Gerard Kempen (Nijmegen) and Karin Harbusch (Koblenz)

Comparing Linguistic Judgments and Corpus Frequencies as Windows on Grammatical Competence: A Study of Argument Linearization in German Clauses

1 Introduction and Preview

When language users grammatically encode a communicative intention, they often avail of a range of linguistic means — each option yielding a member of a set of paraphrases. A rich source of paraphrases is word order variation. In languages where word order is not entirely free, some members of a set of paraphrastic linear orders tend to be judged as more acceptable than others. Such linear order preferences and dispreferences need to be encoded somehow in the grammars of those languages. Two prominent methods to measure linear order preferences are *gradient grammaticality ratings* and *corpus frequency counts*. This raises the question to what extent the two methods yield equivalent results.

In the present chapter, taking German as target language, we explore this question in two linear order domains: NP argument sequences in the midfield of finite subordinate clauses, and right-branching versus left-branching infinitival complements with *zu* 'to' immediately preceding the head verb. In both domains, the two methods for measuring linear order preferences produced non-equivalent outcomes. On the one hand, linear orders that receive similarly high grammaticality ratings may have very different but at least moderate corpus frequencies. On the other hand, linear orders whose corpus frequency is (virtually) zero, may nevertheless elicit variable grammaticality ratings, although at best moderate ones.

We address two questions raised by this "grammaticality-frequency gap." The first one is theoretical in nature: How to explain the observed grammaticality-frequency discrepancies in terms of the tasks performed by the language users (grammaticality judgment versus sentence production)? The second question concerns a methodological issue: Which preference assessment method provides the clearest view of the language user's internal grammar? More precisely, which method is more successful in avoiding under- and overgeneration, that is, yields data from which a grammar with a higher level of observational adequacy can be induced?

In modern linguistics, language intuitions of native speakers, in particular their ratings of the level of grammaticality of word strings, are viewed as the *via regia* to their linguistic "competence" and serve as the primary source of empirical data for the construction of natural-language grammars. Data gleaned from corpus explorations — the second method — are often held less trustworthy due to the possibility of contamination by "performance" factors. However, as argued among others by Gerken and Bever (1986), Schütze (1996) and more recently by Luka and Barsalou (2005), linguistic intuitions are not immune to contaminations by performance factors either, and hence do not necessarily mirror properties of the language user's internal grammar directly.

Based on the differences we observed between the linear order preferences emerging from the two methods, and in line with a proposal by Schütze (1996), we argue that judgments of moderately grammatical structures are biased by the outcome of another judgment process: judging the structural similarity between to-be-rated target sentences and 'ideal delivery' paraphrases produced by the raters themselves. We conclude that, at least in the linearization domain, the goal of gaining a clear view of the internal grammar of language users is best served by a combined strategy in which grammar rules are founded on structures that elicit moderate to high grammaticality ratings *and* attain moderate to high usage frequencies.

In Section 2, we present a summary of the quantitative data that made us aware of the existence of the grammaticality-frequency gap in the linearization domain under discussion. More details can be found in earlier publications (Kempen and Harbusch, 2003, 2005). A theoretical account of the data is proposed in Section 3. Finally, Section 4 summarizes the main conclusion and adds some implications and qualifying remarks.

2 The Distant Relationship between Grammaticality and Frequency: A Summary of Empirical Observations

2.1 Linear Order of NP Arguments in the Midfield of Finite Subordinate Clauses

In German, the order of argument NPs — Subject (S), Indirect Object (I), and Direct Object (O) — is relatively flexible; none of the six possible sequences is absolutely unacceptable (Müller, 1999). Keller (2000) observed a great deal of variation in the rated grammaticality of the different argument orders in lexically identical subordinate clauses of German. In one of his experiments, for example, native speakers of German assigned a high acceptability score to (1a) and a very low one to (1b) (see Table 1 for details). String (1c) received an intermediate rating. The argument NPs in (1a-c) are non-pronominal ("full"). Keller also elicited acceptability/grammaticality ratings for three comparable clause types where one argument NP was pronominal: *er* 'he' (S*p*), *ihm* 'him$_{DAT}$' (I*p*), or *ihn* 'him$_{ACC}$' (O*p*). Two examples, both of intermediate grammaticality, are shown in (2). Thus, in this experiment, the informants were presented with four "families" of clauses. In one family, all NPs were full; in three other families, every sentence had one pronominal NP: S*p*, I*p*, or O*p*. Each family occurred with six different argument permutations. This gives the 24 ordering patterns listed in the first column of Table 1.

(1) (a) S–I–O dass der Produzent dem Regisseur den Schauspieler vorschlägt
that the$_{NOM}$ producer the$_{DAT}$ director the$_{ACC}$ actor proposes
'that the producer proposes the actor to the director'
(b) O–I–S dass den Schauspieler dem Regisseur der Produzent vorschlägt
(c) I–S–O dass dem Regisseur der Produzent den Schauspieler vorschlägt
(2) (a) I–S*p*–O dass dem Regisseur er den Schauspieler vorschlägt
(b) S–O–I*p* dass der Produzent den Schauspieler ihm vorschlägt

The finding that we focus on in this chapter is shown in the left-hand panel of Figure 1 (for details see Keller, 2000). It depicts the relationship between the mean grammaticality *score* (Y-axis) and the grammaticality *ranking* (X-axis) for six argument permutations, combined for the four clause families. The graph was calculated as follows. The grammaticality rating scores reported by Keller have undergone a logarithmic transformation (with base 10) — in accordance with usual practice in Magnitude Estimation experiments (Bard, Robertson and Sorace, 1996). We started by undoing this transformation. Then, within each of the four clause families, we ranked the six permutations from the highest (rank 1) to the lowest (rank 6) average grammaticality rating (see column 2 of Table 1). Next, we computed an overall mean rating score for the six ranking positions — by adding, for each rank position in the four families, the untransformed ratings and dividing the sum by four. Finally, we applied the same logarithmic transformation to the six overall mean rating scores. As shown in Figure 1, the average scores decrease more or less regularly when going from high to low rank positions; only between the second and third rank, the slope is somewhat steeper than between the other positions.

In order to find out whether grammaticality ratings are interchangeable with corpus frequencies as measures of linear order preferences, we determined the frequencies of argument orderings in the same type of clauses as used by Keller (2000). As text sources served the NEGRA and TIGER treebanks, which contain written materials (Skut *et al.*, 1997), and the VERBMOBIL[1] corpus for spoken language. The frequency patterns emerging from these different corpora turned out to be remarkably similar (for details of the calculations, see Kempen and Harbusch, 2005).

[1] URL at the time of writing: http://www.ims.unistuttgart.de/projekte/verbmobil/Dialogs/

Table 1. Grammaticality ratings, corpus frequencies, and anteriority ranks for 24 ditransitive argument ordering patterns. Column 1: argument orders and NP shapes (full vs. pronominal). Columns 2 and 3: grammaticality values and their rank order (from Keller, 2000, Experiment 6). Column 4: frequencies in NEGRA and TIGER corpora. Column 5: frequencies in VERBMOBIL corpus. Columns 6 through 8: anteriority ranks (explained in Section 3). In only one case (marked by the gray line), the grammaticality rank of an argument permutation (in column 2) does not correspond to its mean anteriority rank (in column 8). Table reproduced from Kempen and Harbusch (2005). In the spoken corpus (see coulmn 5), the zero frequency of S–I–O and S–O–I permutations in the first clause family relates to the fact that VERBMOBIL consists exclusively of dialog turns, with many occurrences of pronominal NPs (first- and second-person pronouns).

Ordering	Grammaticality		Frequency		Anteriority ranks		
pattern	rank	judgment	written	spoken	subject	pronoun	mean
S–I–O	1	.2083	54	0	1	–	1
S–O–I	2	.0994	5	0	1	–	1
I–S–O	3	-.0716	0	0	2	–	2
O–S–I	4	-.2038	0	0	2	–	2
I–O–S	5	-.2667	0	0	3	–	3
O–I–S	6	-.2736	0	0	3	–	3
S*p*–O–I	1	.1519	4	1	–	1	1
S*p*–I–O	2	.1386	13	4	–	1	1
I–S*p*–O	3	-.1463	0	0	–	2	2
O–S*p*–I	4	-.2081	0	0	–	2	2
I–O–S*p*	5	-.2936	0	0	–	3	3
O–I–S*p*	6	-.3471	0	0	–	3	3
S–I*p*–O	1	.1471	30	6	1	2	1.5
I*p*–S–O	2	.1144	29	4	2	1	1.5
S–O–I*p*	3	-.0516	0	0	1	3	2
O–I*p*–S	4	-.2164	0	0	3	2	2.5
I*p*–O–S	5	-.2612	0	0	3	1	2
O–S–I*p*	6	-.2810	0	0	2	3	2.5
S–O*p* –I	1	.1938	3	1	1	2	1.5
O*p*–S–I	2	.1235	12	0	2	1	1.5
S–I–O*p*	3	-.1876	0	0	1	3	2
O*p*–I–S	4	-.2247	0	0	3	1	2
I–S–O*p*	5	-.2694	0	0	2	3	2.5
I–O*p*–S	6	-.3550	0	0	3	2	2.5

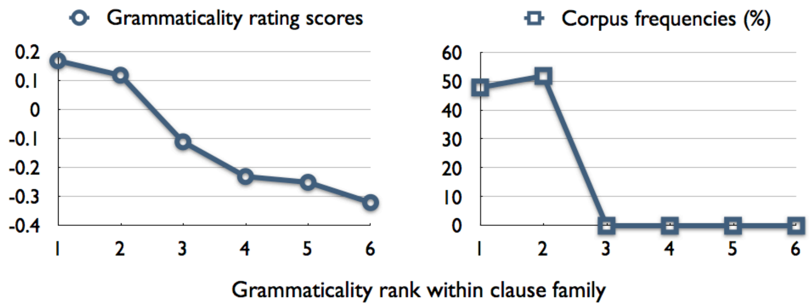

Figure 1. The grammaticality-frequency gap. For explanation see text of Section 2.

The right-hand panel of Figure 1 displays the relative frequencies corresponding to the rank positions of the argument permutations in the left-hand panel. We arrived at these numbers as follows. For each of the 24 argument ordering patterns, we counted the number of finite ditransitive subordinate clauses with the same structure (i.e., same argument sequence, same clause family; see columns 4 and 5 in Table 1). This number was expressed as a percentage of the total number of observations in the same clause family (thus assigning the same weight to each of the four clause families). For every rank position, we averaged the percentages over the clause types. The result is displayed in the right-hand panel of Figure 1.

To our surprise, we could not find a single matching clause for two thirds of the ordering patterns: 48 percent of the argument orderings attested in the corpus embodied the argument ordering that received the highest grammaticality rating in its clause family; and the remaining 52 percent matched the ordering with the one-but-highest mean rating. The percentages for all four remaining orderings were zero. In sum, the ditransitive clauses we did find in the three corpora, all exhibited an argument sequence that obtained the highest or one-but-highest grammaticality score in its family. This finding reveals one side of the grammaticality-frequency gap. *Each of the corpora only contained argument NP orderings in the high-grammaticality range; linear orders of moderate grammaticality were absent and did not outnumber low-grammaticality orders.*

For the other side of the gap, we now turn to a different construction.

2.2 Right- versus Left-Branching Infinitival *zu*-Complement Clauses

Many German verbs that take an infinitival *zu*-complement[2] allow it to branch leftward or rightward. Examples are *versprechen* 'promise' and *versuchen*

[2] As for terminology, we consider complement clauses as arguments (as opposed to adjuncts) of the verbs that govern them.

'try'. The subordinate clauses in (3a) and (3b) contain left- and right-branching complements, respectively. Rambow (1994) reported informally collected grammaticality judgments for 30 linear order variants of the constituents within this clause. Two members of this set are (4a) and (4b), both judged highly grammatical. Seuren (2003) obtained less informal grammaticality ratings for the same set of clauses, with essentially the same results. We checked whether the informants in the two studies systematically assigned higher grammaticality ratings to either left-branching or right-branching constructions. However, no such directional preference became apparent. Without going into the details, we conclude that, for native speakers of German, right- and left-branching *zu*-complements are about equally highly grammatical.

(3)	(a)	weil	niemand	[S das Fahrrad	zu reparieren]	versucht/verspricht
		because	nobody	the bike	to repair	tries/promises
		'because nobody tries/promises to repair the bike'				
	(b)	weil niemand versucht/verspricht [S das Fahrrad zu reparieren]				
(4)	(a)	weil niemand [S [S das Fahrrad zu reparieren] zu versuchen] verspricht				
		'because nobody promises to try to repair the bike'				
	(b)	weil niemand verspricht [S zu versuchen [S das Fahrrad zu reparieren]]				

The corpus frequencies of the corresponding right-and left-branching structures exhibited a rather different pattern. In the NEGRA and TIGER treebanks as well as in the VERBMOBIL and W-PUB[3] corpora, both *versuchen* and *versprechen* have a strong preference of around 85 percent for right-branching *zu*-complements. A similar directional bias holds for other complement-taking verbs as well, e.g. *erwägen* 'consider', *beginnen* 'begin', *drohen* 'threaten'.

This observation is the second side of the grammaticality-frequencies gap: *Argument orderings of similarly high grammaticality turn out to have rather different corpus frequencies*. Without going into the details, we add here that this characterization also holds for the NP arguments of the four clause families discussed in the previous subsection (The reader may wish to check this in Table 1, by inspecting the corpus frequencies in the first two rows of every clause family.)

3 Explaining the Grammaticality-Frequency Gap

Taking the grammaticality judgments as direct reflections of the informants' internal grammar would yield a more lenient grammar than a grammar based

[3] The W-PUB corpus, which resides at the Institut für Deutsche Sprache in Mannheim, Germany, consists of plain text without any annotations. Using the COSMAS-II search tool, we searched for combinations of the past-participle form (which nearly always occurs in clause-final position) of complement-taking verbs with *zu* 'to' preceding or following it. We analyzed by hand random samples of 200 sentences per verb, or if fewer than 200 could be found, all available examplars.

only on structures that actually occur in a corpus. For instance, a judgment-based grammar would generate moderately acceptable strings (1c), (2a) and (2b), although perhaps only as marginal cases. A frequency-based grammar would rule them out completely. Likewise, a corpus-based grammar would assign higher acceptability scores to right-branching than to left-branching *zu*-complements, whereas a judgment-based grammar would generate the two structures with equal quality indices. This state of affairs creates the following problem for the grammarian who aims to avoid over- and undergeneration by the to-be-developed grammar: Which type of data yields the closest approximation to the set of strings generated by the language user's internal grammar (the highest level of observational adequacy)?

The arguments against relying on corpus (i.e. performance) data are well-known and need not be repeated here (for an overview, see Schütze, 1996). In the two studies summarized in Section 2, the following performance factors inherent in corpus data may have obscured the view of the native speakers' linguistic competence. The much higher incidence of right-branching than left-branching clausal complements is easily explainable in terms of processing load. The right-branching structures enable the grammatical encoding process to release the complement-taking verb from working memory at an earlier point in time than their left-branching counterparts do. The production frequencies in the NP argument study can be understood by invoking the notion of incremental production. As we argued elsewhere (Kempen and Harbusch, 2003), pronominal constituents are easier to compute than full (non-pronominal) NPs, and hence are ready to receive an ordinal position in the emerging clause at earlier points in time. Thus, they can trigger linear-order rules that license more anterior (leftward) positions. The same holds for Subject NPs, at least for those fulfilling the role of topic. It is reasonable to assume that topical information becomes available to the grammatical encoder prior to the information that is to be expressed in the comment/predicate, including direct and indirect objects. Indeed, all linear orders with frequencies greater than zero had the Subject NP and/or a pronominal NP as earliest possible constituent(s) (for details, see the data in Table 1).

The assumption that grammaticality judgments are direct reflections of the internal grammar (hence, can serve as a data source for the induction of a grammar that aims at avoiding over- or undergeneration), can yield an account of the grammaticality-frequency gap, at least in principle. In such an account, the internal grammar should generate pairs consisting of a sentence and a numerical index indicating the sentence's level of grammaticality. The scores should correlate highly with empirically established grammaticality ratings by native speakers. Any discrepancies between these ratings and corpus frequencies are explained in terms of performance factors. Examples are the above strategies of incremental production and of minimizing working memory load.

An important problem incurred by this account concerns *learnability*. Remember that a substantial proportion of argument NP orders that received moderate grammaticality ratings, had zero-frequencies (compare the graphs in Figure 1). But how could a native speaker acquire the syntactic knowledge underlying these structures — knowledge licensing the judgment “not too bad” —, if they have (virtually) never been encountered?

Another account of the grammaticality-frequency gap was recently proposed by Featherston (2005). His Decathlon Model assumes that grammatical encoding is a two-stage process. The first stage generates a set of sentences in accordance with the prevailing grammar rules and constraints. The members of the set are ordered according to quality (in his terminology: “rule violation costs”). The second stage selects the best member — the sentence with lowest violation cost — for overt production (with some allowance for occasional selection errors, when a suboptimal member is chosen).

However, this model is open to two psycholinguistic objections. The calculation of a numerical quality score for a candidate output sentence can only be finalized after this sentence has been encoded *in its entirety*. Hence, the selection of the best candidate can only begin after the first stage of the generation process has encoded one or more *complete* sentential output candidates. This rules out incremental production. Furthermore, from a psycholinguistic viewpoint it is likely that the grammatical encoder quickly zooms in on a single paraphrase (or at most a very small number of paraphrases) to express the communicative intention, and does not elaborate and keep track of the full set of paraphrases offered by lexicon and grammar. For instance, it probably does not compute all six permutations of Subject, Direct and Indirect Object in the midfield of a ditransitive finite clause. Both criticisms can also be leveled at conceivable models based on Optimality Theory, which also presupposes parallel generation of complete, multiple paraphrases.

In view of these shortcomings, we propose a third theoretical alternative. It lends more weight to corpus frequencies as a window on the language user’s competence. The proposal consists of two parts. First, let us assume that at least a *moderate* corpus frequency is needed for a linear order variant to be “taken seriously” by the human grammar induction component, that is, to attain and maintain the status of a more or less stable rule of grammar. (The required frequencies need not be the same in all syntactic domains; see Section 4 below.) Consider the sigmoid curve in Figure 2, which renders one possible hypothesis concerning the relationship between corpus frequency and grammaticality. Low-frequent argument orderings fail to achieve the status of grammar rule and are rejected as ill-formed. Onwards from the lower-left turning point of the curve, structures with moderate corpus frequencies do succeed in being acknowledged by the grammar induction component and thus acquire at least marginal rule status. With rising corpus frequencies, linear order patterns rapidly develop into stable rules, and their tokens are judged as higly grammatical. Rightward of the upper-right turning point, grammaticality ratings approach an asymptotic maximum value. Presumably, both the left- and the right-branching *zu*-complements and the moderately to highly frequent argument NP orderings have achieved the status of stable grammar rules, but not the latter's zero-frequency counterparts.

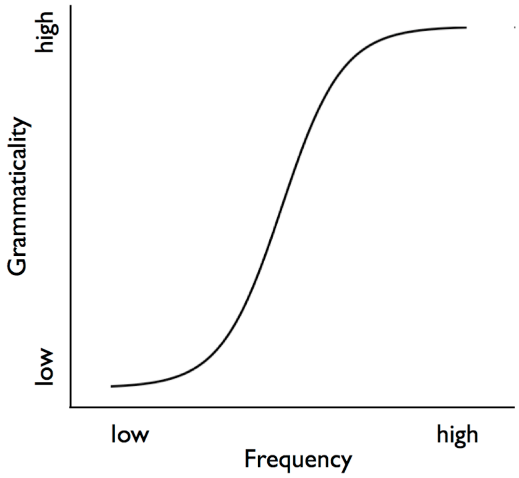

Figure 2. Rated grammaticality of linear order variants as a function of their corpus frequency: A hypothetical mapping.

A recent experiment by Luka and Barsalou (2005) provides direct evidence for the assumption that "mere exposure" to moderately (un)grammatical structures has a positive effect on grammaticality ratings. The effect shows already after one or a small number of exposures to such structures and — crucially — generalizes to exemplars that contain different content words.

The second part of our proposal deals with the following problem: If none of the linear orders in the zero/low-frequency range have been incorporated into the informants' internal grammar, why aren't they all judged to be equally ill-formed? A plausible solution is based on a suggestion by Schütze (1996, p. 177). After an extensive survey of the empirical literature on grammaticality judgments, he suggests:

> "[P]roduction strategies might be involved [in the grammaticality judgment process], in two distinct ways. First, a yes/no judgment might be arrived at by attempting to generate the sentence in question. If our production mechanism cannot do so, then we judge the sentence ungrammatical [...] Second, production strategies might be employed in the scalar rating, locating, explaining, and correcting of errors. Intuitively, it seems plausible that all of these activities involve comparison to a correct or predicted version of the sentence, and so it must be generated somehow. [...] In rating a marginal sentence, for example, one might first extract the intended meaning, then generate a grammatical sen-

> tence that is the expression of that meaning, then compare the two to decide how far off the original sentence was."

The comparison process that Schütze is alluding to in the second part of this quotation, requires a measure of similarity between the to-be-rated original sentence and the "ideal delivery" version self-generated by the informant. In view of Keller's (2000) finding that a high proportion of the variance of the grammaticality rating scores is attributable to two ordering constraints — "pronominal before full NPs" and "Subject before Direct and Indirect Object" — we devised a similarity metric based on "anteriority" (early position) of pronominal and Subject NPs. Ditransitive orderings where these NPs indeed occupy early positions, are therefore more similar to "ideal delivery" than orderings where these constituents have moved to the right. For each member of a family of clauses, we calculated the average "anteriority rank" of the Subject and pronominal NPs in an argument ordering. We predict high negative correlations between the anteriority rank and grammaticality: the lower the anteriority rank of an ordering (that is, Subject and pronominal NPs in early positions), the higher the grammaticality score. Comparison of the resulting mean anteriority scores with the grammaticality ratings for each member of a family of clauses reveals a strong correlation. (See the grammaticality ratings (column 2) and the anteriority ranks (column 8) of Table 1.)

This result may be viewed as a confirmation of Schütze's hypothesis that the grammaticality judgment process may be confounded with extraneous judgment processes and hence do not necessarily mirror properties of the native speaker's internal grammar directly.

4 Conclusion and Discussion

After a sketch of the grammaticality-frequency gap, i.e the discrepancies between the rated grammaticality and the corpus frequency of linear order paraphrases, we proposed an account in terms of a two-factor theory. First, we hypothesized that the grammatical induction component needs a sufficient number of exposures to a syntactic pattern to incorporate it into its repertoire of more or less stable rules of grammar. The moderately to highly frequent argument NP orderings and the left- and the right-branching *zu*-complements are likely have attained this status, but not their zero-frequency counterparts. This is why the latter argument sequences cannot be produced by the grammatical encoder and are absent from the corpora. We suggested that factors related to the cognitive processing load imposed by the different variants (the extent to which they occasion incremental production; early release from working memory of partial structures) are responsible for the different frequencies of the paraphrases. Second, we assumed that an extraneous judgment process biases the ratings of moderately grammatical linear order patterns: Confronted with such structures, the informants produce an "ideal delivery" variant of the to-be-rated target sentence and evaluate the similarity between the two versions. A high similarity score yielded by this judgment then exerts a positive bias on the grammaticality rating — a score that should not be mistaken for an authentic grammaticality rating.

We conclude that, at least in the linearization domain studied here, the goal of gaining a clear view of the internal grammar of language users is best served by a combined strategy in which grammar rules are founded on structures that elicit moderate to high grammaticality ratings *and* attain at least moderate usage frequencies.

At the end of this paper we wish to address three possible objections against the ideas expressed in the foregoing. The most important one pertains to the corpus extraction method we used and calls into question the very existence of the grammaticality-frequency gap. It consists of two questions: Were the corpora big enough to justify the existence of the gap in the argument NP study; and are the ditransitive clauses extracted from the corpora sufficiently similar to the clauses rated by Keller's (2000) informants? Our reply to the former question is simple: The fact that we observed similar frequency patterns in three independent corpora lends extra credibility to the reality of the grammaticality-frequency gap, and we expect the discrepancy to survive larger corpus studies. But we hasten to add that studies targeting other languages and other sets of syntactic paraphrases are needed.

The second question hints at a potential problem having to do with the fact that in Keller's ditransitive clauses all NPs referred to humans (this for good reasons: in order to control for animacy). In the corpora we interrogated, we found hardly any exemplars meeting this criterion; the typical pattern is for the Subject and the Indirect Object to refer to human/animate entities (or groups of such entities) and for the Direct Object to inanimate entities. So, it might be the case that human or animate reference of the Direct Object NP improves the quality of a permutation. This, in turn, could be the reason why zero-frequency argument permutations receive moderate grammaticality ratings. However, we think the objection can be countered as follows. As is well-known, animacy influences the position of NPs in clauses: NPs with animate reference tend to have earlier positions than comparable NPs with inanimate reference. We confirmed this tendency for German in related corpus study (Kempen and Harbusch, 2004). Hence, permutations with early Direct Objects are predicted to have higher grammaticality scores than comparable permutations with late Direct Objects. This prediction is falsified by Table 1, though. We assigned an anteriority score to the Direct Object in the 24 argument patterns, and averaged these scores over the six permutations per clause family. It turns out that within the zero-frequency permutations, there is no tendency for early Direct Objects to attract higher grammaticality ratings than late Direct Objects. If anything, the effect is opposite.

The second issue concerns a difference between the theoretical account presented above, and the one that we proposed recently in Kempen and Harbusch (2005). The two proposals are similar in that they both include the idea of a bias due to an extraneous similarity judgment; they differ with respect to the alleged reason why the moderately grammatical linear orders do not show up in actual usage. In our 2005 paper, we held a grammaticality threshold responsible for the presence or absence of a permutation: Linear orders that do not reach this threshold, so we hypothesized, are rejected by the grammatical encoder and therefore cannot end up in a text corpus. However, Featherston (2005) put forward an empirical argument against this asssumption. He investigated grammaticality-frequency gaps in several different syntactic domains, collecting not only corpus data but also grammaticality ratings (through Magnitude Estimation). He observed (*o.c.*, pp. 200-201) that different grammaticality thresholds appear to be operative in different syntactic domains. This finding is at variance with our earlier proposal. The theory described in the present paper replaces this threshold by a condition on learnability. Structures should be encountered frequently enough by the grammar induction component to make it into the grammar. The probability of being incorporated depends not only on usage frequency but also on other factors, e.g. similarity to competing structures, on the number of such structures, on the complexity of the structure involved, etc. — that is, on any factor capable of improving or worsening the learnability of the structure.

Finally, our theoretical proposal to combine grammaticality judgments with corpus frequency data in order to gain a less biased view of the internal grammar invites a criticism concerning the utilization of complex syntactic structures in designing natural-language grammars. On the one hand, complex structures are rare and therefore will not be considered by the grammarian; hence, linguistic grammars will not be able to generate complex structures. On the other hand, the informant who is presented with complex structures and experiences processing problems, will tend to give low grammaticality scores to these structures despite the fact that the grammar does generate them — thereby creating yet another confound between grammaticality ratings and extraneous judgments (here: difficulty of understanding or producing).

This objection is well-founded and requires a refinement of our proposal. We add the stipulation that the grammarian should only consider grammaticality ratings for structures that clearly do not overtax the sentence processing capacities of typical informants. Notice that this restriction does not entail grammars that only generate finite languages: Judgments by informants about other properties of syntactic structures — e.g. constituency and cohesion — can provide justification for the introduction into the grammar of recursive or iterative devices that enable the generation of infinite languages.

References

Bard, E. G., D. Robertson and A. Sorace (1996). Magnitude estimation of linguistic acceptability. Language, 72, 32–68.

Featherston, S. (2005): The Decathlon Model of empirical syntax. In: S. Kepser and M. Reis (Eds.), Linguistic Evidence — Empirical, Theoretical, and Computational Perspectives. Berlin: Mouton De Gruyter, 187-208.

Gerken, L. and T. Bever (1986): Linguistic intuitions are the result of interaction between perceptual processes and linguistic universals. Cognitive Science, 10, 457-476.

Keller, F. (2000): Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. PhD. Thesis, University of Edinburgh, Edinburgh, UK.

Kempen, G. and K. Harbusch (2003): Word order scrambling as a consequence of incremental sentence production. In H. Härtl and H. Tappe (Eds.), Mediating between concepts and language — Processing structures. Mouton De Gruyter, Berlin, Germany, 141–164.

Kempen, G. & Harbusch, K. (2004). A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment. In: Pechmann, T. & Habel, C. (Eds.), *Multidisciplinary approaches to language production*. Berlin: Mouton De Gruyter. 173-181.

Kempen, G. and K. Harbusch (2005): The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In: S. Kepser and M. Reis (Eds.), Linguistic Evidence — Empirical, Theoretical, and Computational Perspectives. Berlin: Mouton De Gruyter, 329-349.

Luka, B.J. and L.W. Barsalou (2005): Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. Journal of Memory and Language, 52, 436-459.

Müller, G. (1999): Optimality, markedness, and word order in German. Linguistics, 37, 777–815.

Rambow, O. (1994): Formal and Computational Aspects of Natural Language Syntax. PhD. Thesis, University of Pennsylvania, Philadelphia PA.

Schütze, C. (1996): The empirical base of linguistics: Grammaticality judgments and linguistic methodology. Chicago IL: The University of Chicago Press

Seuren, P. (2003): Verb clusters and branching directionality in German and Dutch. In: P. Seuren and G. Kempen (Eds.), Verb constructions in German and Dutch. Amsterdam: Benjamins, 247-296.

Skut, W., B. Krenn, T. Brants and H. Uszkoreit (1997): An annotation scheme for free word order languages. In: Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP), Washington D.C., 27–28.