uniform distribution for INDEF > DEF.

As DAT > ACC and DEF > INDEF are considered the unmarked constituent order (cf. Section 2), they can be expected to be easier to process for language users since they are more familiar with this conventionalized order. However, if the information profile for ACC > DAT or INDEF > DEF was smoother than for the default order, this could potentially lead to an inverse, marked order of objects to reduce processing difficulty. If this is true, a higher $\text{DORM}_{\text{case}}$ (i.e., a less optimal information profile for DAT > ACC) should increase the likelihood of ACC > DAT. And, along the same lines, a higher DORM_{giv} (i.e., a smoother information profile for INDEF > DEF) should increase the probability of INDEF > DEF.

5 Results

5.1 $\text{DORM}_{\text{diff}}$: Object order and the information profile

To explore the relevance of information-theoretic principles for object order in the German middle field, we inspect the information profiles of the original sen-

[12] Note that $\text{DORM}_{\text{case}} = -(\text{DORM}_{\text{ACC > DAT}} - \text{DORM}_{\text{DAT > ACC}})$, and, similarly, $\text{DORM}_{\text{giv}} = -(\text{DORM}_{\text{INDEF > DEF}} - \text{DORM}_{\text{DEF > INDEF}})$. Hence, as long as it is used consistently, the order of minuend and subtrahend is irrelevant, and we arbitrarily decided for the orders DAT > ACC and DEF > INDEF as the minuends.

tences and their generated variants. For the data described in Section 3 and their corresponding variants, $\text{DORM}_{\text{diff}}$ lies between −33.16 and 23.90, with slightly more than half of the values (52.7%) being smaller than zero. On average, the DORM value of the original constituent order is significantly lower than the DORM value of the generated variants: $\text{DORM}_{\text{diff}} = -0.17$ ($t = -6.88$, $p < 0.001$).[13] The effect size (Cohen's $d = 0.06$) is smaller than 0.2, which is traditionally assumed to indicate a small effect (Winter 2020), but the result suggests that natural language indeed follows information-theoretic principles, as writers tend to produce sentences with information profiles smoother than the ones that would result from an also plausible, but inverse object order.

Table 3: Mean $\text{DORM}_{\text{diff}}$ values for different object orders; (i)–(iv) refer to the four groups of possible combinations (*** $p < 0.001$; for the complete statistics, see Table 4).

	DEF>INDEF	INDEF>DEF	all
DAT>ACC	(i) −0.12 ***	(ii) −0.68 ***	−0.18 ***
ACC>DAT	(iii) 0.10	(iv) −0.23	−0.01
all	−0.12 ***	−0.63 ***	−0.17 ***

As Table 3 shows, this observation holds independently of the observed order in the original sentence of dative and accusative or definite and indefinite object.[14] Looking first at the right-most column (*all*), we see that for the unmarked order DAT > ACC (first row), which appears in the majority of sentences of the original data set (Section 3), the mean $\text{DORM}_{\text{diff}}$ is −0.18. For ACC > DAT, the mean $\text{DORM}_{\text{diff}}$ is also negative (−0.01) even though it is not significantly different from zero. Looking at the bottom row (*all*), we see that for the marked order INDEF > DEF, $\text{DORM}_{\text{orig}}$ is on average −0.63 lower than $\text{DORM}_{\text{variant}}$. For the unmarked order DEF > INDEF, the difference (−0.12) is negative, too, and also significantly different from zero.

Regarding the four possible combinations of case and givenness order (i.e., groups (i)–(v) in the inner part of Table 3), we see that three out of four groups show negative $\text{DORM}_{\text{diff}}$ values on average, the order ACC_{def} > $\text{DAT}_{\text{indef}}$ being an exception with a $\text{DORM}_{\text{diff}}$ of 0.10. Only the two groups with default case order DAT > ACC result in highly significant differences, both for the unmarked

[13] Statistical calculations have been performed with R (R Core Team 2018). We used two-tailed Welsh t-tests for these calculations.

[14] The complete statistics are presented in Table 4.

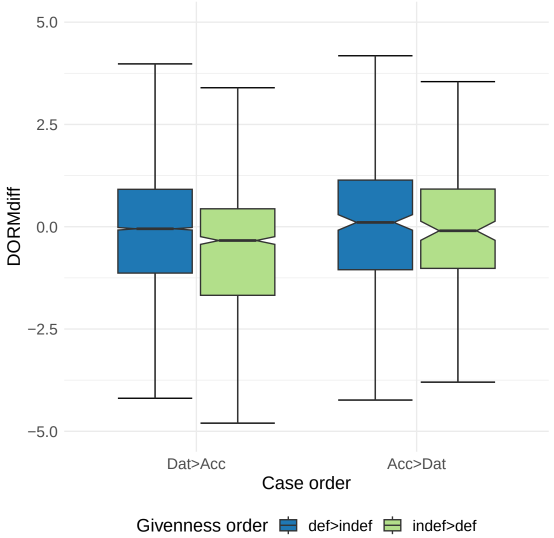

Figure 4: $\mathrm{DORM}_{\mathrm{diff}}$ by object order and givenness order (not displaying outliers). The boxes show the interquartile range from first to third quartile, with a black line for the median $\mathrm{DORM}_{\mathrm{diff}}$. The notches indicate the confidence intervals for the median. The four boxes correspond, from left to right, to the groups (i)–(iv), respectively.

givenness order, DEF > INDEF, with a mean of −0.12 as well as for the marked givenness order with a mean of −0.68. In the two cases where the original order is ACC > DAT, no significant differences are found between the $\mathrm{DORM}_{\mathrm{diff}}$ values. This can possibly be attributed to the small amount of data that is available in these groups (cf. Table 2).

Figure 4 shows additional details about the distribution of the four combinations of case and givenness. If t is negative, $\mathrm{DORM}_{\mathrm{orig}}$ is lower on average than $\mathrm{DORM}_{\mathrm{variant}}$, which indicates a more uniform information profile for the original sentence. Traditionally, values of $0.2 \leq d < 0.5$ are interpreted as a small effect. In three out of four conditions, the majority of values lie below zero. However, this difference is significant only in the left group (DAT > ACC) and, in particular, for the marked order INDEF > DEF (green box).

We can interpret the observed trends as follows: In many cases, the information profil of the original sentence and its variant are rather similar, which is shown by many values close to zero and the small effect sizes (see Table 4). However, if sentences show the default case order (DAT > ACC), this is associated with a more uniform information profile, which may explain the large preponderance of this order in modern German (cf. Section 3). At the same time, original sentences generally show a more uniform distribution of information than possible

Table 4: Results of two-sided one sample t-tests for DORM$_{\text{diff}}$

	DORM$_{\text{diff}}$	t	df	Cohen's d	p
all	−0.17	−6.88	12755	0.06	<0.001 ***
DAT > ACC	−0.18	−6.98	12252	0.06	<0.001 ***
ACC > DAT	−0.01	−0.11	502	0.00	0.91
DEF > INDEF	−0.12	−4.37	11329	0.04	<0.001 ***
INDEF > DEF	−0.63	−8.26	1425	0.22	<0.001 ***
(i) DAT.DEF > ACC.INDEF	−0.12	−4.55	10998	0.04	<0.001 ***
(ii) DAT.INDEF > ACC.DEF	−0.68	−8.20	1253	0.23	<0.001 ***
(iii) ACC.DEF > DAT.INDEF	0.10	0.63	330	0.03	0.53
(iv) ACC.INDEF > DAT.DEF	−0.23	−1.36	171	0.10	0.17

variant sentences — even if the realized order violates the unmarked order of case or givenness though the effect is only significant in the DAT > ACC order. So the preference of language users for smooth information profiles, as predicted by the UID hypothesis, may license deviations from the default case or givenness order.

5.2 DORM$_{\text{case}}$ and DORM$_{\text{giv}}$: Case and givenness order

To inspect possible effects of the information profile on the order of dative and accusative object and definite and indefinite object, we use logistic regression analyses in R (R Core Team 2023). We start with case order and run a logistic regression with DORM$_{\text{case}}$, givenness status, and the number of constituents in the sentence as well as all two-way-interactions as predictors.[15]

Case order is sum-coded: DAT > ACC received the coding 1 and ACC > DAT was sum-coded as −1. Thus, positive estimates in the main effects indicate the order DAT > ACC. As givenness status, we use the definiteness of the dative, which was also sum-coded to increase the precision of the model (Gries 2021).[16] A definite dative was coded as −1, an indefinite dative as +1. While we control for

[15] glm(formula = Dat > Acc ~(DORM$_{\text{case}}$ + Dat$_{\text{definiteness}}$ + n_Constituents)^2, family = binomial(), data = constituents_sample); for the complete final regression model, see Table 5. Furthermore, we include the two-way interactions of the three factors. Since DORM$_{\text{case}}$ and DORM$_{\text{giv}}$ are strongly correlated ($r = 0.73$), we choose to only include one of them as a predictor in each regression analysis.

[16] The objects always exhibit opposing definiteness (cf. Section 3). If the dative object is definite, the accusative object is indefinite, and vice versa. We arbitrarily selected the definiteness of the dative object as predictor. With the accusative as predictor, results would simply be reversed.

object length in that both objects consists of the same number of words, the number of constituents varies between sentences. It seems plausible that a long sentence with a high number of constituents is harder to process than a sentence with fewer constituents. When the amount of information in a sentence already threatens to strain the working memory, the default order DAT > ACC might be preferred to ease overall sentence processing. There might also be an interaction between the information profile of the sentence and its length. However, the order ACC > DAT only occurs once in sentences that have more than 40 constituents (cf. Figure 5). We, consequently, run the logistic regression on a sample of the whole data excluding sentences with more than 30 words.

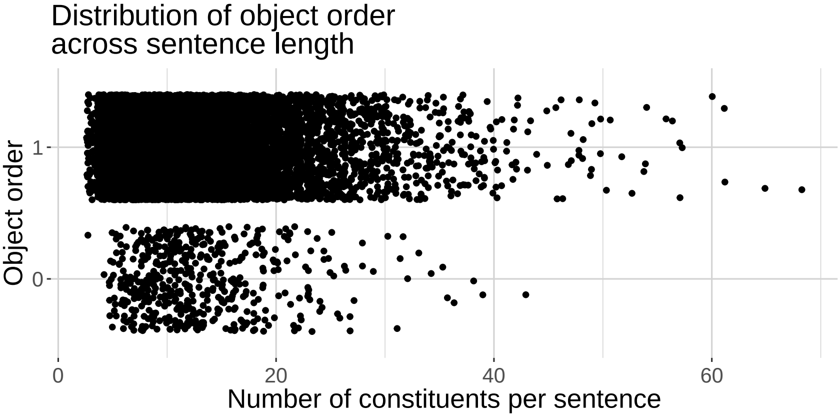

Figure 5: Distribution of object order in the sentences of various length, shown by the number of constituents

Then, we perform *backward model selection* (Gries 2021), excluding one interaction or one main effect at a time, depending on the *p*-value of the predictor. We start with the interactions and first exclude those with the highest non-significant *p*-value. To find out whether the exclusion led to an improvement of the model, a *likelihood ratio* test with the anova function in R (R Core Team 2018) is performed. It allows model comparison by capturing how well the model explains the data (Winter 2020). This process is repeated until only significant effects or main effects involved in a significant interaction remain in the model. As soon as the *likelihood ratio* test shows a significant difference between the models, the process of backward model selection is completed. The final model then corresponds to the model before the exclusion of the last predictor and is used to interpret the results.

5.2.1 Case order

Table 5: Logistic regression with DORM$_{case}$, definiteness of the dative object and the number of constituents in the sentence to predict case order dative > accusative

Variable	Estimate	SE	z	p	
Intercept	2.90	0.12	23.34	<0.001	***
DORM$_{case}$	−0.06	0.02	−3.58	<0.001	***
DAT$_{def}$	−1.83	0.12	−14.695	<0.001	***
Constituents	−0.014	0.009	−1.60	0.11	
DAT$_{def}$:Constituents	0.03	0.01	3.64	0.001	***

Table 5 shows the results of the regression analysis for case order. According to the model, DORM$_{case}$ ($z = -3.58$, $p < 0.001$) has a highly significant influence on case order.[17] A higher DORM$_{case}$ reduces the likelihood of observing DAT > ACC. An increase of DORM$_{case}$ means that the information profile of ACC > DAT is smoother than that of DAT > ACC. Hence, a more uniform, smoother distribution for the order ACC > DAT increases the likelihood of observing this marked order in the sentence. And vice versa, a more uniform distribution of DAT > ACC increases the likelihood of this default order.

The second predictor, the definiteness of the dative, also significantly influences case order ($z = -14.695$, $p < 0.001$). In accordance with information structure, an indefinite dative reduces the likelihood of observing the order DAT > ACC. If the dative object is indefinite, it is more likely to follow the accusative (when controlling for other factors, including information density). This result can also hint at an explanation for the positive DORM$_{diff}$ value in Table 4 as the influence of the givenness seems to be stronger than the influence of the DORM$_{case}$.

The raw number of constituents in the sentence does not significantly influence the order of objects. In the interaction with an indefinite dative ($z = 3.64$, $p < 0.001$), we can see that a definite dative is still a significant predictor for the DAT > ACC constituent order. However, in long sentences, the likelihood of an indefinite dative preceding a definite accusative increases slightly.

We conclude from these results that the information profile, indeed, influences object order as we hypothesized. Language users are more likely to produce the

[17] The model comparison with anova showed a p-value of 0.11. However, we cannot reduce the model any further because the number of constituents interacts with the definiteness of the dative.

order of objects that results in the more uniform distribution of information. This holds independently of general preferences for the unmarked order DAT > ACC: If placing the accusative before the dative object smoothes the information profile, language users are more likely to produce the marked order ACC > DAT.

What we also see from the regression analysis is that an indefinite dative tends to trigger the order ACC > DAT, i.e., it favors maintaining the default order DEF > INDEF (our proxy for given before new). This finding provides evidence for the influence of information status on object order, as described in Section 2. The effect is larger than for information density, though, which may explain violations of givenness order if that is associated with a more uniform information profile. Also, the interaction of definiteness and the number of constituents shows that, potentially, the importance of givenness decreases with increasing sentence length.

5.2.2 Givenness order

In addition to the investigation of case order, we run a second logistic regression analysis to inspect the effects of information distribution on givenness order. Similar to above, we include DORM$_{giv}$, givenness status, and the number of constituents as predictors[18] and performed a backward model selection, as described above. As shown in Table 6,[19] DORM$_{giv}$ is a significant predictor for givenness order ($z = -3.58$, $p < 0.001$). An increase in DORM$_{giv}$ reduces the likelihood of the DEF > INDEF order. A high DORM$_{giv}$ indicates that the information profile of the DEF > INDEF order is less smooth than the information profile of the INDEF > DEF order. Hence, similar to above, a more uniform, smoother information profile for INDEF > DEF increases the likelihood of observing this marked order. And vice versa, a more uniform distribution of DEF > INDEF increases the likelihood of this default order.

These results may provide insights into the relationship between information theory and information structure. In general, we expect both concepts to make similar predictions regarding the order of objects. Placing a given object before a new object, as preferred by information structure, could help to ease processing of the new object by lowering its surprisal and smoothing the information

[18] glm(formula = def > indef ~(DORM$_{giv}$ + Dat$_{definiteness}$+n_Constituents)^2, family = binomial(), data = constituents_sample); for the complete final regression model, see Table 6. The DEF > INDEF order is coded as 1, the INDEF > DEF order as −1. A definite dative was coded as −1, an indefinite dative as +1. Since DORM$_{case}$ and DORM$_{giv}$ are strongly correlated (r=0.73), we choose to only include one of them as a predictor in each regression analysis.

[19] The model comparison with anova had a p-value of 0.11.

Table 6: Logistic regression with $\mathrm{DORM}_{\mathrm{giv}}$ and number of constituents in the sentence to predict givenness order definite > indefinite

Variable	Estimate	SE	*z*	*p*
Intercept	1.81	0.12	14.68	<0.001 ***
$\mathrm{DORM}_{\mathrm{giv}}$	−0.06	0.02	−3.58	<0.001 ***
$\mathrm{Dat}_{\mathrm{def}}$	−2.72	0.05	−54.45	<0.001 ***
Constituents	−0.03	0.01	−3.40	<0.001 ***

signal. Indeed, we find that language users prefer placing definite, i.e., given objects before indefinite, i.e., new objects if that is associated with a more uniform distribution of information. If, however, the information profile of INDEF > DEF is smoother, this can license a deviation from the default information structure.

Surprisingly, the number of constituents in a sentence ($z = -3.40$, $p < 0.001$) also influences the givenness order: An increase in sentence length predicts the marked order INDEF > DEF. Above, we argued that longer sentences should favor an unmarked object order to counterbalance the effort required for processing the high amount of information in the sentence. Instead, in long sentences, the less frequent givenness order seems to be preferred. In the first regression analysis, we already found this effect for the interaction of an indefinite dative and the number of constituents (Table 5). Here, the effect is predicted independently of case order, which was excluded during *backward model selection.*

Perhaps there are other influences on givenness order in longer sentences. As explained in Section 2, definiteness is only a proxy for givenness that we selected because it does not require complex additional annotations. However, our operationalization is independent of the context in which a constituent occurs, whereas givenness, as defined by Prince (1981), Gundel et al. (1993), and Riester & Baumann (2017), can only be determined from the actual context. The longer the sentence, the more context is given in the sentence itself, probably leading to discrepancies between definiteness and givenness. In particular, longer sentences may include more referents and, therefore, require finer increments of givenness than a binary distinction of *given/definite* vs. *new/indefinite.* In future work, we will explore such effects with a more advanced annotation of givenness.

6 Discussion

The results from the previous section can be interpreted as a confirmation of our assumption that information-theoretic features influence the order of objects in the German middle field (cf. Section 5.1): Small but significant effects of DORM$_{\text{diff}}$ show up within the groups (i) and (ii) with default case order DAT > ACC. No significant effects occur within the groups (iii) and (iv), possibly due to the small size of these groups. Independent of the group size, we could show in Section 5.2 that DORM, i.e., the smoothness of the information profile, can indeed predict the object order in the middle field. Speakers choose the order that results in the most uniform information profile. This holds both for the case order and for the givenness order (cf. Tables 5 and 6).

As we saw in Figure 2, there is a clear preponderance of the unmarked order DAT > ACC. Even though recipients are not consciously aware of the default order, it seems reasonable that they will unconsciously expect the most frequent order of dative preceding accusative. So, if the sentence exhibits the default case order, less cognitive capacity would be consumed for processing the grammar (i.e., case order), according to Futrell et al. (2021). Instead, this capacity would then be free, for example, to process deviations from default givenness order. Similarly, facing the default givenness order (given before new, as reflected by the determiner) would facilitate processing of the unusual case order ACC > DAT. This view is supported by the fact that there is a tendency towards the order DEF > INDEF in sentences with ACC > DAT (cf. Figure 2).

In future work, we want to extend and refine the approach from this pilot study. In particular, we plan to develop improved language models. So far, we used lemma-based bigram models to estimate the probability of observing specific words (and constituents) in different possible orders. Such models reflect lexical or content-based surprisal and can reveal whether a change in object order results in processing advantages on the lexical level. Compared to language models based on word forms, the use of lemmas has the advantage of reducing data sparsity by mapping different word forms to the same lemma. However, this also comes at the price of lemmas being less informative than word forms. In the context of our investigation, this especially concerns case information, which is overtly realized by German determiners but has not been included in our language models. A model based on word forms instead of lemmas could capture the fact that during reading (or listening), the case of objects can already be recognized on the basis of the determiner, helping to reduce entropy early on. Especially in sentences that violate the default order, this could be particularly relevant for processing.

In this pilot study, we have also excluded indefinite plural noun phrases, as they do not have an explicit article in German – and, as a consequence, are shorter than equivalent definite noun phrases, which makes it difficult to compare DORM$_{\text{diff}}$ values across different types of noun phrases. Integrating indefinite plurals into the analysis may give additional insights into the relevance of information-theoretic concepts for object order. For example, we have seen in exemplary observations that the proportion of the default order given>new seems to be even higher for object pairs with an indefinite plural object. Following our aforementioned considerations, this might be due to the missing determiner: because case is marked at the determiner, the recipient cannot easily infer the case of an object realized as an indefinite plural noun phrase without a determiner. In these cases, the meaning of the word and its grammatical case must be processed simultaneously, which might increase the strain on the working memory. Maintaining the default order could be especially beneficial for processing such cases.

Besides the mentioned enhancements, we plan to experiment with language models beyond *n*-grams. Depending on the sentence, the main verb can be located in the left or right sentence bracket, i.e., before or after the objects in the middle field. We assume that it makes a difference whether the main verb was already uttered or not, and that this should affect expectations and, thus, object surprisal. Overall, the majority of verbs in Geman are simple transitive verbs, requiring an accusative object only. In contrast, ditransitive verbs or verbs requiring a dative object are less frequent. If the main verb is located in the left sentence bracket, it is evident at an early stage whether a dative object is to be expected in the sentence. Hence, a dative object located in the middle field should be processed rather easily. In contrast, auxiliaries in the left bracket do not set up any expectations for a dative object. In this case, it might help the recipient to narrow down possible expectations of the verb in the right sentence bracket if the dative object (which is less frequent than an accusative object) occurs first. Due to the limited context, simple bigram models cannot capture such effects, and we plan to experiment with skip-gram models or models based on content words only. Implementing dependency-based models that take into account the relations between object head nouns and full verbs could also shed light on the direct influence of verb valency.

A topic related to the issue of language models is the calculation of surprisal and DORM values. We proposed to investigate the effects of information density on object order by comparing information profiles of original sentences and variant sentences in which we swapped the two objects. We call this the *corpus of variants* method because it allows us to directly inspect the differences between plausible alternative word orders, while keeping other factors constant. However, swapping two objects creates only punctual changes in the information profile of the entire sentence, leading to rather small DORM$_{\text{diff}}$ values. Calculating DORM values only for the local context of the modified parts of the sentence (e.g., as in Figure 3) may return different results and, perhaps, reflect more closely the unfolding of the information flow and resulting effects on local decisions between different structures.

We calculate the DORM$_{\text{diff}}$ values by subtracting the variant DORM values from the original DORM values. We argued in Section 4.2 that a negative DORM$_{\text{diff}}$ value indicates a smoother information profile for the original variant. Since the DORM$_{\text{diff}}$ values are influenced by the length of the sentence, as stated above, the most relevant part of the resulting figures is the algebraic sign, i.e., whether the DORM$_{\text{diff}}$ value is negative or positive. Thus, it should be possible to interpret and use the DORM$_{\text{diff}}$ values as a categorical variable instead of a numerical variable (though we would sacrifice the visibility of gradual changes in doing so).

One area where this study could be further enhanced is by exploring alternative measures for givenness, instead of relying on definiteness as a proxy. We chose this operationalization because it does not require additional complex annotations. However, the binary distinction of *given/definite* vs. *new/indefinite* may not be accurate enough, especially in longer sentences or longer contexts in general. We plan to work on creating more nuanced annotations of givenness and inspect how this influences the order of objects in the middle field. Furthermore, we intend to also include objects with the same givenness status in the investigation to confirm that the information profile has an influence on the object order without being also influenced by the givenness.

Finally, it is yet an open question how the current order preferences have been established. In future work, we want to extend the experiments to historical German. In historical language stages of German, the word order was generally more flexible than in modern German. Crucially, this also holds for dative and accusative objects, which showed much more variation with respect to their relative order than nowadays. However, similar factors as in modern German already played a role, in particular givenness (Rauth 2020). Hence, in the long term, we are interested in investigating how information density relates to object order variation in historical German. Furthermore, a diachronic analysis could provide insight into the historical development of object order and reveal which role information density might have played diachronically, ultimately resulting in the clearly-preferred order of objects (dative before accusative) as we observe them for modern German.

Using the proposed methods, we will investigate how the object order in historical data can be explained. In a second step, we will trace the development to modern German and inspect relevant factors that contributed to the formation of modern standard object order. A prerequisite is that we can control for other factors besides length, in particular animacy, which plays an important role in language and cognitive processing.

7 Conclusion

In this paper, we have motivated the order of dative and accusative objects in the German middle field with information-theoretic concepts, while controlling for the factor length.

Overall, the corpus data shows an exceedingly strong bias for the unmarked orders (DAT > ACC in 96% and DEF > INDEF in 88% of the cases). As we hypothesized, the corpus sentences are in general characterized by a more uniform information profile than the generated swapped variants. This is true for corpus sentences with the default order DAT > ACC. This observation is confirmed by logistic regression models in which lower $\text{DORM}_{\text{case}}$ and DORM_{giv} values increase the likelihood of the marked orders (accusative before dative, new before given). We thus argue that deviations from the default orders can be explained by more uniform information profiles, which improve overall sentence processing.

In future work, we will extend the proposed approach to historical data. We plan to investigate how the modern order preferences have been established and which role information-structural and information-theoretical factors may have played in this process.

Acknowledgements

The work reported in this paper has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project ID 232722074 – SFB 1102 “Information Density and Linguistic Encoding”.

Our thanks goes to the reviewers and editors of this volume for many helpful comments, furthermore to Magdalena Meiser for helping with the light verb constructions. All remaining errors are of course ours.

References

Aylett, Matthew & Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56. DOI: 10.1177/00238309040470010201.

Behagel, Otto. 1932. *Deutsche Syntax: Eine geschichtliche Darstellung*, vol. 4: Wortstellung, Periodenbau (Germanische Bibliothek). Heidelberg: Winter.

Bohnet, Bernd. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 89–97. Beijing, China.

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith & Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation* 2(4). 597–620. DOI: 10.1007/s11168-004-7431-3.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots. Linguistics in search of its evidential base*, 75–96. Berlin, New York: De Gruyter Mouton. DOI: 10.1515/9783110198621.75.

Cuskley, Christine, Rachael Bailes & Joel Wallenberg. 2021. Noise resistance in communication: Quantifying uniformity and optimality. *Cognition* 214. 104754. DOI: 10.1016/j.cognition.2021.104754.

Eisenberg, Peter. 2020. *Grundriss der deutschen Grammatik*. 5., aktualisierte und überarbeitete Auflage. Berlin: J. B. Metzler.

Engel, Alexandra, Jason Grafmiller, Laura Rosseel & Benedikt Szmrecsanyi. 2022. Assessing the complexity of lectal competence: The register-specificity of the dative alternation after give. *Cognitive Linguistics 2022* 69. 727–766. DOI: 10.1515/cog-2021-0107.

Faaß, Gertrud & Kerstin Eckart. 2013. SdeWaC: A corpus of parsable sentences from the web. In Iryna Gurevych, Chris Biemann & Torsten Zesch (eds.), *Language processing and knowledge in the web*, 61–68. Berlin, Heidelberg: Springer.

Fenk-Oczlon, Gertraud. 1983. Ist die SVO-Wortfolge die natürlichste? *Papiere zur Linguistik* 29. 23–32.

Féry, Caroline & Manfred Krifka. 2008. Information structure. Notional distinctions, ways of expression. In Piet van Sterkenburg (ed.), *Unity and diversity of languages*, 123–136. Amsterdam: Benjamins.

Frey, Werner. 2004. A medial position for topics in German. *Linguistische Berichte* 198. 153–190.

Futrell, Richard, Edward Gibson & Roger Levy. 2021. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science* 44(3). e12814. DOI: 10.1111/cogs.12814.

Gries, Stefan Th. 2021. *Statistics for linguistics with R. A practical introduction.* Berlin, Boston: De Gruyter Mouton. DOI: 10.1515/9783110718256.

Gundel, Jeanette K., Nancy Hedberg & Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69. 274–307.

Hale, John. 2001. A probabilistic Early parser as a psycholinguistic model. *Proceedings of the second meeting of the North American chapter of the Association for Computional Linguistics (NAACL'01).* 159–166. DOI: 10.3115/1073336.1073357.

Hoberg, Ursula. 1981. *Die Wortstellung in der geschriebenen deutschen Gegenwartssprache.* München: Hueber.

Jeffreys, Harold. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London Series A: Mathematical and Physical Sciences* 186. 453–461.

Lenerz, Jürgen. 1977. *Zur Abfolge nominaler Satzglieder im Deutschen.* Tübingen: Narr.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177. DOI: 10.1016/j.cognition.2007.05.006.

Levy, Roger & T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schölkopf, John Platt & Thomas Hofmann (eds.), *Advances in neural information processing systems 19: Proceedings of the 2006 conference*, 849–856. Cambridge: The MIT Press. DOI: 10.7551/mitpress/7503.003.0111.

Ortmann, Katrin. 2021. Automatic phrase recognition in historical German. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, 127–136. Düsseldorf. https://aclanthology.org/2021.konvens-1.11.

Petrov, Slav, Leon Barrett, Romain Thibaux & Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 433–440. Sydney, Australia. DOI: 10.3115/1220175.1220230.

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In P. Cole (ed.), *Radical pragmatics*, 223–254. New York: Academic Press.

ProGram2.0. 2018. *Liste der Funktionsgefüge.* https://program.idf.uni-heidelberg.de/fvg/liste.

R Core Team. 2018. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

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

Rauth, Philipp. 2020. *Ditransitive Konstruktionen im Deutschen: Geschichte und Steuerung der Objektabfolge im Mittelfeld*. Tübingen: Stauffenburg.

Riester, Arndt & Stefan Baumann. 2017. The RefLex scheme – annotation guidelines. *SinSpeC* 14.

Schiller, Anne, Simone Teufel, Christine Stöckert & Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27(3). 379–423.

Speyer, Augustin. 2011. Die Freiheit der Mittelfeldabfolge im Deutschen – ein modernes Phänomen. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 133. 14–31.

Speyer, Augustin. 2013. Mündlichkeitsnähe als Faktor für die Objektstellung im Mittel- und Frühneuhochdeutschen. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 135. 1–36.

Speyer, Augustin. 2015. Object order and the thematic hierarchy in Older German. In Jost Gippert & Ralf Gehrke (eds.), *Historical corpora: Challenges and perspectives*, 101–124. Tübingen: Narr.

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister & Kathrin Beck. 2017. *Stylebook for the Tübingen treebank of written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen. https://www.sfs.uni-tuebingen.de/resources/tuebadz-stylebook-1707.pdf.

Winter, Bodo. 2020. *Statistics for linguists: An introduction using R*. New York: Routledge, Taylor & Francis Group.

Wöllstein, Angelika. 2010. *Topologisches Satzmodell*. Heidelberg: Winter.

Wöllstein, Angelika. 2014. *Topologisches Satzmodell*. 2nd edn. Heidelberg: Winter.