Glossa a journal of general linguistics

Ellsiepen, Emilia and Markus Bader. 2018. Constraints on Argument Linearization in German. *Glossa: a journal of general linguistics* 3(1): 6. 1–36, DOI: https://doi.org/10.5334/gjgl.258

RESEARCH

Constraints on Argument Linearization in German

Emilia Ellsiepen and Markus Bader

Goethe Universität Frankfurt, Norbert-Wollheim-Platz 1, 60629 Frankfurt, DE

Corresponding author: Emilia Ellsiepen (emilia.ellsiepen@gmail.com)

In this article we present experimental findings on the acceptability of different argument orders in the German middle field. Our study pursues two goals: First, to evaluate a number of surface constraints on German argument order that have been proposed in the literature, and second, to shed new light on how gradient constraints jointly determine sentence acceptability. In four experiments, we investigated the impact of surface constraints relating to animacy, thematic roles, definiteness and case. While we are able to confirm an influence of most constraints under investigation, the resulting constraint hierarchy does not coincide with any hierarchy put forward so far in the literature, to the best of our knowledge. With regard to gradience, our results can be accounted for either by an OT variant incorporating a notion of markedness, or by a fully quantified model using constraint weights. For the latter, however, we provide evidence against uniform penalties associated with constraint violations.

Keywords: word order; German; magnitude estimation; weighted constraints

1 Introduction

The order of verb arguments in the German middle field[1] is known to exhibit a fair degree of variability. At the same time, not all possible orders are equally acceptable or can occur in all contexts. The rich literature on the topic has identified several factors that influence relative acceptability, but the applicability and importance of individual factors is not generally agreed upon. This paper presents a series of experiments that tease apart the contribution of individual factors to word order acceptability.

As an example of the flexibility of argument order in the German middle field, consider a subordinate clause containing a subject, a dative object and an accusative object. (1) shows three out of six possible orders.

(1) a. ..., dass [der Vater] [dem Sohn] [das Buch] geschenkt hat.
that the.**NOM** father the.**DAT** son the.**ACC** book given has
'..., that the father has given the son the book.'

b. ..., dass [der Vater] [das Buch] [dem Sohn] geschenkt hat.
the.**NOM** the.**ACC** the.**DAT**

c. ..., dass [dem Sohn] [der Vater] [das Buch] geschenkt hat.
the.**DAT** the.**NOM** the.**ACC**

All orders are considered to be grammatical, but they are not equally acceptable. The order in (1a), where the subject precedes the dative object, which in turn precedes the

[1] The middle field of a German sentence is that part of the sentence that starts after the finite verb in main clauses and after the complementizer in embedded clauses and ends before the verb in clause-final position.

accusative object, is the canonical order for ditransitive verbs in German. In accordance with this, Pechmann et al. (1996) obtained experimental results with a range of different procedures showing that acceptability is highest for sentences like (1a). For sentences with non-canonical order, acceptability decreased to varying degrees. For example, the decrease brought about by switching the order of the two objects in (1b) was less severe than the decrease caused by putting an object in front of the subject as in (1c) (see also Rösler et al. 1998 for neuropsychological evidence). Taking findings like these as starting point, we systematically evaluate the explanatory power of a collection of factors proposed in the literature.

A second topic addressed by our experiments is the issue of "gradience in grammar", to borrow the title of a book providing a broad overview of research concerned with the relationship between grammar and gradience (Fanselow et al. 2006). A growing body of research shows that the acceptability of linguistic structures is a gradient property (see Schütze & Sprouse 2014, for a recent overview and further references). For example, rather than categorizing sentences as grammatical or ungrammatical, native speakers are able to make fine-grained judgments using experimental procedures like magnitude estimation or rating on a Likert scale. Even when forced to make binary decisions, as in the classical grammaticality judgment task, a continuum between strictly unacceptable and perfectly acceptable is observed when results are averaged across participants, items, or both (Bader & Häussler 2010a; Weskott & Fanselow 2011; Fukuda et al. 2012).

Following Bard, Robertson & Sorace (1996: 33), we distinguish between "*grammaticality*, a characteristic of the linguistic stimulus itself, *acceptability*, a characteristic of the stimulus as perceived by a speaker, and the *acceptability judgment* which is the speaker's response to the linguist's inquiries." With regard to the judgment of acceptability, all experiments presented below make use of the method of magnitude estimation (Bard, Robertson & Sorace 1996; Cowart 1997), which provides a fine-grained measure of relative acceptability. Although there are ongoing discussions as to which judgment procedure is best suited for measuring acceptability (see Schütze & Sprouse 2014 for an overview), obtaining acceptability judgments does not seem to be controversial as such. As pointed out by Schütze & Sprouse (2014: 28), "[a]cceptability is just like other percepts (e.g., brightness, loudness, temperature) in that there are no methods for directly measuring the percept as it exists within a participant's mind." Acceptability judgments are thus akin to judgments in psychophysics, from which the magnitude estimation procedure has been borrowed, and do not require from participants to engage in introspection in the sense of reporting what is going on in one's own mind (see Goodwin 2003 for a historical discussion of this distinction).

The focus of our research lies on the relationship between grammaticality and acceptability. That these two notions must be kept distinct is shown, among others, by the observation that some sentences are grammatical but unacceptable (e.g., garden-path sentences or multiply center-embedded sentences) whereas other sentences are ungrammatical but acceptable (grammatical illusions; see overview in Phillips, Wagers & Lau 2011).

While there is no disagreement that acceptability is a gradient property, the source of the observable gradience is an unresolved issue, which has become an area of extensive research (Fanselow et al. 2006; Schütze & Sprouse 2014). In this paper, we follow a widespread conception of linguistic intuitions and assume that the acceptability of a sentence is the joint product of linguistic knowledge – the competence grammar – and processing mechanisms that put the linguistic knowledge to use. According to this conception, gradient acceptability can have its source in the grammar, in the processing

mechanisms, or in both. The inherent difficulty of deciding what causes an observed acceptability difference is known as the *source ambiguity problem* (see Hofmeister et al. 2013 for further discussion). It seems uncontroversial that variation in processing complexity can cause different degrees of acceptability. For example, garden-path sentences differ widely in acceptability, depending on how easy or difficult it is for the human parser to recover from a garden-path (see Fodor & Ferreira 1998). The need to compute non-local dependencies provides a different source of processing complexity that affects the acceptability of sentences (Gibson 2000). Non-local dependencies can make a sentence unacceptable under certain circumstances, for example, when too many nested non-local dependencies must be processed, as in sentences containing multiply center-embedded clauses. Less severe manipulations of sentence complexity, however, typically result in fine-grained degrees of acceptability (e.g., Warren & Gibson 2002).

In principle, all gradience observed for acceptability judgments could be due to variations in processing complexity. However, we do not know of any extant processing model that comes with the aspiration to provide a general account of gradience in grammar. Instead, processing models try to account for the incremental on-line processing of sentences. In selected cases, this aim subsumes findings concerning sentence acceptability (e.g., research on island constraints; cf. Hofmeister & Sag 2010; Sprouse, Wagers & Phillips 2012), but this is still far from providing a general account of gradient acceptability.[2] Future processing models may be more powerful in this regard, making it possible to attribute gradient acceptability to the performance mechanisms in toto. Meanwhile, however, we should also consider the grammar itself as a further source of gradient acceptability. Note that even in classical generative syntax, this possibility has been entertained from time to time. For example, Chomsky (1955/1975, 1965) hypothesized that different syntactic violations are associated with different degrees of unacceptability (see Schütze 1996, for discussion). Later work in the framework of the Government and Binding Theory assumed that violations of the Empty Category Principle (ECP) decrease acceptability to a larger extent than violations of the Subjacency Principle.

The seminal work by Bard, Robertson & Sorace (1996) and Cowart (1997) led to a growing number of studies that apply experimental methods to syntactic issues. Inspired at least partly by this strand of research, grammar formalisms have been developed that provide a principled way to assign fine-grained degrees of grammaticalness to all sentences of a language by integrating numeric information into the grammar. Such grammar formalisms go beyond the classification of sentences into a small number of degrees of grammaticality (possibly just two, grammatical and ungrammatical), and they do no longer restrict degrees of grammaticality to unacceptable sentences. Among these formalisms are certain variants of Optimality Theory (OT), including the OT-predecessor Harmonic Grammar (HG) (Legendre, Miyata & Smolensky 1990; Pater 2009) and Linear Optimality Theory (LOT) (Keller 2000, 2006). With standard OT (Prince & Smolensky 1993/2004), these OT variants share the assumption that constraints are violable, in the sense that a violation does not necessarily rule out a candidate. In contrast to standard OT, however, constraints are not applied in a discrete way. Instead, each constraint is associated with a numeric weight. These weights are used in turn to assign a harmony H to each sentence S in a given candidate set, where harmony is defined as the negative weighted sum of all grammatical constraints C_i that are violated by S.

[2] An exception is Hawkins (2006), where the efficiency theory of Hawkins (2004) is proposed as a model of gradient acceptability. However, due to its programmatic nature and its limited coverage, this proposal is not applicable to the syntactic constructions under considerations in this paper.

(2) **Harmony of a candidate sentence S**

$$H(S) = -\sum_i w(C_i)\, v(S, C_i)$$

where $w(C_i)$: the weight of constraint C_i,
$v(S, C_i)$: the number of violations of constraint C_i in sentence S

By itself, the notion of harmony does not yet provide a graded notion of acceptability. As discussed in Pater (2009), one way to make use of harmony is by defining a binary notion of grammaticality. The candidate with the highest harmony least severely violates the constraints of a given constraint hierarchy and can thus be declared the winning candidate, similarly to determining the winning candidate by Standard OT's evaluator. However, one can also go a step further and equate harmony with grammaticality, resulting in a gradient notion of grammaticality. In conjunction with appropriate linking assumptions, harmony can then be related to frequency of occurrence and/or to acceptability. In this paper, the focus lies on the relationship between harmony and acceptability (for the relationship between harmony and frequency, see Goldwater & Johnson 2003 and references in Pater 2009; see also section 6). Linear Optimality Theory (LOT) takes a particularly strong position in this regard. According to LOT, the grammar itself constitutes a second source of gradient acceptability, in addition to the gradience rooted in the performance mechanisms. More specifically, each constraint violation results in a decrease in acceptability proportional to the weight associated with the constraint. The second goal of this paper is to test the generality of violation costs predicted by LOT within the domain of word order in German.

In the next section, we review major accounts of word order variation and introduce the individual constraints under investigation. In section 3, we present four acceptability judgment experiments that assess the constraints and establish a ranking among them. In section 4, we use the experimental results to evaluate different quantitative models that use weighted constraints. In section 5, we relate our final constraint hierarchy to existing categorical accounts and compare the experimental results to evidence from corpus studies.

2 Constraints on order

Most accounts of German word order identify one order for a given sentence as the canonical or unmarked order (see Lenerz 1977; Höhle 1982 and much subsequent work). This unmarked order is characterized by high acceptability and the greatest focus potential: It can carry wide focus as well as narrow focus on subconstituents, whereas other orders are restricted to narrow focus. Despite the importance of focus, sentence acceptability is often judged without a context in the literature, with the intuition that highest acceptability out of context corresponds to wide focus potential. We follow this tradition in our experiments and test relative acceptability out of context. With regard to deviating word orders, the correspondence between acceptability and focus potential is less clear. Gradience may directly reflect mismatches between focus potential and context. In this view, a degraded sentence should become fully acceptable if a licensing context is provided. Alternatively, gradience may be related to performance factors or generated by the grammar in addition to specific focus patterns.

Syntactic accounts of word order variation differ in several ways. One point of divergence concerns the formal means used to generate word order variants. Some accounts assume that argument order variation is derived by a syntactic movement operation called *scrambling*, which moves constituents to the left of other phrases. A prominent account of this type has been proposed by Haider & Rosengren (2003), who

assume that each sentence has an underlying base order determined by the semantics of the verb (see also Haider 2010). By applying scrambling, deviations from the base order become possible. According to Haider (2010), this operation is strictly optional and not triggered by syntactic features. He also assumes that all word orders obtained by scrambling carry narrow focus, which is consistent with a lower acceptability out of context. Furthermore, it has been shown that scrambling is subject to certain restrictions, e.g., indefinite NPs may not scramble (Lenerz 2001). A rather different scrambling-based account of word order variation has been proposed by Müller (1999). This account will be discussed in detail below (further movement-based accounts to word order variation in German are, among others, Frey 1993; Meinunger 2000). In contrast to accounts involving movement, pure base-generation accounts (e.g., Fanselow 2001, 2003) assume that all word orders, whether giving rise to narrow or wide focus, are base generated.

Another question on which models of word order variation diverge concerns the role of *surface constraints* such as 'animate > inanimate' or 'definite > indefinite'. In some models, surface constraints do not play any role and are just descriptive generalizations (e.g., Haider & Rosengren 2003; Haider 2010). This is different for competition based models, including but not limited to the above mentioned OT, HG and LOT. In such models, surface constraints are used to determine which orders are grammatical and/or preferred (Uszkoreit 1987; Müller 1999; Heck 2000; Keller 2000). Independent evidence for surface constraints comes from psycholinguistic studies on language production, suggesting them to play a role in choosing between different orders (see Jaeger & Norcliffe 2009 for a recent summary).

Different sets of surface constraints have been proposed in the literature (see Table 1 for an overview). One reason for this variability is that the relevant properties are often confounded. The source of the canonical order of the example in (1), repeated below, is such a case.

(3) Ich glaube, dass der Vater dem Sohn ein Buch geschenkt hat.
I think that the.NOM father the.DAT son a.ACC book given has
a. Case: **nominative** > **dative** > **accusative**
b. Thematic roles: **agent** > **recipient** > **theme**
c. Animacy: **animate** > **animate** > **inanimate**
d. Definiteness: **definite** > **definite** > **indefinite**

Table 1: Constraint sets used by selected accounts of German word order.

Uszkoreit (1986)	**Jacobs (1988)**	**Heck (2000)**
Agent < Theme	Agent < Non-Agent	(NOM< ACC)[i]
Agent < Goal	Recipient[ii] < Patient	Definite < Indefinite
Goal < Theme	Definite < Indefinite	Animate < Inanimate
		Agent < Non-Agent
Uszkoreit (1987)	**Hoberg (1997)**	**Müller (1999)**
NOM < NON-NOM	Animate < Inanimate	NOM < NON-NOM
DAT < ACC	NOM < ACC < DAT	Definite < Indefinite
		Animate < Inanimate
		DAT < ACC

[i] implicitly as part of the generator GEN.
[ii] Jacobs actually refers to all thematic roles that are usually expressed by dative in German, i.e., recipient, animate goal, benefactive etc.

The particular order of subject, indirect object and direct object in the subordinate clause could be due to a constraint in syntax (nominative > dative > accusative) or to a constraint concerning the thematic roles assigned by the specific verb (agent > recipient > theme). In addition, the lexical and discourse semantic properties of the two objects – animacy and definiteness – also favor the order dative before accusative object, according to constraints like 'animate > inanimate' and 'definite > indefinite'. Our goal is to tease apart these factors and investigate the interplay between them. In the remainder of this section, we present the constraint set under investigation and outline two different competition models of word order variation. The OT model of Müller (1999) is close to traditional accounts in that it accommodates gradience by establishing an ordinal ranking of alternative word orders based on markedness. Keller's LOT, on the other hand, is a fully quantified competition model, where acceptability can be derived from the grammar, while grammaticality in the traditional sense becomes an obsolete term.

2.1 Constraint candidates

In our experiments, we assess six surface constraints that can be assigned to three different categories, namely lexical-semantic constraints, syntactic constraints and discourse constraints:

(4) **Lexical-semantic constraints**
- a. ANI: animate > inanimate
- b. AG: agent > non-agent
- c. REC: recipient/goal/benefactive > theme

(5) **Syntactic constraints**
- a. NOM: nominative > non-nominative
- b. DAT: dative > accusative

(6) **Discourse constraints**
- a. DEF: definite > indefinite

ANI (animate > inanimate) Constraint (4a) refers to the animacy status of arguments and states that animate NPs should precede inanimate ones. This pattern is in line with the general tendency to put dative NPs, which are prevalently animate, before accusative NPs in the German middle field. Some authors regard this constraint as one factor within a set of competing surface constraints (Lenerz 1977; Hoberg 1981; Müller 1999). Other authors dismiss this constraint and consider it a mere confound of case or thematic roles (Uszkoreit 1986; Haider & Rosengren 2003).

AG (agent > non-agent) Constraint (4b) refers to the thematic roles of arguments and states that an agentive NP should precede all other argument NPs. While this constraint could be partly responsible for the canonical order of subject before object in German, it might also be confounded with either the syntactic constraint equivalent NOM in (5a), or with ANI, as agents are mostly animate entities. This constraint is part of Uszkoreit's (1986) set of constraints[3] and is entailed by Haider & Rosengren's (2003) account of base order in terms of thematic roles.

REC (recipient/goal/benefactive > theme) Constraint (4c) is an alternative to ANI in accounting for the predominant order of indirect before direct object in German,

[3] Uszkoreit (1986) breaks down the constraint in (4b) into the two constraints "agent precedes theme" and "agent precedes goal". This divide does not result in predictions different from ours.

making reference to thematic roles. It is advocated in slightly different forms in Haider & Rosengren (2003) and Uszkoreit (1986).

NOM (nominative > non-nominative) Constraint (5a) refers to the syntactic status of arguments and states that nominative subjects precede accusative or dative objects. While this constraint has been taken to be the source of the canonical SO word order in German and is often considered more important than other constraints (Müller 1999), it has also been noted that it may be violated in certain contexts without a cost (Lenerz 1977).

DAT (dative > accusative) Constraint (5b) is the syntactic counterpart of REC and states that dative NPs precede accusative NPs. This constraint has been proposed by different authors (Uszkoreit 1987; Müller 1999), sometimes in addition to seemingly competing constraints like ANI. While it appears to be at odds with traditional case hierarchies, e.g., the accessibility hierarchy of Keenan & Comrie (1977), it could be triggered by an underlying constraint requiring structural case (accusative) to appear adjacent to the verb, as argued in Heck (2000).

DEF (definite > indefinite) Constraint (6a) states that definite NPs should precede indefinite NPs. It was proposed in this simple surface oriented form in Lenerz (1977) and Müller (1999). Alternatively, it is conceptualized as a result of banning indefinite NPs from scrambling (Büring 2001; Lenerz 2001).

2.2 Capturing markedness within OT: Müller (1999)

Some of the surface constraints described above form part of the account of Müller (1999), which extends standard OT to accommodate markedness. In accordance with much research in syntactic theory, standard OT classifies sentences as either grammatical or ungrammatical and offers no direct way to differentiate between degrees of grammaticality or markedness. In Müller's variant of OT, alternative word orders can be grammatical and nevertheless differ in terms of markedness, in the sense of a dispreference of a marked structure compared to other less marked structures.

Müller's account is also notable for combining minimalist syntax and OT. Like Haider & Rosengren (2003), Müller assumes that argument order variation is derived by scrambling. The underlying base order is not determined by semantic properties of the verb, however, but is assumed to be the uniform configuration "subject > direct object > indirect object" at D-structure. Deviations from this configuration can be achieved by scrambling the indirect or direct object when this is licensed by one of the constraints in (7). The last constraint, PER, counteracts existing lower ranked faithfulness constraints (STAY, PAR-MOVE) by favoring linearizations where the order of arguments is reversed relative to the base-generated order.

(7) SCR-CRIT (Müller 1999): In the VP domain,
- a. NOM ("nominative constraint"): [+NOM] precedes [–NOM] >
- b. DEF ("definiteness constraint"): [+definite] precedes [–definite] >
- c. ANI ("animacy constraint"): [+animate] precedes [–animate] >
- d. FOC ("focus constraint"): [–focus] precedes [+focus] >
- e. DAT ("dative constraint"): [+dative] precedes [+accusative] >
- f. ADV ("adverb constraint"): [+NP] precedes [+adv] >
- g. PER ("permutation constraint"): If α c-commands β at level L_n, then α does not c-command β at level L_{n+1}

Any word order resulting from a triggered scrambling operation is grammatical, in addition to the base-generated order, which is always grammatical. This is implemented by means of a subhierarchy, subsumed under the constraint SCR-CRIT, which may be replaced by

any of its subconstraints at evaluation. Formally then, every structure is grammatical that is optimal for at least one replacement of SCR-CRIT.

Müller (1999) accounts for relative grammaticality, or markedness, by imposing an internal ranking (>) on the constraints in the subhierarchy. To find the unmarked candidate for a given content, SCR-CRIT is replaced with the *whole* subhierarchy and evaluation takes place as usual. As one or more constraints in the subhierarchy might favor a scrambled word order, the base order is not in general the unmarked one, contrary to Haider & Rosengren (2003). For example, the direct object is base generated before the indirect object, but the dative constraint favors the scrambled order and renders it the unmarked candidate, if the candidates do not differ on higher ranked dimensions. The *candidate set* is defined by Müller (1999) as consisting of <D-structure, S-structure> pairs, which share the same numeration, i.e., the same lexical content. He also seems to assume the same meaning, which results in the set of <D-structure, S-structure> pairs with the same D-structure.

All unmarked structures that are identified by the extended evaluation including the whole subhierarchy, are expected to be of approximately equal, high acceptability, as long as they do not differ substantially with respect to complexity. For the remaining candidates, an order of markedness can be established using the whole subhierarchy and the concept of suboptimality introduced in Keller (1996): Candidate C_i is more marked than Candidate C_j, iff C_i is suboptimal to C_j. This definition imposes a ranking in terms of markedness, but does not quantify the differences in acceptability. In addition, it does not allow to compare sentences from different candidate sets, i.e., sentences that do not directly compete because they differ in content or information structure.

2.3 Weighted constraints: Keller (2000)

Linear Optimality Theory (LOT) proposed in Keller (2000) provides a fully quantified model of constraint application. Similar to Müller (1999), Keller uses surface constraints to account for word order preferences in German, in particular testing the relative importance of the equivalents of Müller's NOM, DAT and FOC constraints. In his theoretical framework, each constraint is associated with a numeric weight and the harmony of a sentence is defined as the weighted sum of the constraint violations, as in the formula in (2).

Keller assumes that differences in harmony are directly reflected in acceptability differences that can be observed empirically. When two candidate structures S_1 and S_2 differ only in that candidate S_1 violates constraint C_i whereas S_2 does not, then S_1 will be less acceptable than S_2 to an extent that is proportional to the numeric weight of C_i. The decrease in acceptability that is caused by a violation of C_i should be approximately the same in all candidate sets containing two candidates differing only with regard to C_i.

What is not defined directly in this model is the effect that the violation of a constraint has on the optimal candidate, that is, the candidate with the highest harmony. When a constraint C_i is violated in the optimal candidate, this should also lead to a decrease in acceptability, but this decrease cannot be measured because there is no better candidate that differs from the optimal one only in not violating C_i (otherwise, this candidate would have an even higher harmony and thus would be the optimal candidate). An obvious way to quantify the effect that violating C_i has on an optimal candidate would be to compare two optimal candidates that differ only with respect to C_i. In the formal statement of LOT given in Keller (2000), this is not allowed because the relationship between harmony and acceptability is limited to sentences within the same candidate set, that is, different realizations of the same content. Below, we discuss an informal proposal by Keller that allows a comparison across candidate sets under certain narrow conditions.

Keller's account makes interesting and testable predictions. First, constraint violations are cumulative. Therefore, in contrast to OT's notion of strict domination, multiple violations of low-weighted constraints might outweigh a single violation of a higher-weighted constraint.[4] Secondly, the violation of a constraint leads to the same numeric acceptability decrease in every candidate set. In contrast to Müller's proposal, we can thus quantify the degree of markedness directly and are not limited to a categorical ranking of candidates. These predictions are summarized in the *Uniform Penalty Hypothesis* given in (8). The weaker version only includes the comparison of candidates from a single candidate set, as stated in (8a), and corresponds to Keller's formalization. The strong version, which includes both (8a) and (8b), claims that comparisons across candidate sets are possible if all constraints and their weights are known.

(8) **Uniform Penalty Hypothesis**
Each constraint C_j is associated with a fixed penalty p. Whenever C_j is violated, acceptability decreases proportional to p.

a. **Within candidate set comparison:** When two sentences instantiating two candidates from a single candidate set A differ only in the violation of Constraint C_j, they will differ in acceptability proportional to $w(C_j)$.

b. **Across candidate set comparison:** When two optimal sentences, one from a candidate set A and one from a candidate set B, differ only in the violation of Constraint C_j and the corresponding change in meaning, but are matched in terms of complexity otherwise, they will differ in acceptability proportional to $w(C_j)$.

As noted above, the Uniform Penalty Hypothesis contrasts with standard OT, where constraint violations by the optimal candidate do not affect acceptability. These include violations common to all candidates in the same candidate set and violations of constraints ranked below the fatally violated constraint.

3 Experiments

We conducted four magnitude estimation experiments in order to assess the individual importance and applicability of the constraints presented above as well as their interplay. The experiments were designed to tease apart the contributions of each constraint and to establish a ranking between them. In the interest of intelligibility, we postpone the discussion of constraint weights and the Uniform Penalty Hypothesis to the next section.

Experiment 1 teases apart the influence of the often confounded constraints ANI, REC, DAT and NOM on the order between the dative recipient object and the theme in ditransitive clauses. Experiment 2 replicates the effects of Experiment 1 for sentences with inanimate themes and contrasts ANI with DEF. In Experiment 3, ANI and DEF are contrasted with AG. In Experiment 4, we test the influence of NOM on the order between accusative and nominative.

3.1 Experiment 1

In Experiment 1, we test the applicability and ranking of ANI, REC, DAT, and NOM with regard to the order of arguments in the German middle field. We concentrate on ditransitive verbs and manipulate the order of the two objects, the animacy of the direct object, and the voice of the verb. Let us first consider cases where ANI and REC make the

[4] To achieve cumulativity in standard OT, the concept of constraint conjunction can be used, where two lower ranked constraints form a unified constraint which is ranked higher than the individual ones. In contrast to Keller's model, this would only affect a subset of constraints.

same predictions, that is, sentences containing an animate recipient and an inanimate theme. For active sentences, ANI, REC, and DAT all predict a preference for orders where the dative object precedes the accusative:

(9) Der Internatsleiter sagte, ('The warden said,')

a. **animate-recipient > inanimate-theme**
dass man dem Erzieher den Bericht gebracht hat.
that one.NOM the.DAT educator the.ACC report brought has
'that someone brought the report to the educator.'

b. **inanimate-theme > animate-recipient**
dass man den Bericht dem Erzieher gebracht hat.
that one.NOM the.ACC report the.DAT educator brought has

In order to tease apart the influence of ANI and REC on the one hand and DAT and NOM on the other hand, Experiment 1 includes the passive counterparts of the sentences in (9). The examples below only include the dative 1st order, the full example can be found in Table 2.

(10) **animate-recipient > inanimate-theme**
dass dem Erzieher der Bericht gebracht wurde.
that the.DAT educator the.NOM report brought was
'that the report was brought to the educator.'

In contrast to the active sentence, DAT is not applicable here, as there is no accusative object. Furthermore, NOM now competes with ANI and REC: If the dative 1st order is still preferred, we can conclude an influence of ANI or REC and their ranking above NOM. A preference for dative 2nd, on the other hand, would be in line with a high-ranked NOM constraint as advocated by Müller (1999) and others.

Table 2: Example item from Experiment 1.

Der Internatsleiter sagte, ('The warden said',)				
Cond	**Animacy**	**Voice**	**Order**	**Stimulus continuation**
1	INANI	ACT	DAT 1ST	dass **man dem Erzieher den Bericht** gebracht **hat**.
2			DAT 2ND	dass **man den Bericht dem Erzieher** gebracht **hat**.
				that one the.ACC report the.DAT educator brought has
				'that someone has brought the report to the educator.'
3		PASS	DAT 1ST	dass **dem Erzieher der Bericht** gebracht **wurde**.
4			DAT 2ND	dass **der Bericht dem Erzieher** gebracht **wurde**.
				that the.NOM report the.DAT educator brought was
				'that the report was brought to the educator.'
5	ANI	ACT	DAT 1ST	dass **man dem Erzieher den Jungen** gebracht **hat**.
6			DAT 2ND	dass **man den Jungen dem Erzieher** gebracht **hat**.
				that one the.ACC boy the.DAT educator brought has
				'that someone has brought the boy to the educator.'
7		PASS	DAT 1ST	dass **dem Erzieher der Junge** gebracht **wurde**.
8			DAT 2ND	dass **der Junge dem Erzieher** gebracht **wurde**.
				that the.NOM boy the.DAT educator brought was
				'that the boy was brought to the educator.'

In a last step, we tease apart REC and ANI by including sentences with two animate NPs:

(11) **animate-recipient > animate-theme**
dass man dem Erzieher den Jungen gebracht hat.
that one the.DAT educator the.ACC boy brought has
'that someone brought the boy to the educator.'

For active sentences, where we expect a strong preference for dative 1st orders with inanimate themes, a modulation of this preference caused by the animacy manipulation would be strong evidence in favor of ANI. If the preference is completely neutralized, REC and DAT can be dismissed. The passive versions instantiate a competition of REC and NOM and will thus enable us to establish a ranking between them.

3.1.1 Method

Participants. Sixty-four students from the Goethe University Frankfurt took part in the experiment. All were native speakers of German. The experiment took about 20–25 minutes to complete. For the whole experimental session, which included one or two additional experiments and took approximately one hour, they received either partial course credit or a compensation of 8 Euro.

Materials. Thirty-two experimental items were created. A full example item can be found in Table 2. Each experimental sentence started with a short introductory main clause, followed by a complement clause that contained two NPs and an indefinite subject pronoun, which was dropped in the passive conditions. The dative object NP was always animate. Three variables were manipulated in a factorial design: Order between the dative NP and the theme NP (Order: dative 1st, dative 2nd), voice (Voice: active, passive), and animacy of the theme NP (Animacy: animate, inanimate), where this NP was accusative in the active conditions and nominative in the passive conditions. Our focus in constructing the items was on the plausibility of the sentences for both lexicalizations of the theme NP. In addition, we controlled for length as measured by the number of syllables. Three pairwise t-tests confirmed that there were no significant differences between animate and inanimate second NP ($t(31) = -0.53$, $p > .1$), between dative NP and animate second NP ($t(31) = 1.21$, $p > .1$), or between dative NP and inanimate second NP ($t(31) = 1.68$, $p > .1$). Three additional experiments and 29 unrelated sentences with mild to severe grammaticality problems served as filler items, amounting to 159 test sentences in the whole experiment.

Procedure. All participants were tested individually in the lab. The ME procedure closely followed the description of the ME method in Bard, Robertson & Sorace (1996) and consisted of a customization phase, where participants were acquainted with the method by judging the length of lines and the acceptability of ten training sentences, and the experimental phase.

In each phase, participants first assigned a numerical value to the reference stimulus (either a line or a sentence). Afterwards, the experimental stimuli were displayed one by one, and participants judged each stimulus relative to the reference stimulus, which remained visible throughout the experiment. The reference sentence (12), almost literally taken from Keller (2000: sentence B.18 on page 377), is a sentence with non-canonical word order.

(12) Ich glaube, dass den Bericht der Chef in seinem Büro gelesen hat.
I believe that the.ACC report the.NOM boss in his office read has
'I believe that the boss read the report in his office.'

3.1.2 Results

All statistical analyses were conducted using the statistics software R, Version 3.1.3 (R Development Core Team 2015).

Preprocessing. The raw output of the magnitude estimation procedure was treated as follows: First, every rating was normalized by dividing it by the value assigned to the reference sentence to put all judgments on the same scale. Second, these normalized values were log-transformed, as is common practice with ME data and ensures the data to be approximately normally distributed. Third, to reduce individual differences, we performed a z-transformation by subject. The same procedure was used in Hofmeister et al. (2013). Additionally, we conducted the following procedure to identify influential data points: Using the r-package influence.ME (Nieuwenhuis, te Grotenhuis & Pelzer 2012), we calculated Cook's Distance for individual data points. For the reported means (model summaries and graphs), we excluded all data points exceeding a Cook's distance of 4/(Number of data points), accounting for 4% of the data. Excluding influential data points is justified by our incentive to compare estimates across experiments. In addition, we conducted an analysis including influential data points and report any difference in significance. We used the same procedure for the results from the following experiments. Table 3 shows mean z-transformed log ratios by condition (influential data points removed).

Analysis. We analyzed the results with linear mixed effect models using the R package lme4 (Bates et al. 2015). We entered Order, Voice, Animacy and all interactions as fixed effects into the model, using effect coding (i.e., the intercept represents the unweighted grand mean, fixed effects compare factor levels to each other). In addition, we included random effects for items and subjects with maximal random slopes supported by the data, largely following the strategy proposed in Bates et al. (2015). In Table 4, we report

Table 3: Mean z-transformed log ratios in Experiments 1 and 2.

Experiment 1	**animate**		**inanimate**	
	active	**passive**	**active**	**passive**
dative 1st	0.110	0.049	0.362	0.353
dative 2nd	−0.023	0.148	0.073	0.264
Experiment 2	**Def align**		**Def noAlign**	
	active	**passive**	**active**	**passive**
dative 1st	0.523	0.532	0.367	0.386
dative 2nd	−0.273	−0.124	0.134	0.264

Table 4: Linear mixed model fit by maximum likelihood for Experiment 1, annotated with p-values from likelihood ratio tests.

	Coefficiant	**Std. Error**	**t value**	**p (LRT)**
(Intercept)	0.167	0.028	5.959	
Animacy	−0.096	0.018	−5.218	<.001***
Order	0.051	0.013	3.841	<.01**
Voice	−0.037	0.012	−3.072	.05·
Animacy:Order	−0.044	0.013	−3.449	<.01**
Animacy:Voice	0.009	0.010	0.904	.37
Order:Voice	0.054	0.011	4.943	<.001***
Animacy:Order:Voice	00.004	0.010	0.384	.70

Formula: zlogratio ~ Order * Voice * Animacy + (Order * Voice + Animacy|subject) + (Order * Animacy|sentence).

the full model summary as well as likelihood ratio tests, which assess the contribution of single factors or interactions. Where necessary, we report pairwise comparisons (Tukey's test).

Results and discussion. Order had a significant effect on judgments ($\chi(1) = 9.676$, $p < .01$), with dative 1st resulting in higher judgments than dative 2nd on average, in line with the allegedly canonical order in German. The main effect of Animacy ($\chi(1) = 20.336$, $p < .001$) indicates overall lower acceptability of sentences containing an animate theme. This effect is unexpected with regard to our constraint set and we suspect it to be related to plausibility issues. Interestingly, there was also an interaction between Order and Animacy ($\chi(1) = 10.337$, $p < .01$). In Figure 1, we can see that the significant preference for dative 1st orders that we see for inanimate themes is reduced for animate themes in the active voice and even reversed in the passive voice[5]. We conclude from this that ANI is an active constraint in German, but not the only one causing the preference for the canonical order. The persisting preference for dative 1st for animate themes in active sentences must be due to either DAT or REC. We also found a marginal main effect of Voice ($\chi(1) = 3.742$, $p = .05$) indicating higher acceptability for passive sentences on average and an interaction between Voice and Order ($\chi(1) = 21.679$, $p < .001$). While there is a preference for dative 1st in active sentences, in passive this preference is reduced for inanimate themes, and even reversed for animate themes. This effect indicates an influence of the case constraint NOM.

Now that we have established an effect of ANI and NOM as well as an influence of either DAT or REC, let us turn to their relative importance under a strict domination view as entertained by OT. In order to do so, Table 5 presents a collective tableau listing the constraint violations of all sentences tested in Experiment 1. Candidate sets are separated by a solid line and are in general defined similar to Müller (1999), i.e., consisting of the different orders for the same lexical content. In this view, passive and active sentences do not compete, as they differ with regard to the realization of the subject.

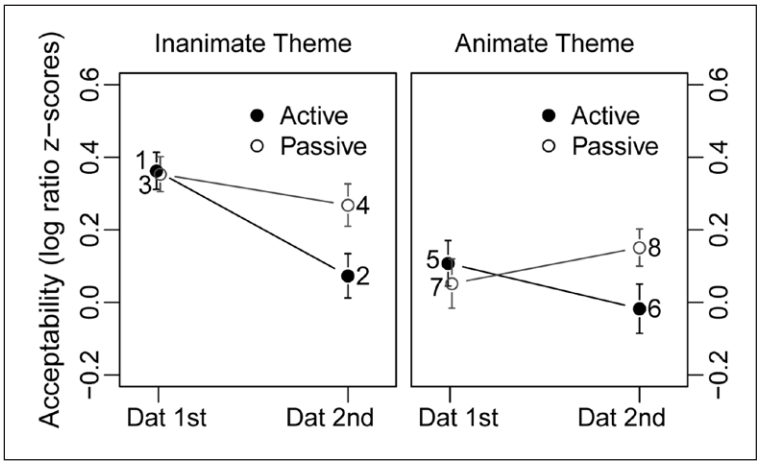

Figure 1: Z-transformed log ratios of magnitude estimation judgments by condition for Experiment 1. Error bars represent 95% confidence intervals.

[5] For animate themes, the preference for dative 1st in active was significant, while the numerical preference for dative 2nd in passive did not reach significance ($p = .15$).

Table 5: Violation profiles for sentences from Experiment 1; candidate sets are separated by solid lines.

Condition	Candidates	ANI	NOM	DAT	REC
1	$IO_{ANI\text{-}DAT\text{-}REC}$ $DO_{INANI\text{-}ACC\text{-}THEME}$				
2	$DO_{INANI\text{-}ACC\text{-}THEME}$ $IO_{ANI\text{-}DAT\text{-}REC}$	*		*	*
3	$IO_{ANI\text{-}DAT\text{-}REC}$ $S_{INANI\text{-}NOM\text{-}THEME}$		*		
4	$S_{INANI\text{-}NOM\text{-}THEME}$ $IO_{ANI\text{-}DAT\text{-}REC}$	*			*
5	$IO_{ANI\text{-}DAT\text{-}REC}$ $DO_{ANI\text{-}ACC\text{-}THEME}$				
6	$DO_{ANI\text{-}ACC\text{-}THEME}$ $IO_{ANI\text{-}DAT\text{-}REC}$			*	*
7	$IO_{ANI\text{-}DAT\text{-}REC}$ $S_{ANI\text{-}NOM\text{-}THEME}$		*		
8	$S_{ANI\text{-}NOM\text{-}THEME}$ $IO_{ANI\text{-}DAT\text{-}REC}$				*

The relative ranking of NOM and REC is a clear case: Since in the animate theme passive conditions (7 and 8 in Table 5) these constraints, and only those, directly compete, we can conclude that the preference for dative 2nd here indicates the ranking NOM > REC.

We can further conclude that ANI is ranked above NOM as the violation of ANI is the only differentiation between the passive conditions with animate themes on the one hand (7 and 8 in Table 5) and inanimate themes on the other (3 and 4 in Table 5) and the order preference switches here. We thus arrive at the partial ranking in (13).

(13) Partial constraint ranking established by Experiment 1
ANI > NOM (, DAT) > REC

Note that we do not have direct evidence of the position of DAT in this ranking, which could be anywhere from above ANI to below REC, as all violations coincide with REC violations and DAT is never in direct competition with one of the other constraints. If we assume *weighted* constraints, the most likely position would be close to NOM, as the distance between dative 1st and dative 2nd for animate themes seems to be comparable in active (DAT violation) and passive (NOM violation).

3.2 Experiment 2

In this experiment, we aim to replicate the effects of Order and Voice on acceptability that we found in Experiment 1. Furthermore, we want to integrate DEF in the preliminary hierarchy. In the hierarchy of Müller (1999) given in (7), DEF is higher-ranked than ANI. This ranking derives from the claim that sentence (14a) is unmarked and sentence (14b) marked. That is, when DEF and ANI are in conflict, the order that respects DEF but violates ANI wins the competition.

(14) From Müller (1999)

a. daß der Verkäufer den Wein einem Kunden empfahl
that the.NOM salesman the.ACC wine a.DAT costumer recommended
'that the salesman recommended the wine to a costumer'

b. ?daß der Verkäufer einem Kunden den Wein empfahl
that the.NOM salesman a.DAT costumer the.ACC wine recommended

In order to keep the complexity moderate, we only used inanimate theme sentences here and manipulated three factors. Order and Voice were parallel to Experiment 1. In contrast to Experiment 1, one of the two full NPs was indefinite. We manipulated whether definiteness aligned with animacy or not. When definiteness and animacy are aligned (i.e., the animate NP is definite and the inanimate NP indefinite), ANI and DEF make the

same predictions in favor of the dative 1st order. When animacy and definiteness are not aligned (i.e., the inanimate NP is definite and the animate NP is indefinite), ANI and DEF pull in different directions: ANI predicts dative 1st to be preferred, while DEF makes the opposite prediction.

(15) Paul hat berichtet, ('Paul reported,')

a. **animate > inanimate**
dass man **dem** Zeugen **ein** Foto gezeigt hat.
that one.NOM **the**.DAT witness **a**.ACC picture shown has
'that someone showed the witness a picture.'

b. **animate > inanimate**
dass man **einem** Zeugen **das** Foto gezeigt hat.
that one.NOM **a**.DAT witness **the**.ACC picture shown has
'that someone showed a witness the picture.'

The preference in the conflicting condition will therefore allow us to establish a ranking between ANI and DEF and evaluate Müller's proposal that DEF is higher ranked than ANI.

3.2.1 Method

Participants. Seventy-two students from the University of Konstanz took part in the study. All were native speakers of German. The experiment took about 20–25 minutes to complete. For the whole experimental session, which included an additional experiment and took approximately one hour, they received either partial course credit or a compensation of 6 Euro.

Materials and procedure. Forty experimental items were created. The general form of the experimental sentences was similar to Experiment 1. In contrast to Experiment 1, the theme NP was always inanimate and one of the NPs was indefinite. In addition to Order and Voice, we manipulated alignment of definiteness (Def: align, noAlign). In the align condition, the recipient NP was definite and the theme NP indefinite. In the noAlign condition, in contrast, the theme NP was definite and the recipient NP indefinite, thereby creating a conflict between ANI and DEF. A full example item can be found in Table 6. The procedure was the same as in Experiment 1. 30 sentences from an unrelated experiment and 18 unrelated sentences with mild to severe grammaticality problems served as filler items, amounting to 88 test sentences in the whole experiment.

3.2.2 Results and discussion

Table 3 shows mean z-transformed log ratios by condition and Table 7 the summary of the linear mixed effect model. We replicate the effects of Order ($\chi(1) = 79.61$, $p < .001$) and the interaction between Order and Voice ($\chi(1) = 5.88$, $p < .05$) from Experiment 1. In addition, we see a main effect of Voice ($\chi(1) = 6.96$, $p < .01$) due to higher judgments in passive, on average. Similar to the inanimate theme conditions in Experiment 1, the interaction was not a crossing one. A general preference for dative 1st orders was modulated by Voice, corroborating the partial ranking ANI > NOM established by Experiment 1. Pairwise comparisons showed a significant difference between passive and active sentences only for the dative 2nd align conditions.

The main effect of DEF was significant ($\chi(1) = 16.95$, $p < .001$) due to higher judgments for noAlign conditions than align conditions. The effect seems to be driven by the conditions where both DEF and ANI are violated (2 and 4 in Table 8 and Figure 2), which is perceived as particularly bad. The interaction between Order and Def was highly significant ($\chi(1) = 59.369$, $p < .001$). As is apparent from Figure 2 and confirmed by pairwise comparisons, dative 1st conditions are rated lower in noAlign, where ANI

Table 6: Example item from Experiment 2.

Paul hat berichtet, ('Paul reported',)				
Cond	**Definiteness**	**Voice**	**Order**	**Stimulus continuation**
1	align	ACT	DAT 1ST	dass **man dem** Zeugen **ein** Foto gezeigt **hat**.
2			DAT 2ND	dass **man ein** Foto **dem** Zeugen gezeigt **hat**.
				that one a.ACC photo the.DAT witness shown has
				'that someone has shown a photo to the witness.'
3		PASS	DAT 1ST	dass **dem** Zeugen **ein** Foto gezeigt **wurde**.
4			DAT 2ND	dass **ein** Foto **dem** Zeugen gezeigt **wurde**.
				that a.ACC photo the.DAT witness shown was
				'that a photo was shown to the witness.'
5	noAlign	ACT	DAT 1ST	dass **man einem** Zeugen **das** Foto gezeigt **hat**.
6			DAT 2ND	dass **man das** Foto **einem** Zeugen gezeigt **hat**.
				that one the.ACC photo a.DAT witness shown has
				'that someone has shown the photo to a witness.'
7		PASS	DAT 1ST	dass **einem** Zeugen **das** Foto gezeigt **wurde**.
8			DAT 2ND	dass **das** Foto **einem** Zeugen gezeigt **wurde**.
				that the.ACC photo a.DAT witness shown was
				'that the photo was shown to a witness.'

Table 7: Linear mixed model fit by maximum likelihood for Experiment 2, annotated with p-values from likelihood-ratio tests.

	Coefficiant	**Std. Error**	**t value**	**p (LRT)**
(Intercept)	0.221	0.036	6.136	
Definiteness	0.058	0.013	4.443	<.001***
Order	0.229	0.019	12.281	<.001***
Voice	−0.037	0.0123	−2.995	<.01**
Definiteness:Order	−0.137	0.013	−10.259	<.001***
Definiteness:Voice	0.002	0.009	0.250	.80
Order:Voice	0.031	0.009	3.370	<.01**
Definiteness:Order:Voice	−0.005	0.009	−0.544	.59

Formula: zlogratio ~ Order * Voice * Definiteness + (1 + Order + Order:Definiteness | subject) + (1 + Definiteness | sentence) + (0 + Order + Voice + Order:Voice | sentence).

Table 8: Violation profiles for sentences from Experiment 2; candidate sets are separated by solid lines.

Condition	**Candidates**	**ANI**	**DEF**	**NOM**	**DAT**
1	$IO_{ANI\text{-}DAT\text{-}DEF}$ $DO_{INANI\text{-}ACC\text{-}INDEF}$				
2	$DO_{INANI\text{-}ACC\text{-}INDEF}$ $IO_{ANI\text{-}DAT\text{-}DEF}$	*	*		*
5	$IO_{ANI\text{-}DAT\text{-}INDEF}$ $DO_{INANI\text{-}ACC\text{-}DEF}$		*		
6	$DO_{INANI\text{-}ACC\text{-}DEF}$ $IO_{ANI\text{-}DAT\text{-}INDEF}$	*			*
3	$IO_{ANI\text{-}DAT\text{-}DEF}$ $S_{INANI\text{-}NOM\text{-}INDEF}$			*	
4	$S_{INANI\text{-}NOM\text{-}INDEF}$ $IO_{ANI\text{-}DAT\text{-}DEF}$	*	*		
7	$IO_{ANI\text{-}DAT\text{-}INDEF}$ $S_{INANI\text{-}NOM\text{-}DEF}$		*	*	
8	$S_{INANI\text{-}NOM\text{-}DEF}$ $IO_{ANI\text{-}DAT\text{-}INDEF}$	*			

and DEF are in conflict (compare 5 to 1 and 7 to 3), whereas dative 2nd conditions are rated better in noAlign (6 and 8). The preference does not switch, however: Pairwise comparisons reveal a significant difference between noAlign-dative 1st and noAlign-dative 2nd in active sentences; in passive sentences this difference is marginal.

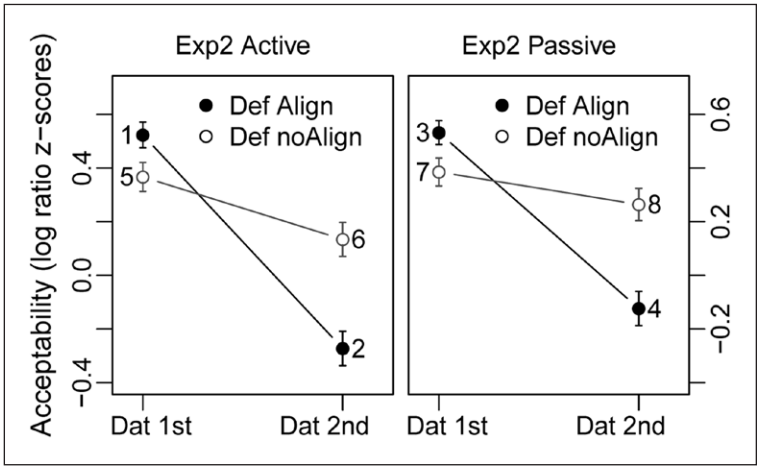

Figure 2: Z-transformed log ratios of magnitude estimation judgments by condition for Experiment 2. Error bars represent 95% confidence intervals.

In sum, the judgment given in Müller (1999) for the sentence pair in (14) could not be confirmed, and contrary to the ranking of these constraints in Müller's constraint hierarchy we must conclude that ANI is higher ranked than DEF.

(16) Partial constraint ranking established by Experiments 1 and 2
ANI > DEF, NOM, DAT > REC

3.3 Experiment 3

In Experiment 3, we turn our attention to the order between subject and direct object in transitive structures. As has often been noted, German has a strong preference for subject-object order, but this preference does not hold across the board (Lenerz 1977, and much subsequent work). For example, as confirmed by Experiments 1 and 2, passive sentences with an animate dative object and an inanimate theme subject show a preference for object-subject order. The otherwise strong subject-object preference can therefore not be attributed to a purely syntactic constraint requiring the subject to precede all other arguments. A prototypical sentence with preferred OS order has a subject that is inanimate and not an agent. In order to test whether the thematic role constraint AG or the animacy constraint ANI is the driving factor for the canonical SO order observed with standard transitive verbs, Experiment 3 contrasts AG with ANI by investigating agentive verbs that take animate as well as inanimate subjects.[6] In addition, the influence of DEF is taken into account by varying the definiteness of the agentive subject. The object in all sentences is a definite animate NP. The sentences in (17) illustrate the contrasts that are tested in Experiment 3.

(17) Mir ist erzählt worden, ('I was told')
a. **+definite +animate subject**
dass der Spekulant den Winzer ruiniert hat.
that the.NOM venturer the.ACC wine-grower ruined has
'that the venturer ruined the wine grower.'

[6] The term "agent" is sometimes reserved for intentional causers of an action. Inanimate causers would then be just causers. We use the term "agent" in the more general sense subsuming both intentional and non-intentional causers.

b. **+definite –animate subject**
dass das Feuer den Winzer ruiniert hat.
that the.NOM fire the.ACC wine-grower ruined has
'that the fire ruined the wine grower.'

c. **–definite +animate subject**
dass ein Spekulant den Winzer ruiniert hat.
that a.NOM venturer the.ACC wine-grower ruined has
'that a venturer ruined the wine grower.'

d. **–definite –animate subject**
dass ein Feuer den Winzer ruiniert hat.
that a.NOM fire the.ACC wine-grower ruined has
'that a fire ruined the wine grower.'

The experiment is organized in three subexperiments. Subexperiment 3a contrasts AG with ANI to test whether ANI is in general more influential than thematic roles or whether the highest thematic role takes precedence. This is achieved by varying the animacy of the agentive subject, (17a) vs (17b), while definiteness is kept constant. If we find a preference for SO in general, we can conclude that AG is ranked higher than ANI, while the opposite holds if OS is preferred for inanimate subjects. Subexperiment 3b contrasts AG with DEF by varying the definiteness of an animate agentive subject, (17a) vs (17c). In Subexperiment 3c, we combine both violations (ANI and DEF) to see whether violations accumulate and gang up against a presumably higher ranked AG by comparing a definite inanimate subject to an indefinite inanimate subject, i.e., (17b) vs (17d).

3.3.1 Method

Participants and procedure. 108 students from the Goethe University Frankfurt took part for partial course credit. All were native speakers of German. The subexperiments were conducted between participants. In Subexperiments 3a and 3b, we tested 28 participants and in Subexperiment 3c 52 participants. The procedure was the same as in Experiments 1 and 2.

Materials. Twenty-four experimental items were created. An experimental sentence started with a short introductory main clause, followed by a transitive complement clause. Three factors were manipulated, animacy of the subject (SubjAni: animate, inanimate), definiteness of the subject (SubjDef: definite, indefinite), and word order (Order: SO, OS). In Subexperiment 3a, animacy was varied within the level definite of definiteness, in Subexperiment 3b, definiteness was varied within the level animate of animacy and finally in Subexperiment 3c, definiteness was varied within the level inanimate. A full example item can be found in Table 9.

Three additional experiments and 19 unrelated sentences with mild to severe grammaticality problems served as filler items, amounting to 115 test sentences in the whole experiment. The additional experiments varied between subexperiments.[7]

3.3.2 Results and discussion

We conducted separate analyses for the three subexperiments. Table 10 shows mean z-transformed log ratios by condition and Table 11 the summary of the linear mixed effect model. Prior to analysis, we had to exclude one item from Subexperiments 3a and 3c because of a typo in the materials. In Subexperiment 3b, we excluded one participant, who gave the same response for all experimental and filler items.

[7] Because of the different fillers, z-scores vary considerably between experiments. In plotting the results, we therefore revert to untransformed log ratios, while inferential analyses are still conducted on z-transformed data.

Table 9: Example item from Experiment 3.

Mir ist erzählt worden, ('I was told')				
Exp	**Cond**	**Subject**	**Order**	**Stimulus continuation**
3a	1	INANI DEF	SO	dass **das** **Feuer** **den** Winzer ruiniert hat.
	2		OS	dass **den** Winzer **das** **Feuer** ruiniert hat.
				that the.ACC wine-grower the.NOM fire ruined has
				'that the fire ruined the wine-grower.'
	3	ANI DEF	SO	dass **der** **Spekulant** **den** Winzer ruiniert hat.
	4		OS	dass **den** Winzer **der** **Spekulant** ruiniert hat.
				that the.ACC wine-grower the.NOM speculator ruined has
				'that the speculator ruined the wine-grower.'
3b	1	ANI INDEF	SO	dass **ein** **Spekulant** **den** Winzer ruiniert hat.
	2		OS	dass **den** Winzer **ein** **Spekulant** ruiniert hat.
				that the.ACC wine-grower a.NOM speculator ruined has
				'that the speculator ruined the wine-grower.'
	3	ANI DEF	SO	dass **der** **Spekulant** **den** Winzer ruiniert hat.
	4		OS	dass **den** Winzer **der** **Spekulant** ruiniert hat.
				that the.ACC wine-grower the.NOM speculator ruined has
				'that the speculator ruined the wine-grower.'
3c	1	INANI INDEF	SO	dass **ein** **Feuer** **den** Winzer ruiniert hat.
	2		OS	dass **den** Winzer **ein** **Feuer** ruiniert hat.
				that the.ACC wine-grower a.NOM fire ruined has
				'that the fire ruined the wine-grower.'
	3	INANI DEF	SO	dass **das** **Feuer** **den** Winzer ruiniert hat.
	4		OS	dass **den** Winzer **das** **Feuer** ruiniert hat.
				that the.ACC wine-grower the.NOM fire ruined has
				'that the fire ruined the wine-grower.'

Table 10: Mean z-transformed log ratios in Experiment 3.

	Experiment 3a		**Experiment 3b**		**Experiment 3c**	
	INANI DEF	**ANI DEF**	**ANI INDEF**	**ANI DEF**	**INANI INDEF**	**INANI DEF**
SO	0.641	0.663	0.418	0.432	0.795	0.773
OS	−0.010	−0.390	−0.282	−0.653	0.536	0.273

Order had a significant effect on acceptability judgments in all three subexperiments ($\chi(1) > 34$, $p < .001$) in that SO sentences were rated significantly better than OS sentences. In Subexperiment 3a, SubjAni had a marginal main effect, too ($\chi(1) = 3.743$, $p = .05$), with inanimate subjects resulting in higher acceptability on average. There was a significant interaction between the two factors ($\chi(1) = 17.116$, $p < .001$). As apparent from the first panel in Figure 3, this interaction was not reversing the SO preference for inanimate subjects, but reducing the difference between SO and OS. Pairwise comparisons revealed that there was no difference between animate and inanimate subject within SO, but a significant difference within OS. Similarly, in Subexperiment 3b, there was a main effect of SubjDef ($\chi(1) = 25.976$, $p < .001$) and an interaction with Order ($\chi(1) = 18.177$, $p < .001$), with pairwise comparisons identifying no difference between the two SO conditions, but a significant difference within OS. The pattern is repeated in Subexperiment 3c, with a main effect of Subj ($\chi(1) = 4.604$,

Table 11: Linear mixed model fit by maximum likelihood for Experiment 3, annotated with p-values from likelihood ratio tests.

		Coefficiant	Std. Error	t value	p (LRT)
Experiment 3a	(Intercept)	0.248	0.033	7.558	
	Order	0.399	0.029	14.010	<.001***
	SubjAni	0.100	0.021	4.898	=.05.
	Order:SubjAni	−0.115	0.023	−4.936	<.001***
Experiment 3b	(Intercept)	−0.019	0.046	−0.422	
	Order	0.444	0.053	8.401	<.001***
	SubjDef	−0.091	0.017	−5.230	<.001***
	Order:SubjDef	0.098	0.019	5.035	<.001***
Experiment 3c	(Intercept)	0.628	0.020	31.69	
	Order	0.159	0.019	8.49	<.001***
	Subj	−0.056	0.015	−3.90	<.05*
	Order:Subj	0.046	0.013	3.59	<.001***

Formulae:
a: zlogratio ~ Order * SubjAni + (Order + Order:SubjAni | subject) + (SubjAni + Order:SubjAni | sentence).
b: zlogratio ~ Order * SubjDef + (Order | subject) + (Order | sentence).
c: zlogratio ~ Order * Subj + (Order | subject) + (1 | sentence) + (0 + Order | sentence) + (0 + Subj | sentence).

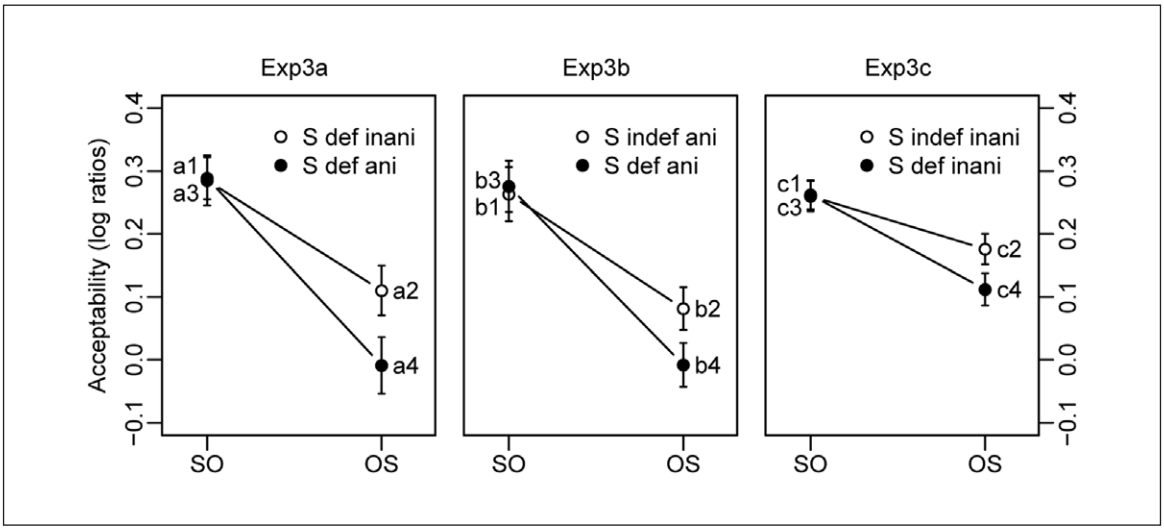

Figure 3: Log ratios of magnitude estimation judgments by condition for Experiment 3. Error bars represent 95% confidence intervals, see footnote 7.

Table 12: Violation profiles for sentences from Experiment 3; candidate sets are separated by solid lines.

Condition	Candidates	AG	ANI	DEF	NOM	REC	PER
a1, c3	$S_{\text{INANI-DEF-NOM-AG}}$ $O_{\text{ANI-DEF-ACC-THEME}}$		*				*
a2, c4	$O_{\text{ANI-DEF-ACC-THEME}}$ $S_{\text{INANI-DEF-NOM-AG}}$	*			*		
a3, b3	$S_{\text{ANI-DEF-NOM-AG}}$ $O_{\text{ANI-DEF-ACC-THEME}}$						*
a4, b4	$O_{\text{ANI-DEF-ACC-THEME}}$ $S_{\text{ANI-DEF-NOM-AG}}$	*			*		
b1	$S_{\text{ANI-INDEF-NOM-AG}}$ $O_{\text{ANI-DEF-ACC-THEME}}$			*			*
b2	$O_{\text{ANI-DEF-ACC-THEME}}$ $S_{\text{ANI-INDEF-NOM-AG}}$	*			*		
c1	$S_{\text{INANI-INDEF-NOM-AG}}$ $O_{\text{ANI-DEF-ACC-THEME}}$		*	*			*
c2	$O_{\text{ANI-DEF-ACC-THEME}}$ $S_{\text{INANI-INDEF-NOM-AG}}$	*			*		

$p < .05$), and an interaction ($\chi(1) = 11.663$, $p < .001$) with no difference between the SO conditions, but a difference within OS. In Figure 3, looking at the raw log ratios, we also observe that in Subexperiment 3c, OS sentences were rated better than in the previous two subexperiments with the two identical conditions inanimate-definite closely corresponding between Subexperiments 3a and 3c.

With regard to our constraint hierarchy, the results together with the violation profiles in Table 12 suggest a high ranked AG dominating both ANI and DEF:

(18) Partial constraint ranking established by Experiments 1, 2 and 3
AG > ANI > DEF, NOM, DAT > REC

3.4 Experiment 4: NOM as NOM > ACC

The preference for dative-before-nominative orders in the passive conditions of Experiments 1 and 2 motivated a relatively low ranking of NOM in the constraint hierarchy in (18). Following Vogel & Steinbach (1995), Heck (2000) argues that subject and accusative object are generated in a fixed order by OT's generator GEN, whereas dative arguments can be rather freely attached inside the clause. In this view, we would have to split NOM into the following two constraints, where (19a) is the highest ranked constraint, whereas (19b) is relatively low ranked.

(19) **Syntactic constraints**
a. NOM_{ACC}: nominative > accusative
b. NOM_{DAT}: nominative > dative

In Experiment 3, we argued that the preference for SO is rooted in the lexical-semantic constraint AG, rather than NOM, as NOM was found to be low-ranked for orders including dative objects. To tease apart the influence of AG and NOM_{ACC}, which were confounded in Experiment 3, we conducted an experiment with accusative object experiencer psych verbs. In addition to word order, we manipulated the definiteness of the accusative object:

(20) Der Bruder hat gehört, ('The brother has heard,')
a. **stimulus-inanimate > experiencer-animate**
dass das/ein Geschenk den Vater getröstet hat.
that the/a.NOM present the.ACC father comforted has.
b. **experiencer-animate > stimulus-inanimate**
dass den Vater das/ein Geschenk getröstet hat.
that the.ACC father the/a.NOM present comforted has.

Thematic roles and animacy favor OS order here, while NOM_{ACC} supports SO order. If SO were preferred, we would have to discard the constraint ordering ANI > NOM in favor of NOM_{ACC} > ANI > NOM_{DAT}.

3.4.1 Method

Participants and procedure. 20 students from the Goethe University Frankfurt took part for partial course credit or received a compensation of 8 Euro. All were native speakers of German. The procedure was the same as in Experiments 1, 2, and 3.

Materials. 24 experimental items were created. An experimental sentence started with a short introductory main clause, followed by a transitive complement clause with an object experiencer psych verb. Definiteness of the subject (SubjDef: definite, indefinite)

and word order (Order: SO, OS) were manipulated. A full example item can be found in Table 13.

The experimental items were mixed with 32 items from an unrelated experiment and 29 fillers with varying degrees of acceptability.

3.4.2 Results

Table 14 shows the summary of the linear mixed effect model. In the inferential analysis, we found a main effect of Order ($\chi(1) = 8.86$, $p < .01$). As can be observed in Figure 4, the main effect originated in a preference for SO orders, thus confirming that with accusative objects, case takes precedence over thematic role and animacy. The main effect of SubjDef was also significant with higher ratings for definite subjects ($\chi(1) = 6.31$, $p < .05$). The interaction was not significant ($\chi(1) < 1$). While the general dispreference for indefinite subjects is unexpected, especially because OS sentences satisfied the DEF constraint, it is possible that sentences with indefinite NPs were in general perceived as somewhat less plausible by the participants.

By splitting NOM in NOM_{ACC} and NOM_{DAT}, we arrive at the following ranking. Constraints in parentheses cannot be placed conclusively given the evidence obtained in our experiments.

(21) Constraint ranking established by Experiments 1, 2, 3 and 4
$\text{NOM}_{\text{ACC}} > (\text{AG}) > \text{ANI} > \text{DEF}, \text{NOM}_{\text{DAT}}, (\text{DAT}) > \text{REC}$

4 Gradient acceptability and weighted constraints

When assessing the constraints from the literature with respect to our experimental findings, we have so far only considered strict domination between constraints. As pointed out in section 2.3, weighted constraints provide an alternative to strict domination. In this section, we use the experimental results to evaluate the explanatory power of weighted constraint models. To this end, we have already formulated the Uniform Penalty

Table 13: Example item from Experiment 4.

Der Bruder hat gehört, ('The brother has heard')			
Cond	**SubjDef**	**Order**	**Stimulus continuation**
1	definite	SO	dass **das Geschenk den Vater** getröstet hat.
2		OS	dass **den Vater das Geschenk** getröstet hat.
			that the.ACC father the.NOM present comforted has
			'that the present has comforted the father.'
3	indefinite	SO	dass **ein Geschenk den Vater** getröstet hat.
4		OS	dass **den Vater ein Geschenk** getröstet hat.
			that the.ACC father a.NOM present comforted has
			'that a present has comforted the father.'

Table 14: Linear mixed model fit by maximum likelihood for Experiment 4, annotated with p-values from likelihood ratio tests.

	Coefficiant	**Std. Error**	**t value**	**p (LRT)**
(Intercept)	0.295	0.035	8.408	
Order	0.155	0.047	3.288	<.05*
SubjDef	0.038	0.015	2.519	<.01**
Order:SubjDef	0.005	0.015	0.359	=.72

Formula: zlogratio ~ Order * SubjDef + (Order|subject) + (Order|sentence).

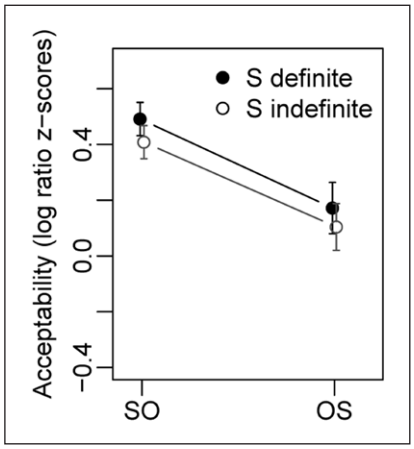

Figure 4: Z-transformed log ratios of magnitude estimation judgments by condition for Experiment 4. Error bars represent 95% confidence intervals.

Hypothesis in section 2.3, according to which each constraint violation leads to a decrease in acceptability proportional to the weight of the constraint. In the strong version of the Uniform Penalty Hypothesis, this should hold even for comparisons across candidate sets, namely when we compare sentences that are optimal within their candidate set, that is, the most acceptable linearization of a given content. In the weak version, the Uniform Penalty Hypothesis only predicts that constraint violations are reflected in the difference between candidates from the same set. Whenever we change a lexical item, we therefore lose the possibility to compare directly. Both versions predict stable constraint weights across candidate sets and experiments.

In a first step, we briefly review the observed patterns descriptively with regard to the strong Uniform Penalty Hypothesis. In a second step, we fit statistical models to the results, this time encoding constraint violations directly and comparing the strong and weak versions of the Uniform Penalty Hypothesis to each other.

4.1 Constraint violations by the optimal candidate

We begin our review with Experiment 3, which is at odds with the strong version of the Uniform Penalty Hypothesis. Let us assume the following partial ranking of the weights: $w_{NOM_{ACC}} > w_{AG} > w_{ANI} > w_{DEF}$. While these weights would clearly lead to a general SO preference, as observed, weighted constraint models make an additional prediction: For inanimate or indefinite subjects, SO should score lower than for animate definite subjects, as ANI or DEF are violated. The OS orders, on the other hand, are not expected to differ in acceptability, as they violate the same constraints: NOM_{ACC} and AG, but not ANI or DEF. What we observed is the opposite. There is no difference in acceptability between SO sentences, regardless of whether they violate ANI, DEF, or both. Surprisingly, however, we see a difference between OS sentences: The more constraints their SO counterpart violates, the better the OS sentence is rated. This pattern is consistent with the weak version of the Uniform Penalty Hypothesis, as the difference between the optimal candidate and its competitor is smaller if the optimal candidate violates a constraint its competitor does not violate.

There is, however, an alternative explanation which also captures the observed pattern. So far, we have only considered positive weights that lead to penalties. Instead, we could argue to include negative weights that lead to a *reward* for a structure that satisfies the constraint *non-vacuously*, i.e., a sentence with arguments of different animacy or definiteness levels (see Pater 2009: 1008, for a discussion of positive and negative weights). In Experiment 3, this would apply to sentences in which OS order accommodates

ANI or DEF. We reformulate the two constraints ANI and DEF in (22) such that they only apply in relevant contexts. In our previous definition, ANI was violated by structures with inanimate > animate, but not by structures with either animate > inanimate or animate > animate. The reward version should apply only to animate > inanimate orders, but not to animate > animate orders

(22) a. ANI-reward : reward any structure in which an animate NP precedes an inanimate NP
b. DEF-reward : reward any structure in which a definite NP precedes an indefinite NP

This account is challenged by the results of Experiment 1. Here, we see a similar pattern in the conditions with an inanimate theme: the two optimal candidates score equally good, but there is a significant difference between the two dative 2nd sentences with passive sentences scoring better than active sentences. When only considering this part of the experiment, we might come to the conclusion that NOM_{DAT} is not an active constraint and therefore remains without penalty, while the difference between dative 2nd orders could be attributed to DAT, which is violated only in the active voice. This argumentation breaks down, however, when we consider the results from the animate theme conditions, where we found clear evidence in favor of NOM_{DAT}. In this case, we cannot argue for a satisfaction reward for NOM_{DAT}, as this would lead to a difference between the dative 1st orders for inanimate themes, because it is satisfied only by the active sentence.

The observed patterns seem to be consistent with the weak Uniform Penalty Hypothesis, but could be captured to a fair degree by a model including negative and positive weights, while the strong Uniform Penalty Hypothesis seems to fare worst. On the other hand, the strong Uniform Penalty Hypothesis is the most simple model that provides us with the most explicit predictions, that is, a predicted acceptability value for each sentence, which can be compared against the empirical acceptability measure. The model derived from the weak Uniform Penalty Hypothesis gives us numerical values for the expected acceptability differences only. In such a model, all optimal candidates could, e.g., score equally good or the differences between optimal candidates could reflect their complexity, plausibility or other factors which are often associated with performance. In comparison with Müller (1999), we still get more fine-grained predictions, as constraint violations should numerically have the same effect in different contexts, although only affecting the difference between candidates from the same candidate set. In this view, there is still a penalty associated with each constraint violation even for optimal candidates, but this penalty does not necessarily affect the absolute acceptability measure, but rather the magnitude of the lead of the optimal candidate over its competitors.

4.2 Modeling weighted constraints

With the mixed results reported above, it seems appropriate to assess the explanatory benefit of weighted constraint models compared to the standard view. While the latter usually focuses on predicting *preferences*, weighted constraint models aim at predicting *quantitative differences*, in particular acceptability differences between suboptimal candidates. In our informal review above, we have already noted that a simple additive model with only positive constraint weights is not able to capture the data. Major adjustments as the ones outlined above are therefore necessary. In order to determine whether these changes, and therefore a numerical model, is motivated by the data, we will build regression models incorporating the different model specifications. There are three aspects of explanatory adequacy that we will assess. The first one is to check whether the qualitative pattern, i.e., the observed preferences, are captured equally well as in the OT model of Müller

(1999) with the modified constraint hierarchy in (21). The second one is to check whether constraint weights remain stable across experiments – a particular well-suited test case for this is to compare the weights of ANI and DEF as derived from Experiments 1 and 2 to their weights as derived from Experiment 3. In the first case, they affect order preferences between theme and recipient in double object constructions, while in the latter case, they affect the order between subject and direct object. Lastly, we can also compare the model fit between the different models directly.

4.2.1 Method

The preprocessing of the results for this analysis differs from the analyses presented above, as we are now not so much interested in the significance of our experimental factors, but rather in the specific coefficient estimates of the constraint weights. We therefore use log ratios rather than their z-transforms, as this transformation depends on the fillers included in the experiment and makes it therefore impossible to compare across experiments. Furthermore, we want to compare the model fits, which is more convenient when using multiple regression, where adjusted R^2 can be calculated straightforwardly. Lastly, we use only condition means for this analysis to exclude some of the variation between experiments that we cannot model, in particular variation at the level of individual participants.

We present models for three data sets. The first one includes all data points from all experiments, the second one combines the data from Experiments 1 and 2, and the third one includes the data from Experiment 3. For all three data sets, we build three different models: the *strong penalty model*, which corresponds to the strong Uniform Penalty Hypothesis and uses only positive weights, the *rewards model*, which includes negative weights for ANI, DEF and NOM_{DAT}, and the *weak penalty model*, which models the acceptability difference within candidate sets rather than absolute values.

For the strong and the weak penalty model, we coded constraint violations in line with the notation in the violation profiles above (5, 12, etc.). For the rewards model, we re-coded ANI, DEF and NOM_{DAT} in such a way that a non-vacuous satisfaction of the constraints was counted instead of a violation. In addition to the constraints we assume, we coded two possible covariate factors that we assume correspond to a deficiency in plausibility, one for the animate theme conditions in Experiment 1 (ConAni), and one for Experiment 4 (Con4). In the weak penalty model, these covariates do not appear, as they become irrelevant when comparing only differences. We also left out the animate covariate in the rewards model – here, its coefficient was negligible, as the different constraint structure predicted already a difference between animate and inanimate themes. In addition, we had to take two decisions concerning the constraints to be included, as the data points were not enough to estimate all of them. The first one was the decision between REC or DAT for the weak penalty model, where we decided for DAT for no particular reason. The second one was for either EXP (experiencer > theme/stimulus) or AG, which in both cases are only possible to estimate for the full data set, then accounting for the difference between the distance between SO and OS orders in Experiment 3 compared to Experiment 4. We included AG, but suspect that its weight is overestimated and should be split between EXP and AG.

4.2.2 Results

All models capture the qualitative pattern fairly well, with two exceptions: The preference for animate theme subjects before animate recipient dative objects in passive in Experiment 1 is only predicted by the weak penalty model, due to a negative weight of NOM_{DAT}. The strong penalty model and the rewards model, in contrast, predict the opposite order to be preferred – although only with a small numerical benefit. The second case is the preference for dative 1st over dative 2nd in the passive condition where Definiteness is not aligned with Animacy in Experiment 2. The strong penalty model estimates a higher

weight for DEF than for ANI, thus predicting the opposite order. Note, however, that exactly these two contrasts were only numerically observable and did not reach significance in the pairwise comparisons.

For the other two criteria, the generality of constraint weights and the model fit, we refer the reader to Table 15, where all relevant models are summarized. First, we have some remarks relevant to all three model types (strong penalty, rewards and weak penalty). As described above, we fitted models for the complete data set as well as for two subsets. For the subset models, not all constraints were included, because the experiments in question did not contain the relevant manipulation. This also affected the coefficient of NOM_{ACC}, which is systematically lower for the whole data set than when fitted to Experiment 3 only. This does not indicate a different weight in different contexts, but is due to the additional constraint AG, which could only be estimated by comparing Experiment 3 to Experiment 4. When the two coefficients are summed, they match the NOM_{ACC} coefficient in Experiment 3 perfectly for the weak penalty model and at least more closely for the strong penalty and rewards models.

The generality – or stability – of constraint weights can best be assessed comparing the coefficients of ANI and DEF across experiments. For the strong penalty models, we see a huge amount of variation: If we only consider Experiment 3 (Model 3), ANI has a weight of zero[8] and DEF has an insignificant coefficient of –.02. In Experiments 1 and 2 (Model 2), on the other hand, both coefficients reach significance and are considerably larger. As expected, this discrepancy leads to medium-sized coefficients for the full data

Table 15: Statistical models for comparison of different quantitative accounts.

	strong penalty models			rewards models			weak penalty models		
	Model 1 all data	**Model 2 Exp 1 & 2**	**Model 3 Exp 3**	**Model 4 all data**	**Model 5 Exp 1 & 2**	**Model 6 Exp 3**	**Model 7 all data**	**Model 8 Exp 1 & 2**	**Model 9 Exp 3**
Intercept	0.30*** (0.02)	0.33*** (0.01)	0.27*** (0.04)	0.24*** (0.01)	0.20*** (0.02)	0.27*** (0.01)			
NOM_{ACC}	−0.14** (0.05)		−0.18** (0.04)	−0.15*** (0.04)		−0.25*** (0.01)	−0.20*** (0.03)		−0.25*** (0.01)
ANI	− 0.04 (0.02)	−0.08** (0.02)	0.00 (0.04)	0.07*** (0.02)	0.07** (0.02)	0.11*** (0.01)	−0.10*** (0.02)	−0.09* (0.03)	−0.11* (0.02)
DEF	−0.05* (0.02)	−0.07*** (0.01)	−0.02 (0.05)	0.03* (0.02)	0.04* (0.02)	0.06** (0.01)	−0.07*** (0.01)	−0.08* (0.02)	−0.08* (0.02)
AG	−0.07 (0.05)			−0.04 (0.04)			−0.04 (0.03)		
DAT	−0.05 (0.02)	−0.03 (0.02)		−0.05 (0.02)	−0.05 (0.02)		−0.01 (0.02)	−0.02 (0.03)	
NOM_{DAT}	0.01 (0.02)	−0.00 (0.02)		−0.01 (0.02)	0.02 (0.02)		−0.02 (0.02)	−0.02 (0.03)	
Con4	−0.09* (0.03)			−0.09** (0.03)					
ConAni	−0.09** (0.03)	−0.11*** (0.02)							
R^2	0.85	0.90	0.79	0.88	0.76	0.98	0.97	0.92	0.99
Adj. R^2	0.79	0.85	0.71	0.84	0.68	0.97	0.95	0.84	0.98

[8] Here and later assuming rounding to the second digit.

set (Model 1), where only DEF reaches significance, which contrasts with the ranking in (21). The situation is slightly better for the rewards models: There are no zero coefficients and animacy is significant for both data subsets (Models 5 and 6). While the exact numeric value changes, it could still be considered of comparable magnitude, if we compare it with NOM_{ACC} (much higher) or NOM_{DAT} (lower). By far the most stable pattern can be found for the weak penalty models: Here, ANI and DEF are significant for both subsets (Models 8 and 9) and the numerical difference between coefficients is small for ANI and even zero for DEF.

Lastly, we want to consider the model fits of the different model types. If we look at adjusted R^2, the weak penalty model again fares best for the full data set with an almost perfect fit of .95. The other two models also reach respectable adjusted R^2 values for the full data set, indicating that they do capture a fair amount of variation. It is interesting to compare the values between data sets again: As discussed above, in Experiments 1 and 2, there was only one obvious case where a constraint violation was not associated with a penalty. Accordingly, the model fit of the strong penalty model for this subset is relatively high. For Experiment 3, however, where we have multiple cases of cost-free violations, the model fit is considerably worse. For the rewards model, we see the opposite pattern. Experiment 3, where we have cost-free violations and a benefit for the suboptimal candidates, is modeled nicely, with a high adjusted R^2. For Experiments 1 and 2, however, where DEF and ANI seem to show a truly additive pattern, the model fit is worse than for the strong penalty model.

4.3 Discussion

The statistical results presented in this section show that fully quantified models of acceptability face problems with regard to our data. In particular, the strong penalty model and the rewards model cannot account for all of the data and show relatively poor stability of constraint weights across experiments. The weak penalty model, which only accounts for differences within candidate sets, on the other hand, shows stable constraint weights and good model fits for all data subsets. In addition to the predictions we can derive from the OT-based markedness model with the hierarchy in (21), we can now make precise numerical predictions with regard to the difference between optimal candidate and suboptimal candidate for unseen cases. If, e.g., we use object experiencer psych verbs as in Experiment 4, but with animate subjects, we expect the benefit of SO over OS to increase by roughly .1. Also, this model accommodates the possibility of ganging-up effects which might become necessary when we extend the model to include more phenomena.

5 Argument order in German

One of the goals of this paper was to evaluate a number of surface constraints on German argument order that have been proposed in the literature. We will now review the constraint set and ranking that is supported by our experimental data, contrast it to existing constraint sets, and finally compare it to findings from corpus studies.

5.1 A constraint set for German

While the majority of existing accounts of argument order in German relies on individual judgments or corpus examples, our work tested and compared a set of constraints by gathering judgment data from native speakers in a principled fashion by using factorial designs and keeping the participants unaware of the manipulated factors. As expected, our results mostly match informal data from earlier accounts with one exception detailed below. Importantly, our method allowed us to disentangle constraints that are often confounded in examples and to assess their individual contribution. As a result, we can

confirm the influence of the lexical-semantic constraint ANI, the syntactic constraints NOM_{ACC}, NOM_{DAT}, and DAT and the discourse constraint DEF on word order acceptability. For REC, we do not find direct evidence, whereas an influence of AG and/or EXP can only be derived by the numerical patterns, not by categorical preferences alone.

The minimal constraint hierarchy necessary to account for the results of Experiments 1–4 is given in (23), where constraints in parentheses cannot be placed conclusively with the evidence provided by our experiments.

(23) Minimal constraint ranking established by Experiments 1, 2, 3 and 4
NOM_{ACC} > ANI > DEF, NOM_{DAT}, (DAT)

If we take into account the numerical information, we arrive at the following constraint ranking along with their estimated weights:

(24) Constraint weights established by Experiments 1, 2, 3 and 4
NOM_{ACC} (.21) > ANI (.10) > DEF (.07) > AG (.04) > NOM_{DAT} (.02) > DAT (.01) > REC

Both hierarchies are incompatible with the subhierarchy of SCR-CRIT in (25) as proposed by Müller (1999).

(25) NOM > DEF > AN > FOC > DAT > ADV > PER

Firstly, his hierarchy predicts dative 2nd to score better than dative 1st in passive sentences, because of the high-ranked constraint NOM. The same holds for all other hierarchies in Table 1 in which NOM is the highest-ranked constraint. Secondly, he ranks DEF above ANI, while we find evidence for the opposite ordering when looking at those sentence pairs where ANI and DEF compete directly (conditions 5 and 2 and conditions 7 and 4 in Figure 1). Contrary to the intuition expressed in example (32) of Müller (1999), given above as (14), sentences with an indefinite animate dative object preceding the definite inanimate accusative object were judged better than their counterparts in our experiment. While our results therefore do not confirm Müller's specific constraint hierarchy, they are compatible with his model architecture, provided the splitting of NOM and the different ordering between ANI and DEF given in (23).

If we commit ourselves to weighted constraints, a striking feature of the constraint set we established is that it incorporates constraints related to thematic roles and case simultaneously, namely AG and NOM. This contrasts with the majority of the literature on word order variation in German, which shows a divide between accounts relying on thematic roles (Uszkoreit 1986; Haider & Rosengren 2003) or case (Uszkoreit 1987; Müller 1999) to identify unmarked or base-generated orders, with exceptions being Hoberg (1997) and to some extent Heck (2000).

In summary, our results show that a suitable set of surface-based constraints can capture observed word order preferences. The constraints in this set make reference to different types of linguistic information, including thematic roles, animacy, case and definiteness. Our results therefore argue against the view that the unmarked base order is determined solely by verb semantics, a view entertained by the scrambling account of Haider & Rosengren (2003). In particular, we were able to show that the effect of animacy cannot be reduced to thematic roles, as we observed significant differences even when using the same verbs with animate/inanimate direct objects.

5.2 Comparison of experimental data to corpus results

In the first major corpus study on argument order in the German middle field, Hoberg (1981) derived the word order template in (26) for NP arguments that are neither pronouns nor part of an idiomatic expression (see also Hoberg 1997).

(26) $(\text{NOM–ACC–DAT})_{\text{animate}}$ – $(\text{NOM–ACC–DAT})_{\text{inanimate}}$

The template in (26) amounts to a constraint hierarchy in which ANI is ranked highest and case decides when there is a tie with respect to ANI. Bader & Häussler (2010b) applied the word order template in (26) to their corpus sample and found that for about 90% of all sentences, order was correctly predicted. The same result was obtained when order was predicted from verb semantics, which roughly corresponds to constraints in terms of thematic roles. Bader & Häussler (2010b) could therefore not decide whether word order in the German middle field is determined by animacy or by thematic roles. A possibility not considered by either Hoberg (1997) or Bader & Häussler (2010b) is the intermingling of case constraints and the animacy constraint. This, however, is exactly what our experimental results argue for. While NOM_{ACC} is higher ranked than ANI, ANI is in turn higher ranked than NOM_{DAT}.

The evidence for ranking NOM_{ACC} higher than ANI comes from the finding that in sentences with an accusative object, SO order is preferred even if this violates the animacy constraint ANI, that is, even if the subject is inanimate and the object animate. In Experiment 3, this was shown for sentences containing an agentive verb. The use of agentive verbs with an inanimate subject and an animate object is rare in comparison to other associations between animacy and syntactic function, but it still occurs with some regularity. A typical authentic example is provided in (27) (from the deWac corpus; Baroni et al. 2009). Note that in this example, both NPs are definite and the first NP is longer than the second, so there are no confounding factors that could explain the order "inanimate subject before animate object."

(27) Man konnte noch nicht mit Sicherheit sagen, ob das Feuer oder der
one could yet not with certainty say whether the fire or the
Rauch die Astronauten getötet hatte.
smoke the astronauts killed has
'One couldn't say with certainty yet whether the fire or the smoke had killed the astronauts.'

In an ongoing corpus study that complements a series of production experiments, about 3500 sentences containing one of 24 action verbs were retrieved from the deWac corpus. 193 sentences, or about 5%, contained an inanimate subject and an animate object. In all sentences, the subject preceded the object, thus confirming that NOM_{ACC} is ranked above ANI. Given the rareness of examples of this kind, they probably occurred not at all or only with very low frequency in corpora of the sizes investigated by Hoberg (1997) or Bader & Häussler (2010b), resulting in a premature acception of the schema in (26). The ranking of ANI above NOM_{DAT} is already contained within the word order template in (26), for which – as pointed out above – a large body of corpus evidence is available.

Evidence for ranking NOM_{ACC} higher than ANI was also found in Experiment 4, which investigated sentences with an accusative object experiencer verb appearing together with an inanimate subject and an animate object. In this experiment, SO sentences like (28a) received higher ratings than OS sentences like (28b).

(28) a. Ich glaube, dass die Musik den Lehrer fasziniert hat.
I believe that the.NOM music the.ACC teacher fascinated has
'I believe that the music fascinated the teacher.'

b. Ich glaube, dass den Lehrer die Musik fasziniert hat.
I believe that the.ACC teacher the.NOM music fascinated has
'I believe that the music fascinated the teacher.'

In the syntactic literature, there is no consensus concerning object-experiencer psych verbs with an accusative object (for comprehensive discussion, see Verhoeven 2015). Whereas some authors treat such verbs on a par with other verbs selecting a direct object (e.g., Vogel & Steinbach 1995; Fanselow 2000), other authors have claimed that accusative object-experiencer verbs are like dative object-experiencer verbs in that OS order is unmarked for them (e.g., Lenerz 1977; Haider & Rosengren 2003). In a sentence selection experiment, Haupt et al. (2008) found an SO preference for object-experiencer verbs with an accusative object when the subject was definite and the object indefinite whereas OS order was preferred when the object was definite and the subject indefinite. This was true independently of animacy. In a recent corpus study, Verhoeven (2015) found a moderate preference for middle field-internal OS order for accusative object-experiencer verbs when the object was animate and the subject inanimate. When both were animate, in contrast, a strong preference for SO was observed.

In the ongoing corpus study mentioned above, we are also analyzing a set of 24 object-experiencer verbs with an accusative object. As in Verhoeven (2015), all sentences containing a subject or an object pronoun were removed. For the remaining 171 examples with an animate object and an inanimate subject, both occurring within the middle field, 23% of the sentences occurred with OS order and 77% with SO order. For a set of 82 sentences in which both subject and object were animate, a strong SO preference emerged, as in the study of Verhoeven (2015). For sentences with inanimate subject and animate object, the results for the object-experiencer psych verbs clearly differ from the results for the agentive verbs, for which not a single OS sentence was found.

For agentive verbs and for object-experiencer verbs with two animate arguments, the corpus findings of Verhoeven and our corpus findings are almost identical. For object-experiencer psych verbs with an animate object and an inanimate subject, in contrast, they differ. Verhoeven found a moderate preference for OS order whereas we found a moderate preference for SO order. We do not know what accounts for this difference. One reason could be that psych verbs in general are more variable with regard to order than other verbs. For psych verbs, an OS preference turned up in Verhoeven's study only when both arguments were contained in the middle field. When one of the arguments was contained within the prefield, SO order was preferred. For agentive verbs, in contrast, a strong preference for SO order was observed independently of where the arguments were located. In addition, in a production experiment involving accusative object-experiencer psych verbs, Temme & Verhoeven (2016) found a strong context effect on order. Given that the number of sentences with inanimate subject and animate object was neither large in Verhoeven (2015) nor in our study, the preceding discourse context may have had particularly strong effects. More research is needed in order to resolve this issue.

6 Conclusion and outlook

This paper has presented four experiments that have investigated the effect of surface constraints on the order of arguments in the German middle field. As shown in Table 1, the literature contains a range of constraint hierarchies that are incompatible with each other both with regard to the proposed constraint set and with regard to the ranking of the

constraints. None of the existing constraint hierarchies was confirmed in its entirety by our results. A main contribution of our experiments is thus that we arrived at a constraint hierarchy that is empirically more adequate than prior hierarchies. With this hierarchy, we were able to show that surface order constraints can capture argument order preferences in German to a great degree. On the more theoretical side, the strong Uniform Penalty Hypothesis was not confirmed, as in some contexts optimal candidates violate constraints without any decrease in acceptability. The weak version, however, which only concerns the differences within candidate sets, captures the observed patterns quite well. Müller's variant of standard OT is able to account for the qualitative pattern as well, provided we use an altered constraint set and ranking.

Given that we can capture the observed acceptability differences between different orders in the grammar, the controversial question remains whether we should do so. One question here is whether these constraints cannot be placed in other submodules, such as a processing component or a pragmatic component. With regard to this question, we already mentioned in the introduction that it is uncontroversial that some effects on acceptability can be attributed to the processing mechanisms responsible for sentence comprehension. However, we do not see any systematic differences in complexity between our experimental conditions (although there certainly are differences between the experiments) which could explain our experimental results with recourse to sentence processing. Our sentences did not involve local ambiguities and they were not syntactically complex due to center embedding or any other kind of non-local dependency formation. It is therefore unlikely that the acceptability differences that we observed in our experiments reflect different processing costs.

A further possibility to consider is whether information-structural properties of the sentences under consideration may be responsible for the observed acceptability differences. As we already mentioned in section 2, following the pioneering work of Lenerz (1977) and Höhle (1982), the distinction between marked and unmarked word orders is often defined in terms of the focus potential of a sentence, where sentences with unmarked word order allow both wide and narrow focus whereas sentences with marked order allow narrow focus only. Since we tested our sentences out of context, it is possible that the acceptability differences yielded by our experiments reflect the fact that for sentences allowing only narrow focus, contextual requirements were not full-filled. This is surely a possibility that deserves further investigation, in particular by testing whether differences in acceptability disappear when sentences are presented with appropriate contexts. Note however that, with the exception of DEF, which is inherently connected to information structure, our constraints are independent of information structure. Thus, even if the lack of contextual licensing is the source of the reduced acceptability of marked word orders, constraints of the sort considered in this paper will still be necessary in order to define whether a word order is marked or not. One way to achieve this could be to combine surface-oriented constraints as considered here with constraints regulating the mapping between syntax and prosody as proposed in such work as Büring (2001) and Samek-Lodovici (2005).

A further general question raised by our experiments (and also by other experiments measuring fine grades of acceptability) is whether the sometimes subtle acceptability differences that we observed are sufficient to motivate the inclusion of our constraints into the grammar. This question is connected to the kind of data we use and the issue of gradience in general. If we ask for judgments of two different sentences using magnitude estimation, we will in the vast majority of cases observe a difference in acceptability – irrespective of whether they differ with regard to well-formedness. This problem is addressed in our work by using minimal pairs to exclude effects of word frequency,

syntactic complexity etc. and in running inferential analyses to see which differences are actually significant. In addition, we might look at the effect size to make our results comparable to other studies. Indeed, as far as comparisons of sentences which differed only with regard to one constraint are concerned, we only observed very small to small effect sizes using Cohen's d. If multiple constraint violations coincide, however, the effects become medium to large (Cohen's $d > 1.2$). Using effect sizes would thus urge us to include the whole set of constraints, because in combination they do produce large effects.

As a final point, we would like to discuss some consequences of the experimental data presented in this paper with regard to the relationship between corpus frequencies and acceptability. At the theoretical level, our results add to the existing evidence that there is no straightforward correlation between the acceptability and the frequency of syntactic structures (Featherston 2005; Kempen & Harbusch 2005; Arppe & Järvikivi 2007; Bader & Häussler 2010a). This is shown most clearly by the finding that in Experiments 1 and 2, unmarked passive sentences were no less acceptable than unmarked active sentences, despite the fact that passive sentences are much less frequent than active sentences (see Bader 2012 for the case of ditransitive verbs). The lack of a correlation between frequency and acceptability at the level of whole sentence structures does not preclude that the weights of individual constraints reflect corpus frequencies. There exist several proposals as to how constraint weights can be learned from corpus frequencies (e.g., Goldwater & Johnson 2003; Jäger 2007). In this case, we would not expect a correlation between the overall frequency of a structure and its acceptability, but a correlation between frequency differences and acceptability differences. For example, our results revealed that the constraint NOM_{ACC} has a greater weight, and is thus higher ranked, than the constraint NOM_{DAT}. This comports with the finding from corpus studies that the ratio of SO to OS sentences is much higher for sentences with an accusative object than for sentences with a dative object (see Hoberg 1981; Kempen & Harbusch 2005; Bader & Häussler 2010b). Deriving constraint weights from frequency counts is an attractive option because it would answer an unresolved question raised by the assumption of weighted constraints – namely the question of where constraint weights come from. However, this is a controversial issue that is beyond the scope of the current paper, both for reasons of space and because the frequency data necessary for addressing this issue are not readily available. Whether the complete constraint hierarchy established in this paper can be derived from corpus frequencies must therefore be left as a question for future research.

Additional File

The additional file for this article can be found as follows:

- **Appendix.** This appendix lists the complete sentence material used in Experiments 1–4. For each experiment, we first give a table showing the complete design of the experiment. The following sentence list shows only a subset of the conditions. Conditions that are not shown can be reconstructed from the complete design given in the table. DOI: https://doi.org/10.5334/gjgl.258.s1

Abbreviations

NOM = nominative, DAT = dative, ACC = accusative, DEF = definite, INDEF = indefinite, ANI = animate, INANI = inanimate, S = subject, IO = indirect object, DO = direct object, SO = subject before object, OS = object before subject, NP = noun phrase, OT = Optimality Theory, HG = Harmonic Grammar, LOT = Linear Optimality Theory

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (Project DFG-BA1598). We would like to thank Yvonne Portele, Vasiliki Koukouloti, Alice Schäfer and two anonymous reviewers for their helpful comments.

Competing Interests

The authors have no competing interests to declare.

References

Arppe, Antti & Juhani Järvikivi. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131–159. DOI: https://doi.org/10.1515/CLLT.2007.009

Bader, Markus. 2012. The German *bekommen* passive: A case study on frequency and grammaticality. *Linguistische Berichte* 231. 249–298.

Bader, Markus & Jana Häussler. 2010a. Toward a model of grammaticality judgments. *Journal of Linguistics* 46(2). 273–330. DOI: https://doi.org/10.1017/S0022226709990260

Bader, Markus & Jana Häussler. 2010b. Word order in German: A corpus study. *Lingua* 120(3). 717–762. DOI: https://doi.org/10.1016/j.lingua.2009.05.007

Bard, Ellen Gurman, Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1). 32–68. DOI: https://doi.org/10.2307/416793

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation Journal* 23(3). 209–226. DOI: https://doi.org/10.1007/s10579-009-9081-4

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Büring, Daniel. 2001. Let's phrase it! Focus, word order, and prosodic phrasing in German double object constructions. In Gereon Müller & Wolfgang Sternefeld (eds.), *Competition in syntax*, 69–105. Berlin & New York: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110829068.69

Chomsky, Noam. 1955/1975. *The logical structure of linguistic theory*. Chicago: University of Chicago Press.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Cowart, Wayne. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage Publications.

Fanselow, Gisbert. 2000. Optimal Exceptions. In Barbara Stiebels & Dieter Wunderlich (eds.), *The Lexicon in Focus*, 173–209. Berlin: Akademie Verlag. DOI: https://doi.org/10.1515/9783050073712-009

Fanselow, Gisbert. 2001. Features, theta-roles, and free constituent order. *Linguistic Inquiry* 32. 405–437. DOI: https://doi.org/10.1162/002438901750372513

Fanselow, Gisbert. 2003. Free constituent order: A minimalist interface account. *Folia Linguistica* 37(1–2). 191–232. DOI: https://doi.org/10.1515/flin.2003.37.1-2.191

Fanselow, Gisbert, Caroline Féry, Ralf Vogel & Matthias Schlesewsky. (eds.) 2006. *Gradience in grammar: Generative perspectives*. New York: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199274796.001.0001

Featherston, Sam. 2005. The decathlon model of empirical syntax. In Marga Reis & Stephan Kepser (eds.), *Linguistic evidence. Empirical, theoretical and computational perspectives*, 187–208. Berlin: de Gruyter. DOI: https://doi.org/10.1515/9783110197549.187

Fodor, Janet Dean & Fernanda Ferreira. (eds.) 1998. *Reanalysis in sentence processing*. Dordrecht: Kluwer. DOI: https://doi.org/10.1007/978-94-015-9070-9

Frey, Werner. 1993. *Syntaktische Bedingungen für die semantische Interpretation: über Bindung, implizite Argumente und Skopus*. Berlin: Akademie-Verlag.

Fukuda, Shin, Grant Goodall, Dan Michel & Henry Beecher. 2012. Is Magnitude Estimation worth the trouble. In Jaehoon Choi, E. Alan Hogue, Jeffrey Punske, Deniz Tat, Jessamyn Schertz & Alex Trueman (eds.), *Proceedings of the 29th West Coast Conference on formal linguistics*, 328–336. Somerville, MA: Cascadilla Proceedings Project.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita & Wayne O'Neil (eds.), *Image, language, brain. Papers from the first Mind Articulation Project Symposium*, 95–126. Cambridge, MA: MIT Press.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint ranking using a maximum entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111–120. University of Stockholm.

Goodwin, C. James. 2003. Psychology's experimental foundations. In Stephen F. Davis (ed.), *Handbook of research methods in experimental psychology*, 3–23. Malden, MA: Blackwell. DOI: https://doi.org/10.1002/9780470756973.ch1

Haider, Hubert. 2010. *The syntax of German*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511845314

Haider, Hubert & Inger Rosengren. 2003. Scrambling: Non-triggered chain formation in OV languages. *Journal of Germanic Linguistics* 15. 203–266. DOI: https://doi.org/10.1017/S1470542703000291

Haupt, Friederike S., Matthias Schlesewsky, Dietmar Roehm, Angela D. Friederici & Ina Bornkessel-Schlesewsky. 2008. The status of subject–object reanalyses in the language comprehension architecture. *Journal of Memory and Language* 59(1). 54–96. DOI: https://doi.org/10.1016/j.jml.2008.02.003

Hawkins, John. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199252695.001.0001

Hawkins, John. 2006. Gradedness as relative efficiency in the processing of syntax and semantics. In Gisbert Fanselow, Caroline Féry, Ralf Vogel & Matthias Schlesewsky (eds.), *Gradience in grammar: Generative perspectives*, 207–226. New York: Oxford University Press.

Heck, Fabian. 2000. Tiefenoptimierung. *Linguistische Berichte* 184. 441–468.

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

Hoberg, Ursula. 1997. Die Linearstruktur des Satzes. In Gisela Zifonun, Ludger Hoffmann & Bruno Strecker (eds.), *Grammatik der deutschen Sprache*, 1495–1680. Berlin: de Gruyter.

Hofmeister, Philip, Florian Jaeger, Inbal Arnon, Ivan A. Sag & Neal Snider. 2013. The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes* 28(1–2). 48–87. DOI: https://doi.org/10.1080/01690965.2011.572401

Hofmeister, Philip & Ivan A. Sag. 2010. Cognitive constraints and island effects. *Language* 86(2). 366–415. DOI: https://doi.org/10.1353/lan.0.0223

Höhle, Tilman N. 1982. Explikation für "normale Betonung" und "normale Wortstellung". In Werner Abraham (ed.), *Satzglieder im Deutschen. Vorschläge zur syntaktischen, semantischen und pragmatischen Fundierung*, 75–153. Tübingen: Narr.

Jacobs, Joachim. 1988. Probleme der freien Wortstellung im Deutschen. *Sprache und Pragmatik* 5. 8–37.

Jaeger, Florian & Elisabeth J. Norcliffe. 2009. The cross-linguistic study of sentence production. *Language and Linguistics Compass* 3(4). 866–887. DOI: https://doi.org/10.1111/j.1749-818X.2009.00147.x

Jäger, Gerhard. 2007. Maximum entropy models and Stochastic Optimality Theory. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling & Chris Manning (eds.), *Architectures, rules, and preferences: A Festschrift for Joan Bresnan*, 467–479. Stanford, CA: CSLI.

Keenan, Edward L. & Bernard Comrie. 1977. Noun phrase accesssability and universal grammar. *Linguistic Inquiry* 8. 63–99.

Keller, Frank. 1996. *Extraction from complex noun phrases. A case study in graded grammaticality*. Stuttgart: MA Thesis University of Stuttgart.

Keller, Frank. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Edingburgh: University of Edinburgh dissertation.

Keller, Frank. 2006. Linear optimality theory as a model of gradience in grammar. In Gisbert Fanselow, Caroline Féry, Ralf Vogel & Matthias Schlesewsky (eds.), *Gradience in grammar: Generative perspectives*, 270–287. New York: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199274796.003.0014

Kempen, Gerard & Karin Harbusch. 2005. The relationship between grammaticality ratings and corpus frequencies: A case study into word-order variability in the midfield of German clauses. In Marga Reis & Stephan Kepser (eds.), *Linguistic evidence. Empirical, theoretical and computational perspectives*, 329–349. Berlin: de Gruyter. DOI: https://doi.org/10.1515/9783110197549.329

Legendre, Géraldine, Y. Miyata & Paul Smolensky. 1990. Can connectionism contribute to syntax? Harmonic Grammar, with an application. In Karen Deaton, Manuela Noske & Michael Ziolkowski (eds.), *Proceedings of the 26th regional meeting of the Chicago Linguistic Society*, 237–252. Chicago: Chicago Linguistic Society.

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

Lenerz, Jürgen. 2001. Word order variation: Competition or co-operation? In Gereon Müller & Wolfgang Sternefeld (eds.), *Competition in syntax* 49. 249–281. (Studies in Generative Grammar). Berlin & New York: Mouton de Gruyter.

Meinunger, Andre. 2000. *Syntactic aspects of topic and comment*. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/la.38

Müller, Gereon. 1999. Optimality, markedness, and word order in German. *Linguistics* 37. 777–818. DOI: https://doi.org/10.1515/ling.37.5.777

Nieuwenhuis, Rense, Manfred te Grotenhuis & Ben Pelzer. 2012. Influence. ME: Tools for detecting influential data in mixed effects models. *R journal* 4(2). 38–47. http://journal.r-project.org/archive/2012-2/RJournal_2012-2_Nieuwenhuis~et~al.pdf.

Pater, Joe. 2009. Weigthed constraints in generative linguistics. *Cognitive Science* 33. 999–1035. DOI: https://doi.org/10.1111/j.1551-6709.2009.01047.x

Pechmann, Thomas, Hans Uszkoreit, Johannes Engelkamp & Dieter Zerbst. 1996. Wortstellung im deutschen Mittelfeld. In Sascha Felix, Siegfried Kanngießer & Gert Rickheit (eds.), *Perspektiven der kognitiven Linguistik*, 257–299. Opladen: Westdeutscher Verlag. DOI: https://doi.org/10.1007/978-3-663-07678-0_11

Phillips, Colin, Matthew Wagers & Ellen W. Lau. 2011. Grammatical illusions and selective fallibility in real-time language comprehension. In Jeffrey T. Runner (ed.), *Experiments at the interfaces*, 147–180. Bingley, UK: Emerald.

Prince, Alan & Paul Smolensky. 1993/2004. *Optimality theory. Constraint interaction in generative grammar*. Oxford: Blackwell.

R Development Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/.

Rösler, Frank, Thomas Pechmann, Judith Streb, Brigitte Röder & Erwin Hennighausen. 1998. Parsing of sentences in a language with varying Word Order: Word-by-word variations of processing demands are revealed by event-related brain potentials. *Journal of Memory and Language* 38. 150–176. DOI: https://doi.org/10.1006/jmla.1997.2551

Samek-Lodovici, Vieri. 2005. Prosody-syntax interaction in the expression of focus. *Natural Language & Linguistic Theory* 23. 687–755. DOI: https://doi.org/10.1007/s11049-004-2874-7

Schütze, Carson T. 1996. *The empirical base of linguistics*. Chicago: Chicago University Press.

Schütze, Carson T. & Jon Sprouse. 2014. Judgment data. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 27–50. Cambridge: Cambridge University Press.

Sprouse, Jon, Matthew Wagers & Colin Phillips. 2012. A test of the relation between working-memory capacity and syntactic island effects. *Language* 88(1). 82–123. DOI: https://doi.org/10.1353/lan.2012.0004

Temme, Anne & Elisabeth Verhoeven. 2016. Verb class, case, and order: A crosslinguistic experiment on non-nominative experiencers. *Linguistics* 54(4). 769–813. DOI: https://doi.org/10.1515/ling-2016-0018

Uszkoreit, Hans. 1986. Constraints on order. *Linguistics* 24. 883–906. DOI: https://doi.org/10.1515/ling.1986.24.5.883

Uszkoreit, Hans. 1987. *Word order and constituent structure in German*. Chicago: Chicago University Press.

Verhoeven, Elisabeth. 2015. Thematic asymmetries do matter! A corpus study of German word order. *Journal of Germanic Linguistics* 27(01). 45–104. DOI: https://doi.org/10.1017/S147054271400021X

Vogel, Ralf & Markus Steinbach. 1995. On the (absence of a) base position for dative objects in German. *FAS Papers in Linguistics* 4. 99–131.

Warren, Tessa & Edward Gibson. 2002. The influence of referential processing on sentence complexity. *Cognition* 85. 79–112. DOI: https://doi.org/10.1016/S0010-0277(02)00087-2

Weskott, Thomas & Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87(2). 249–273. DOI: https://doi.org/10.1353/lan.2011.0041

How to cite this article: Ellsiepen, Emilia and Markus Bader. 2018. Constraints on Argument Linearization in German. *Glossa: a journal of general linguistics* 3(1): 6.1–36, DOI: https://doi.org/10.5334/gjgl.258

Submitted: 12 September 2016 **Accepted:** 19 April 2017 **Published:** 12 January 2018