Markus Bader*

Analyzing free variation with harmony – A case study of verb-cluster serialization

<https://doi.org/10.1515/zfs-2020-2020>

Abstract: In German, a verb selected by another verb normally precedes the selecting verb. Modal verbs in the perfect tense provide an exception to this generalization because they require the perfective auxiliary to occur in cluster-initial position according to prescriptive grammars. Bader and Schmid (2009b) have shown, however, that native speakers accept the auxiliary in all positions except the cluster-final one. Experimental results as well as corpus data indicate that verb cluster serialization is a case of free variation. I discuss how this variation can be accounted for, focusing on two mismatches between acceptability and frequency: First, slight acceptability advantages can turn into strong frequency advantages. Second, syntactic variants with basically zero frequency can still vary substantially in acceptability. These mismatches remain unaccounted for if acceptability is related to frequency on the level of whole sentence structures, as in Stochastic OT (Boersma and Hayes 2001). However, when the acceptability-frequency relationship is modeled on the level of individual weighted constraints, using harmony as link (see Pater 2009, for different harmony based frameworks), the two mismatches follow given appropriate linking assumptions.

Keywords: harmony, weighted constraints, graded acceptability, verb clusters, syntactic alternations

1 Introduction

With the advent of corpus linguistics, the relationship between corpus frequencies and acceptability has become a central topic of linguistic research. Syntactic alternations have played a major role in this research. With regard to the relationship between frequency and acceptability, syntactic alternations provide a mixed picture (Featherston 2005b; Arppe and Järvikivi 2007; Bresnan 2007; Kempen and Harbusch 2008; Bader and Häussler 2010a). On the one hand, it has been found that when an alternative A_n of a given alternation occurs more frequently than an alternative A_m , A_n will be rated at least as acceptable as A_m . On the other

***Corresponding author: Markus Bader**, Department of Linguistics, Goethe-University Frankfurt, Frankfurt, Germany, e-mail: bader@em.uni-frankfurt.de

hand, certain mismatches between corpus frequencies and acceptability ratings have repeatedly been found as well. In particular, corpus frequencies have been shown to decline much steeper than acceptability ratings. This is visible in two ways. First, even if some candidate structure is rated only somewhat worse than the most highly rated structure, its frequency can get rather low in comparison to the frequency of the highest rated structure. Second, candidate structures that are of degraded acceptability are typically not produced at all or with a frequency that approaches zero. Nevertheless, acceptability ratings can still show substantial variations when frequencies are at or near zero.

Following earlier work by Featherston (2005b) and Kempen and Harbusch (2008), I will argue in this paper that such mismatches are not necessarily incompatible with the assumption that acceptability and frequency are related in a systematic way. The syntactic alternation that I will focus on concerns the formation of verb clusters in German, in particular verb clusters of the type illustrated in (1).

- (1) *Ich weiß, dass Peter einen Brief [hat schreiben wollen].*
 I know that P. a letter has write want
 'I know that Peter wanted to write a letter.'

The verb cluster in the embedded clause of (1) shows one out of six possible orders among the three verbs. This is the only order that is grammatical according to prescriptive grammars of Standard German (Dudenredaktion 2009). As will be discussed in more detail in the next section, results from acceptability experiments as well as corpus data show that the Standard German order is judged as most acceptable and occurs with highest frequency. However, although the remaining five orders are all ungrammatical in Standard German, they still differ with regard to acceptability and frequency. Furthermore, verb clusters as in (1) have been shown in prior research to exhibit the frequency-acceptability mismatch discussed above (Bader and Häussler 2010a).

Based on existing experimental results in combination with new corpus data, I will argue that the mismatch between frequency and acceptability disappears if frequency and acceptability are related not on the level of whole sentence structures, but on the level of individual syntactic constraints. More specifically, I will argue that in the case of verb cluster formation, acceptability and frequency can be systematically related to each other by making use of *weighted constraints* and the notion of *harmony* (Pater 2009).

A second issue addressed in this paper concerns the relationship between gradient acceptability ratings and binary grammaticality judgments. Binary grammaticality judgments have been the basic research tool of much of theoretical

syntax. If acceptability is a graded property, as is commonly assumed, an obvious question is how binary grammaticality judgments are related to graded acceptability scores. Answering this question can shed new light on how useful binary judgments are for conducting syntactic research. Furthermore, although explicit binary grammaticality judgments are rarely required from us in everyday language use, sometimes we have to make the same kind of binary decision during language production. This is so when we are not sure – either explicitly or implicitly – whether or not it is licit to use a sentence with a given syntactic structure. As in the case of binary grammaticality judgments, the underlying linguistic intuition is gradient, but the final decision is binary – either you produce the sentence or you produce a different one.

The organization of this paper is as follows. Section 2 introduces the basic empirical findings concerning the acceptability and frequency of German 3-verb clusters as in (1). Based on these findings, Section 2 also presents an informal analysis relating frequency to acceptability. Section 3 extends this analysis to the case of 4-verb clusters. Section 4 discusses further experimental findings that argue that verb cluster serialization is a case of free variation. Section 5 discusses how the acceptability and frequency data introduced in Section 2 and 3 can be related to each other by means of weighted constraints. Section 6 discusses how the notion of grammaticality relates to acceptability and harmony. The paper ends with a general discussion in Section 7.

2 Acceptability and frequency in German 3-verb clusters

The empirical domain of the present investigation is the syntax of verb cluster formation in German (for overviews, see Wurmbrand 2006, Wurmbrand 2017). In German, verbs normally select their dependent elements to the left. This is true for nominal and prepositional objects as in (2), but also for verbs selected by another verb as in (3). Here and in the following, the dependency relations between the verbs are indicated by subscripted numbers. V_1 is the hierarchically highest verb, that is, the verb that is not selected by any other verb. V_1 selects V_2 which in turn selects V_3 , and so on.

- (2) ... *dass Peter [ein Buch ← schreibt₁].*
 that P. a book writes
 ‘... that Peter is writing a book.’

- (3) a. ... *dass er* [*es* ← *geschrieben*₂ ← *hat*₁].
 that he it written has
 ‘... that he has written it.’
- b. ... *dass er* [*es* ← *geschrieben*₃ ← *haben*₂ ← *könnte*₁].
 that he it written have could
 ‘... that he might have written it.’
- c. ... *dass* [*es* ← *geschrieben*₄ ← *worden*₃ ← *sein*₂ ← *könnte*₁].
 that it written been be could
 ‘... that it might have been written.’

The general pattern is given in (4).

- (4) a. $V_2 \leftarrow V_1$
 b. $V_3 \leftarrow V_2 \leftarrow V_1$
 c. $V_4 \leftarrow V_3 \leftarrow V_2 \leftarrow V_1$

Although the patterns in (4) account for the vast majority of verb clusters in German, they are not without exceptions. The most common exception concerns clusters of size three or greater in which V_1 is a perfective auxiliary and V_2 a modal verb. In this case the auxiliary must be fronted to the cluster initial position according to normative grammars of Standard German, as illustrated in (5).

- (5) *dass er es* [*hat*_{Aux₁} → [*schreiben*_{V₃} ← *wollen*_{Mod₂}]].
 that he it has write want
 ‘that he wanted to write it.’

Clusters as in (5) are not only special insofar as they provide an exception to the general verb cluster patterns in (4), but also because they are the locus of a large amount of variation that is not found for the large majority of other clusters. First, a fair amount of variation exists across German dialects and varieties, as illustrated in (6) (see, among others, Weiß 1998, Wurmbrand 2006, and Patocka 1997).

- (6) a. Certain variants of Austrian and Bavarian:
 dass er es [[*schreiben*₃ ← *wollen*₂] ← *hat*₁]. *V₃-Mod₂-Aux₁*
- b. Pattern typical for Austrian and Bavarian:
 dass er es [*schreiben*₃ ← [*hat*₁ → *wollen*₂]]. *V₃-Aux₁-Mod₂*
- c. Standard German:
 dass er es [*hat*₁ → [*schreiben*₃ ← *wollen*₂]]. *Aux₁-V₃-Mod₂*
- d. Pattern typical for Swiss German:
 dass er es [*hat*₁ → [*wollen*₂ → *schreiben*₃]]. *Aux₁-Mod₂-V₃*

Table 1: Classification of 3-verb clusters according to auxiliary position and verb-modal order.

	Aux = 1	Aux = 2	Aux = 3
V < Mod	Aux ₁ -V ₃ -Mod ₂	V ₃ -Aux ₁ -Mod ₂	V ₃ -Mod ₂ -Aux ₁
Mod < V	Aux ₁ -Mod ₂ -V ₃	Mod ₂ -Aux ₁ -V ₃	Mod ₂ -V ₃ -Aux ₁

Furthermore, it has often been reported that dialects usually allow for more than one order. This is in opposition to the rules of Standard German as defined in prescriptive grammars, which consider only the order *Aux₁-V₃-Mod₂* in (5) as grammatical.¹ When this issue was subjected to experimental scrutiny, however, it turned out that this kind of variation is not restricted to dialect speakers. In a series of experiments investigating verb clusters ranging in size from 2 to 5 verbs, Bader and Schmid (2009a,b) and Bader et al. (2009) found that non-dialect speakers of German do not adhere strictly to the Standard German pattern: Native speakers of Colloquial German are more liberal than expected by the standards of prescriptive grammars of Standard German in a precisely defined way, as will be discussed next for the case of 3-verb clusters. Corpus data in support of these experimental data can be found in Krasselt (2013) and Niehaus (2014).

A cluster consisting of a lexical verb, a modal verb and an auxiliary can be serialized in six different ways. In order to ease the following discussion, it is useful to classify the six serializations of such a 3-verb cluster according to two factors, as illustrated by the example in (7) and by Table 1.

- (7) a. ...*dass Peter ein Buch* (₁ *hat* ₁) *lesen* (₂ *hat* ₂) *müssen* (₃ *hat* ₃)' V < Mod
that P. a book has read has must has
- b. ...*dass Peter ein Buch* (₁ *hat* ₁) *müssen* (₂ *hat* ₂) *lesen* (₃ *hat* ₃)' Mod < V
that P. a book has must has read has

First, the lexical verb can either precede or follow the modal verb (*V < Mod* or *Mod < V*). Second, the auxiliary can appear in one of three positions within the verb cluster (*Aux = 1* or *Aux = 2* or *Aux = 3*). Standard German requires the lexical verb to precede the modal verb (*V < Mod*) and the auxiliary to occur in initial position (*Aux = 1*). The expectation based on normative grammar thus is that native speakers of German should judge the order *Aux₁-V₃-Mod₂* as grammatical and the remaining five orders as ungrammatical.

¹ Interestingly, this only holds for 3-verb clusters with a modal verb in the perfect tense, as discussed in the text. For 3-verb clusters with a modal verb in the future tense, two orders are allowed even in Standard German, the auxiliary-initial and the auxiliary-final one.

Several methods are in use for assessing the acceptability of a sentence. Most experimental work on verb cluster formation has made use of speeded grammaticality judgments, an experimental procedure that has originally been used to investigate syntactic ambiguity resolution (e. g., Warner and Glass 1987).² In a speeded grammaticality judgment experiment, participants judge sentences as either grammatical or ungrammatical under controlled and timed conditions. In one variant of the method, sentences are presented word-by-word on a computer screen with a rate similar to normal reading rates, and a response deadline of a few seconds is imposed on participants for giving spontaneous judgments. However, the results do not depend on having limited presentation and judgment times. Grammaticality judgment experiments using paper-and-pencil questionnaires have been shown to yield similar results (Bader and Häussler 2010a). In the following, the term “binary grammaticality judgments (BGJ)” will therefore be used for all experiments asking participants to judge sentences as either grammatical or ungrammatical.

As long as accepted standards of experimental research are respected, data obtained by the BGJ procedure are exempted from many concerns that have been leveled against the use of grammaticality judgments within syntactic research (e. g., Wasow and Arnold 2005). For example, experiments involve a group of participants, with the typical group size ranging from 20 to 60 participants, and the participants are naive with respect to the purpose of the experiment. There is one concern, however, that still applies even when grammaticality judgments are given under experimental conditions. This concern is rooted in the observation that acceptability is not a binary property. Speakers of a language are able to judge sentences not just as grammatical or ungrammatical, but can assign finer grades of acceptability. A popular procedure for obtaining such gradient acceptability ratings is the Magnitude Estimation (ME) procedure.³ ME, which originated in psychophysics (Stevens 1975), was adopted for linguistic purposes by Bard et al.

2 When used in the context of experimental procedures, the term “grammaticality judgment” is used when participants are explicitly instructed to judge the grammaticality of sentences. This is usually done by showing participants uncontroversial grammatical and uncontroversial ungrammatical sentences (e. g., violations of subject-verb agreement). Using the term “grammaticality judgment” does therefore not involve any commitment with regard to the theoretical question of what such judgments measure. This issue is discussed later.

3 An alternative method for measuring acceptability in an experimentally sound way is by having participants rate sentences on a numerical scale, typically ranging from 1–5, 1–7, or 1–9. In the last years, this way of measuring acceptability seems to have replaced magnitude estimation as the most popular method in experimental syntax. For verb clusters, most research has been done using ME, which is why only this procedure is discussed in the current paper.

Figure 1: BGJ results (left), ME results (middle), and corpus results (right) for 3-verb clusters.

(1996) and Cowart (1997a). In an ME experiment, participants evaluate sentences relative to a reference sentence on a continuous numerical scale. First, a reference sentence is presented to which the participant assigns an arbitrary numeric value greater zero. All further items are judged in proportion to the reference item. For example, when a participant considers an experimental sentence as twice as acceptable as the reference sentence, the experimental sentence gets a numerical score that is twice the value of the reference sentence. When an experimental sentence is considered as half as acceptable as the reference sentence, it accordingly gets a numerical score that is half the value of the reference sentence. The numbers assigned to experimental sentences are meaningful only in relation to the value that was initially assigned to the reference sentence. All data points provided by a participant are therefore divided by the participant's reference value and the resulting ratio is log-transformed.

The serialization of verbs within complex verb clusters is among the topics that has been addressed experimentally with both of the experimental procedures just discussed. Figure 1 shows representative results for 3-verb clusters as illustrated in (7) and Table 1 (the exact numerical values can be found in Table 2). The BGJ data are taken from Bader and Schmid (2009b) and the ME data from Bader and Häussler (2010a).

The experimental results shown in Figure 1 can be summarized as follows:

- In all experiments, the Standard German order $Aux_1-V_3-Mod_2$ was judged as most acceptable.
- The partially inverted order $V_3-Aux_1-Mod_2$ was judged only somewhat worse, despite being ungrammatical according to normative grammar.
- The remaining orders were judged as unacceptable, but to different degrees: The order without any inversion $V_3-Mod_2-Aux_1$ and the completely inverted

order $Aux_1-Mod_2-V_3$ were judged better than the other two orders $Mod_2-Aux_1-V_3$ and $Mod_2-V_3-Aux_1$, both of which have the modal precede the lexical verb.

Figure 1 reveals a striking resemblance between the ME and the BGJ results. Despite the fact that the two procedures involve rather different tasks – continuous, numerical ratings with the ME procedure and discrete, binary judgments with the BGJ procedure – when the individual judgments are averaged across participants and sentences, the percentages of sentences judged as grammatical vary in the same continuous way as the mean values obtained with the ME procedure. Both procedures thus seem to deliver the same information (for further discussion, see Bader and Häussler 2010a, and Weskott and Fanselow 2011).

In principle, this means that both procedures can be used if one is interested in assessing the acceptability of some syntactic structure, although specific methodological consideration concerning, e. g., reliability and power, may still favor one or the other method (see Weskott and Fanselow 2011; Sprouse and Almeida 2017; Langsford et al. 2018; Linzen and Oseki 2018). From a theoretical point of view, this raises the question of how speakers map – either consciously or unconsciously – graded acceptability ratings onto binary judgments when required to do so. An answer to this question in form of a threshold model based on signal-detection theory has been proposed by Bader and Häussler (2010a) (see also Dillon and Wagers to appear).

The finding that two of the six possible orders seem to be acceptable for native speakers of German could be an artifact of averaging across the judgment data of individual speakers. For example, one group of speakers could accept the order $Aux_1-V_3-Mod_2$ but reject the order $V_3-Aux_1-Mod_2$ and a second group could show the opposite behavior. Because of the known regional variation concerning verb-cluster formation in German, such a possibility cannot be dismissed a priori. Indeed, for do-support in German main clauses, Bader and Schmid (2006) found a bimodal distribution of grammaticality ratings (see also Weber 2018). Whereas one group of speakers rejected do-support, in agreement with prescriptive grammars of Standard German, another group accepted it, in agreement with the regional variant of the participants. For verb-cluster serialization as discussed above, however, further analyses of the data argue against this possibility. First, Bader and Schmid (2009b) found the basic pattern to be independent of the participants' regional background. Second, as shown by Figure 10 of Bader and Häussler (2010a), the basic pattern found in the average data is also visible in the data of individual speakers.

In addition to the experimental results, Figure 1 presents unpublished corpus data from an ongoing corpus analysis of the deWaC corpus of German internet

texts made available by the University of Bologna (Baroni et al. 2009b) (see the appendix for further information on the corpus analysis). The most striking finding visible in the right diagram in Figure 1 is the extreme dominance of verb clusters of type $Aux_1-V_3-Mod_2$ ($Aux = 1$ and $V < Mod$), that is, verb clusters instantiating the order prescribed for Standard German. Verb clusters of this type make up 96.2% of all analyzed verb clusters. The only other order for which a non-negligible number of instances was found is the partially inverted order $V_3-Aux_1-Mod_2$. Of the remaining four orders, three were not attested at all, and the fourth one occurred with a frequency of less than 0.01%. This was the order $V_3-Mod_2-Aux_1$, which instantiates the general pattern (4) of verb serialization in German according to which a selected verb precedes the verb by which it is selected. Were it not for the special rule involving modal verbs in the perfect tense, this order would be grammatical. This suggests that the occasional instances of $V_3-Mod_2-Aux_1$ verb clusters observed in language production are the result of erroneously over-applying the general rule instead of applying the more specific rule for modal verbs.

In the corpus data presented in Bader and Schmid (2009b), the dominance of the Standard German order $Aux_1-V_3-Mod_2$ was even stronger. With 99.4%, the frequency of this structure was only slightly below 100%. This difference is probably due to socio-linguistic variation concerning the importance given to the rules of prescriptive grammar.⁴ Bader and Schmid (2009b) analyzed a corpus of newspaper texts, for which a high pressure to conform to the rules of prescriptive grammar can be assumed. The deWaC corpus analyzed here, in contrast, contains a large variety of texts, ranging from informal texts like internet chats to formal texts like administrative documents. For texts of the former type, the influence of prescriptive grammar is probably lower than for texts of the latter type. Variation of this kind is clearly something which needs further investigation in the current context.

A comparison of the experimental results with the frequency data reveals a clear instance of the two mismatches discussed in the introduction. First, the frequency distribution is much more skewed than the acceptability distribution. Relatively small variations in acceptability go hand in hand with large variations in frequency. Second, cluster orders with zero or near-zero frequency still differ with regard to acceptability. The question then is how acceptability and frequency

⁴ A further source of variation is dialectal variation. On a gross level, the corpus studies under discussion took this kind of variation into account by looking at the nationality of the newspapers (Bader and Schmid 2009b) or the internet sites (current corpus data; based on the country-code top-level domain “de” for Germany and “at” for Austrian) that went into the analysis. For texts from Austrian, the partially inverted order $V_3-Aux_1-Mod_2$ occurs with a much higher frequency of ca. 30% for newspaper texts and 22% for internet texts, as expected given that this order is considered characteristic for much of Austria (cf. Patocka 1997).

Table 2: Evaluation of 3-verb clusters.

		Order	V < Mod Constraint	Aux < Mod Constraint	Aux-First Constraint	ME	BGJ	Corpus
V<Mod	Aux = 1	1-3-2	√	√	√	.22	86	96.2
	Aux = 2	3-1-2	√	√	*	.09	69	3.8
	Aux = 3	3-2-1	√	*	*	-.12	26	<.1
Mod<V	Aux = 1	1-2-3	*	√	√	-.13	12	0
	Aux = 2	2-1-3	*	*	*	-.25	3	0
	Aux = 3	2-3-1	*	*	*	-.29	2	0

can be systematically related to each other despite these mismatches. Ideally, approaching this question would be based on a full-fledged syntactic analysis of verb clusters. However, the syntactic analysis of verb clusters in the West-Germanic languages is a topic of ongoing research, and several intriguing approaches are currently under debate (see, among others, Abels 2016; Barbiers et al. 2018; Salzmann 2013). A discussion of these approaches is beyond the scope of the current paper. Instead of choosing a particular approach, the following analysis of the acceptability-frequency relationship will make use of a small number of simple surface constraints. The status of these constraints is discussed in Section 6.

The first two constraints correspond to the two factors “order between lexical and modal verb” and “position of the auxiliary” introduced in (7) and Table 1. These two constraints – the *V < Mod Constraint* and the *Aux < Mod Constraint* – are shown in (8).

- (8) a. *The V < Mod Constraint*
The complement of a modal verb precedes the modal verb.
- b. *The Aux < Mod Constraint*
When the perfect auxiliary selects a modal verb, it must precede it.

The application of these two constraints to the six orders available for the 3-verb clusters under consideration is shown in Table 2. In addition, this table shows a further constraint (introduced below) and, for ease of reference, the experimental results and corpus data presented in Figure 1. Comparing constraint violations and acceptability ratings reveals the following picture:

- The two orders $Aux_1-V_3-Mod_2$ and $V_3-Aux_1-Mod_2$ violate neither the *V < Mod Constraint* nor the *Aux < Mod Constraint*. These two orders are judged best.
- The two orders which violate both constraints ($Mod_2-Aux_1-V_3$ and $Mod_2-V_3-Aux_1$) receive very low ME ratings and are almost always judged as ungrammatical in the BGJ task.

- The remaining two orders $V_3\text{-}Mod_2\text{-}Aux_1$ and $Aux_1\text{-}Mod_2\text{-}V_3$, which violate exactly one constraint, receive low ratings, but not as low as those observed for orders violating both constraints.

The two constraints discussed so far do not yet capture the complete pattern of experimental results. Most importantly, they do not capture the finding that of the two orders that neither violate the $V < Mod$ Constraint nor the $Aux < Mod$ Constraint, the order with full inversion of the auxiliary – that is, the Standard German order $Aux_1\text{-}V_3\text{-}Mod_2$ – is judged as more acceptable than the order $V_3\text{-}Aux_1\text{-}Mod_2$ with partial inversion of the auxiliary. This difference can be captured by the Aux-First Constraint given in (9).⁵

(9) *The Aux-First Constraint*

A perfect auxiliary selecting a modal verb must appear in cluster-initial position.

The Aux-First Constraint differs from the $Aux < Mod$ Constraint in that it requires the perfect auxiliary to occur not just in any position in front of the modal verb but in the very first position of the verb cluster. When the Aux-First Constraint is strictly obeyed, the Standard German system with only a single grammatical order results. When less weight is given to the Aux-First Constraint, the alternative order with only partial inversion of the auxiliary becomes a second option too, as in Colloquial German and, even more so, in those varieties of German in which the partially inverted order is used as frequently or even more frequently as the Standard German order (see Niehaus 2014, for corpus data on the frequency of this order in different parts of the German speaking countries).

With regard to the relationship between acceptability and frequency, the major finding for 3-verb clusters was that the frequency distribution is much more skewed than the acceptability distribution. This difference between acceptability and frequency is clearly reflected in Table 2. First, whereas all constraints must be violated in order for acceptability to approach the bottom line, the violation of a single constraint can suffice to bring frequency down to zero. Second, the two highest rated orders differ only with regard to the Aux-First Constraint. This implies that violating this constraint causes a relatively small decrement in accept-

⁵ This constraint could also be applied in such a way that it measures the distance of the auxiliary from the left edge of the verb cluster, with each intervening verb leading to a separate violation mark. Because such counting constraints are highly controversial, we apply the Aux-First Constraint in a strictly binary manner, despite certain indications in the data arguing for a counting interpretation of the constraint..

ability but a large decrement in frequency. 96.2 % of all clusters occurred with the highest-rated order $Aux_1-V_3-Mod_2$, leaving only 3.8 % for the still acceptable order $V_3-Aux_1-Mod_2$.

To sum up, the discussion so far suggests that acceptability and frequency can be insightfully related to each other on the level of individual constraints. The next section considers whether the analysis developed so far extends to clusters of size four. Afterward, it will be shown how the informal analysis can be fleshed out in a formal model of constraint weights.

3 Acceptability and frequency in German 4-verb clusters

The most common type of 4-verb cluster with a modal verb in the perfect tense results when the lexical verb within a 3-verb cluster is put into the passive voice. This is illustrated in (10).

- (10) a. ... *dass jemand die Schubkarre hätte₁ reparieren₃ müssen₂*.
 that someone the wheelbarrow had repair must
 ‘... that someone should have repaired the wheelbarrow.’
 b. ... *dass die Schubkarre hätte₁ repariert₄ werden₃ müssen₂*.
 that the wheelbarrow had repaired be must
 ‘... that the wheelbarrow should have been repaired.’

The four elements of a 4-verb cluster can be ordered in 24 different ways. For practical reasons, only a subset of the 24 possible orders will be considered in the following. All orders in which the passive auxiliary precedes the lexical verb (e. g., *hätte werden repariert müssen*) and all orders in which the modal verb is located between lexical verb and passive auxiliary (e. g., *hätte repariert müssen werden*) are excluded. Given these restrictions, the remaining eight orders can be classified with the same factors used for 3-verb clusters, namely the order of modal verb and passive complex (i. e., lexical verb plus passive auxiliary) and the position of the perfect auxiliary. The resulting classification is shown in Table 3.⁶

Figure 2 shows data from a BGJ experiment (taken from Bader and Schmid 2009b), data from an unpublished ME experiment as well as unpublished corpus

⁶ The eight orders shown in Table 3 are the ones for which experimental evidence is available. The remaining orders all seem to be highly ungrammatical, but this still has to be confirmed experimentally. None of these remaining orders showed up in the corpus.

Table 3: Classification of 4-verb clusters according to auxiliary position and verb-modal order.

	V-Pass < Mod	Mod < V-Pass
Aux = 1	Aux ₁ V ₄ Pass ₃ Mod ₂	Aux ₁ Mod ₂ V ₄ Pass ₃
Aux = 2	V ₄ Aux ₁ Pass ₃ Mod ₂	Mod ₂ Aux ₁ V ₄ Pass ₃
Aux = 3	V ₄ Pass ₃ Aux ₁ Mod ₂	Mod ₂ V ₄ Aux ₁ Pass ₃
Aux = 4	V ₄ Pass ₃ Mod ₂ Aux ₁	Mod ₂ V ₄ Pass ₃ Aux ₁

Figure 2: BGJ results (upper row left), ME results (upper row right) and corpus results (lower row) for 4-verb clusters.

data that were obtained from the deWaC corpus in the same way as the corpus data for 3-verb clusters (see the appendix for details). The picture that shows up is rather similar to the one for 3-verb clusters. The three clusters in which the passive complex precedes the modal verb and the perfect auxiliary occupies one of

Table 4: Evaluation of 4-verb clusters.xxx.

	Order	V < Mod Constraint	Aux < Mod Constraint	Aux-First Constraint	ME	BGJ	Corpus
V<Mod	Aux = 1 1-4-3-2	√	√	√	.26	94	95.2
	Aux = 2 4-1-3-2	√	√	*	.17	88	4.1
	Aux = 3 4-3-1-2	√	√	*	.04	80	0.8
	Aux = 4 4-3-2-1	√	*	*	-.31	14	0
Mod<V	Aux = 1 1-2-4-3	*	√	√	-.08	35	0
	Aux = 2 2-1-4-3	*	*	*	-.34	8	0
	Aux = 3 2-4-1-3	*	*	*	-.43	5	0
	Aux = 4 2-4-3-1	*	*	*	-.45	2	0

the three positions in front of the modal verb receive judgment scores of 80 % or higher, whereas all other orders receive scores of 35 % or lower. In the corpus, only the three highest-rated clusters were found, with a steep decline from the most frequent cluster (95.2 % for Aux₁-V₄-Pass₃-Mod₂) to the second frequent cluster (4.1 % for V₄-Aux₁-Pass₃-Mod₂) and the third-frequent cluster (0.8 % for V₄-Pass₃-Aux₁-Mod₂).

Table 4 shows how the three constraints introduced above apply to the eight 4-verb cluster permutations under discussion. In addition, this table contains the ME, BGJ and corpus data from Figure 2. A similar situation obtains as for 3-verb clusters. First, the three orders that violate neither the V < Mod Constraint nor the Aux < Mod Constraint receive the highest judgment scores. The ranking of these three candidates provides evidence that the position of the perfect auxiliary could be modeled by a gradable constraint. With binary judgments, acceptability decreases from 94 % for Aux = 1 to 88 % for Aux = 2 and then to 80 % for Aux = 3; with magnitude estimation, a similar decrease from .26 to .17 to .04. is observed.⁷ Second, when the V < Mod Constraint and the Aux < Mod Constraint are both violated, clusters receive very low grammaticality ratings, which still decrease with the position of the perfect auxiliary: 8 % for Aux = 2, 5 % for Aux = 3 and 2 % for Aux = 4. Third, of the two orders that violate exactly one of the two constraints V <

7 The drop in acceptability from Aux = 2 to Aux = 3, visible both in the BGJ and the ME data, suggests that the Aux-First Constraint in (9) should be formulated as a counting constraint, as considered in footnote 5. That is, the penalty assigned by this constraint may increase with increasing distance from the cluster-initial position. This issue is complicated by the fact that for 5-verb clusters, Bader et al. (2009) found a decrease in acceptability when going from Aux = 3 to Aux = 4, but not when going from Aux = 2 to Aux = 3. More research is necessary before this issue can be settled.

Mod Constraint and Aux < Mod Constraint, the order Aux₁-Mod₂-V₄-Pass₃ (Aux = 1, Mod < V) lies clearly between the highest- and the lowest-rated orders. The other order, V₄-Pass₃-Mod₂-Aux₁ (V < Mod, Aux = 4), was rated rather low, though still above all orders with three constraint violations. This could be taken as evidence that a violation of the Aux < Mod Constraint affects grammaticality more strongly than a violation of the V < Mod Constraint. However, data for the corresponding 3-verb clusters argue against this conclusion. For 3-verb clusters, the experimental results vary somewhat across experiments, with the Aux < Mod Constraint sometimes having a stronger effect than the V < Mod Constraint (Bader and Häussler 2010a, Experiment 2). This issue needs further research.

With regard to corpus frequencies, we also see a picture similar to the one seen for 3-verb clusters. First, all five orders with at least one violation of either the V < Mod Constraint or the Aux < Mod Constraint were absent from the corpus. Second, all three orders that do not violate these two constraints were found in the corpus, but not with equal frequencies. Quite to the contrary, the order not violating the Aux-First Constraint – which is the Standard German order – occurs with an overwhelming frequency of 95.2% whereas the two orders violating this constraint occur rarely, although they still can be found.

4 Verb cluster serialization – a case of free variation?

Before we proceed to modeling the relationship between frequency and acceptability, one further question has to be addressed: are the different verb-cluster orders that were discussed above associated with different semantic or pragmatic properties, or do they have identical meanings so that they are a case of free variation? Answering this question is crucial for modeling how frequency and acceptability are related. To see why, let us shortly look at another instance of word order variation, namely the variation between subject-object (SO) and object-subject (OS) order in the German midfield. An example illustrating this variation is shown in (11).

- (11) a. *Bestimmt wird der Trainer den Ball mitbringen.*
 surely will the manager the ball bring
 ‘Surely, the manager will bring the ball.’
 b. *Bestimmt wird den Ball der Trainer mitbringen.*
 surely will the ball the manager bring
 ‘Surely, the manager will bring the ball.’

Following the seminal work of Lenerz (1977) and Höhle (1982), much research into German syntax has shown that sentences with SO order can be associated with a broad range of focus structures, including broad focus and focus on either the subject or the object. Sentences with OS order, in contrast, are confined to a focus structure with the subject in focus. One consequence of this difference between SO and OS sentences is that when presented out of the blue, SO sentences are highly acceptable but OS sentences are of degraded acceptability (e.g., Bader and Häussler 2010a). With regard to frequency, it has often been found that SO sentences are way more frequent than OS sentences (e.g., Kempen and Harbusch 2005; Bader and Häussler 2010b). To a large degree, this can be considered a consequence of the different context requirements following from the focus structures associated with the two orders: OS order is licit only in a rather restricted set of contexts whereas SO order can be used almost always. As a consequence, modeling the frequency of SO and OS order without taking context into account would not be meaningful.

For verb clusters, a focus-based order restriction has been proposed by Schmid and Vogel (2004) and Wurmbrand (2004). According to them, V_3 - Aux_1 - Mod_2 clusters are only acceptable with narrow focus on the verb. Experimental investigations of this issue have not borne out this claim. Bader and Schmid (2009b) ran two acceptability experiments with visual sentence presentation which did not reveal any effect of whether the subject, the object or the verb was narrowly focused (see also Sapp 2011). However, because focus is not unambiguously signaled in the sentences under consideration when presented visually for reading, this evidence is not conclusive. For that reason, Bader (2020) ran an experiment in which participants heard pre-recorded sentences and judged them as either grammatical or ungrammatical at the end of the sentence. As illustrated in (12), the experiment varied the position of the auxiliary (Aux_1 - V_3 - Mod_2 vs. V_3 - Aux_1 - Mod_2) and the position of the focus (*object focus* vs. *verb focus*)

(12) a. **Object focus**

Ich weiß, dass Luise sogar die THERAPIE
 I know that L. even the therapy
 {*hat verbieten wollen* | *verbieten hat wollen*}.
 has forbid want forbid has want
 ‘I know that Luise wanted to forbid even the therapy’

b. **Verb focus**

Ich weiß, dass Luise die Therapie
 I know that L. the therapy
sogar {*hat VERBIETEN wollen* | *VERBIETEN hat wollen*}.
 even has forbid want forbid has want
 ‘I know that Luise even wanted to forbid the therapy.’

While the position of auxiliary had the expected effect – V_3 - Aux_1 - Mod_2 clusters were rated as somewhat less acceptable than Aux_1 - V_3 - Mod_2 clusters – the position of the focus had no effect, nor was there an interaction between auxiliary position and focus position. Thus, in contrast to the order of subject and object in the midfield, verb cluster serialization does not seem to be subject to information-structural constraints, at least not in the core cases.⁸

Even if the acceptability of the various verb-cluster orders does not depend on focus structure, focus may still be among the factors that determine which order to produce in particular situations. In order to test for this possibility, Bader (2020) had participants repeat the same auditory stimuli that had before been presented in the acceptability judgment experiment. In research on language production, sentence repetition is known under the name of “production from memory”, an experimental procedure already introduced in the early years of modern psycholinguistics (Mehler 1963) and later used in much work on sentence production and working memory (e. g., Bock and Brewer 1974; Lombardi and Potter 1992; McDonald et al. 1993). A recurrent finding of production-from-memory experiments has been that when participants repeat sentences with some delay between memorization and recall, sentences with a less common structure are repeated with a common structure but not vice versa. For example, passive sentences are often repeated as active sentences whereas active sentences are almost never repeated as passive sentences.

In the production-from-memory experiment of Bader (2020), participants had to solve a simple addition problem between memorization and recall. Sentences with Aux_1 - V_3 - Mod_2 cluster, that is, Standard German clusters, were repeated almost always verbatim, independently of what element was focused. Sentences with V_3 - Aux_1 - Mod_2 cluster, in contrast, showed a strong effect of focus. With object focus, these clusters were changed to Aux_1 - V_3 - Mod_2 clusters in about 90 % of all cases, thus following the general schema that less common structures are repeated as more common structures. With verb focus, only 55 % of all V_3 - Aux_1 - Mod_2 clusters were repeated as Aux_1 - V_3 - Mod_2 , which means that about 45 % of all V_3 - Aux_1 - Mod_2 clusters were repeated verbatim. Thus, in contrast to the acceptability experiment, the production experiment supports the intuition of Schmid and Vogel (2004) and Wurmbrand (2004) that V_3 - Aux_1 - Mod_2 order is closely related to narrow focus on the verb. In a similar vein, Vogel et al. (2015) have shown that

⁸ A reviewer provides an example with *verum* focus as counter example to this claim. In this example, a finite perfective auxiliary contained in a verb-final verb cluster bears main stress. I do not find such a configuration acceptable because for me *verum* focus requires stress on the complementizer. This is an issue that calls for an experimental test.

rhythmic well-formedness has no effect on the acceptability of verb clusters like the ones discussed above whereas the choice of a particular order during language production is sensitive to rhythm.

In sum, with regard to contextual licensing, the situation found for verb clusters is different from the one found for the order of subject and object. For the latter, contextual restrictions of a semantic-pragmatic nature play a major role for both the acceptability and the frequency of the alternative orders. For verb clusters, in contrast, the existing evidence indicates that the alternative orders are in free variation as far as contextual restrictions are concerned. That is, in (almost) all contexts, the set of licit verb orders is only limited by purely syntactic constraints. Within these limits, the order of verbs can be freely chosen. Note that this does not imply that orders are chosen on a purely random basis during language production. As discussed above, focus structure and rhythm have been shown to affect the production probabilities of the competing structures. These influences will be set aside in the following attempt to model the relationship between frequency and acceptability, but they clearly would have to be integrated into a full-fledged model of language production.

5 Harmony as link between acceptability and frequency

Several formal frameworks have been developed for modeling the relationship between acceptability and frequency (see the overview in Manning 2003). For reasons of space, this section considers only a family of approaches that use weighted constraints to capture the frequency-acceptability relationship. On a conceptual level, such approaches share certain architectural features with standard Optimality Theory (OT), and because of the familiarity of OT, it provides a good starting point for the following discussion.

The general architecture of an OT grammar is shown in (13).

The generator GEN takes an input and generates a set of competing candidates. These candidates are evaluated with respect to a set of ranked constraints by the evaluator EVAL. The optimal candidate emerging from this evaluation is the output for the given input. Despite all differences, classical generative theory and Standard OT as developed by Prince and Smolensky (1993/2004) share an im-

portant property: The grammar is restricted to discrete symbols. Constraints are therefore either violated or not, and candidates are either grammatical or ungrammatical, with nothing in between.

While Standard OT only works with discrete symbols, its immediate predecessor Harmonic Grammar was a hybrid model in which symbolic constraints were associated with numeric weights (see Smolensky and Legendre 2006). However, at that time there was not much interest in grammar formalisms with weighted constraints. This changed with the development of Stochastic OT (Hayes 2000; Boersma and Hayes 2001), which was explicitly designed to account for graded judgments of acceptability and the acceptability-frequency relationship.⁹ The major innovation of Stochastic OT was the introduction of a constraint hierarchy with numerical weights. Constraint weights are not used directly for evaluating candidates within Stochastic OT, however. Instead, constraint weights influence the selection of the optimal candidate only indirectly in a two-step process: In the first step, a fixed constraint hierarchy is derived, ranking each constraint according to its weight. Before the constraints are ranked, a certain amount of random noise is added or subtracted from each constraint's weight. The resulting fixed constraint hierarchy can therefore vary somewhat from evaluation to evaluation. In the second step, the fixed constraint hierarchy is used to determine an optimal candidate in the same way as it is done in Standard OT.

Because of the noisy mapping from the weighted constraint hierarchy to the fixed constraint ranking, a constraint with a somewhat lower weight can still end up in a higher position in the final ranking than a constraint with a somewhat higher weight. This means that there will not necessarily be a single candidate winning all the time. Instead, the model imposes a frequency distribution over the candidates, with several candidates having the chance to occur with non-zero frequency. Importantly, the resulting frequency distribution is assumed to determine degrees of acceptability. "Our basic premise, then, is that intermediate well-formedness judgments often result from grammatically encodable patterns in the learning data that are rare, but not vanishingly so, the degree of ill-formedness being related monotonically to the rarity of the pattern." (Boersma and Hayes 2001: 73)

By making well-formedness a function of predicted frequency, Stochastic OT predicts identical acceptability ratings for all candidates that do not differ in terms of frequency. In particular, all candidates that occur with zero frequency are predicted to be judged as equally unacceptable. As the above discussion of the data

⁹ A critical assessment of Stochastic OT in its original form can be found in Keller and Asudeh (2002). For later developments of Stochastic OT, see Boersma (2006) among others.

for 3- and 4-verb clusters has shown, relating frequency and acceptability in this way is not borne out empirically. A major finding was that candidate structures with zero or near-zero frequency can still differ systematically with regard to their acceptability. This argues strongly against theories which tie acceptability to the frequencies of candidates, that is, whole sentences in the case under consideration.

Given that the relationship between acceptability and frequency cannot be captured on the level of candidates, let us turn next to approaches that are similar to Stochastic OT by associating constraints with numerical weights, but differ from Stochastic OT in that they relate frequency to acceptability on the level of individual constraints (see the overview in Pater 2009). This is achieved by using numerical constraint weights directly for evaluating different candidate structures. This is done by assigning each candidate a harmony value. The harmony of a candidate is defined in (14) (from Pater 2009: 1006).

$$(14) \quad \text{Harmony: } H(S) = \sum_{k=1}^K s_k w_k$$

According to (14), the harmony of a candidate S is computed by taking the sum of all products that result when for each constraint k , k 's weight ($= w_k$) is multiplied by the number of violations of k by candidate S ($= s_k$). An often-made assumption is that constraint weights are negative numbers, which means that constraint weights act as penalties. A candidate that does not cause any constraint violation has a harmony of 0. Each constraint violation lowers the harmony by an amount that corresponds to the constraint's weight. The candidate with the highest harmony is sometimes called the winning candidate. In the context of language production, the winning candidate is the one that will be selected for production. Whether the winning candidate also has a special role in the context of the grammar will be discussed later. Similar to Stochastic OT, constraint evaluation based on harmony can be turned into a noisy process by assuming that at evaluation time, the weight of each constraint is perturbed by a small amount of random noise. In this way, a candidate that would otherwise have a lower harmony than the most harmonic candidate can still end up with the highest harmony value. How often this will happen depends both on the distance between the candidates when considered without noise and on the amount of noise added before evaluation.

How can harmony help us in accounting for the relationship between acceptability and frequency? One way to approach this question is by assuming that frequency determines constraint weights which in turn determine acceptability. According to this approach, harmony-based grammars share with Stochastic OT the assumption that weights are learned from input data. In fact, one of the rea-

Table 5: Constraint weights and harmony of 3-verb clusters.

Candidate				V < Mod Constraint w = -10	Aux < Mod Constraint w = -4.43	Aux-First Constraint w = -3.59	Harmony
V<Mod	Aux = 1	1-3-2					0.00
V<Mod	Aux = 2	3-1-2				1	-3.59
V<Mod	Aux = 3	3-2-1			1	1	-8.02
Mod<V	Aux = 1	1-2-3	1				-10.00
Mod<V	Aux = 2	2-1-3	1		1	1	-18.02
Mod<V	Aux = 3	2-3-1	1		1	1	-18.02

sons for the renewed interest in harmony-based grammar was Keller and Asudeh’s (2002) criticism that Stochastic OT’s learning algorithm, the Gradual Learning Algorithm, lacks a formal proof of its convergence properties. Several grammar formalisms working with weighted constraints and the notion of harmony have been proposed partly in reaction to this problem (Maximum Entropy Models: Goldwater and Johnson 2003; Jäger 2007; Jäger and Rosenbach 2006; noisy Harmonic Grammar: Boersma and Pater 2008). These models have in common that they come with sound learning algorithms that enable learning constraint weights from a given frequency distribution of the competing candidates.

The remaining question is how constraint weights are related to gradient acceptability ratings. The simplest answer to this question is provided by Linear OT (LOT) (Keller 2000, 2006). According to LOT, the weight associated with a constraint directly reflects how much the acceptability of a sentence is reduced in case the constraint is violated. Since harmony is the sum of all weighted constraint violations (see the formula in (14)), the acceptability of a sentence, as revealed for example by a method like magnitude estimation, is proportional to its harmony value as long as grammar-external factors do not exert an unduly influence. For example, multiple center-embedding can drive acceptability down even when no grammatical constraint is violated and harmony is therefore not reduced. Conversely, sentences may appear acceptable despite violating one or more grammatical constraints, as in the case of grammatical illusions (Phillips et al. 2011).

In order to assess whether a harmony based model is able to learn constraint weights which in turn can successfully predict graded acceptability, I used the Praat program (Boersma and Weenink 2016) for running a simulation learning the weights of the three constraints discussed above. The input for the simulation was the corpus-derived frequency distribution shown in Table 2. The parameters of the simulation were the default parameters of Praat. The resulting constraint

Figure 3: Comparison between experimental acceptability results (magnitude estimation) and acceptability predicted by harmony, for 3-verb clusters (left) and 4-verb clusters (right).

weights are shown in Table 5 together with the harmony values for the six candidate orders available for a 3-verb cluster. The weight of each constraint is given in the top row. The first candidate $Aux_1-V_3-Mod_2$ ($V < Mod$, $Aux = 1$) does not violate any constraint and so its harmony is 0. The second candidate $V_3-Aux_1-Mod_2$ ($V < Mod$, $Aux = 2$) violates a single constraint, namely the Aux-First Constraint. Since this constraint has a weight of -3.59 , the harmony of this candidate is -3.59 . The third candidate $V_3-Mod_2-Aux_1$ ($V < Mod$, $Aux = 3$) violates two constraints and has a harmony of $-4.43 + -3.59 = -8.02$. The harmony values for the remaining constraints are computed accordingly.

In order to test whether these constraint weights capture the experimental data, Figure 3 plots experimental ME results and harmony values overlaid in a single graph, for both 3-verb and 4-verb clusters.¹⁰ For 4-verb clusters, the very same constraint weights were used as for 3-verb clusters. It would also have been possible to run a learning simulation including 3-verb and 4-verb clusters simultaneously, but because 3-verb clusters are much more frequent than 4-verb clusters, 3-verb clusters would have dominated the resulting weights anyway.

Figure 3 shows a close correspondence between the experimentally obtained acceptability scores and the harmony values learned from the corpus data as explained above (3-verb clusters: $R^2 = .92$; 4-verb clusters: $R^2 = .91$). The tight fit between data and model is especially notable for the simplicity of the model which makes use of just three simple constraints for differentiating between the candidates. Furthermore, model fitting relied completely on default values. It must be

¹⁰ Before plotting, the harmony values were linearly rescaled in order to fit the scale of the ME experiment.

left as a task for future research to remove the remaining discrepancies between data and model by including additional factors affecting frequency and acceptability.

The preceding paragraphs have shown one way how the notion of harmony can provide a systematic link between acceptability and frequency. The direction of information flow was from frequency to harmony and then from harmony to acceptability. Harmony can link acceptability and frequency also in the opposite direction. In fact, constraint weights in LOT as presented in Keller (2000, 2006) are not learned from corpus frequencies but are estimated from acceptability values obtained by means of magnitude estimation. Proceeding in this way was motivated, at least in part, by the assumption that acceptability cannot be derived from frequency (see also Keller and Asudeh 2002). In the case of verb clusters, this argument does not hold, as demonstrated above. Nevertheless, instead of deriving constraint weights from frequency counts, frequency counts could also be derived from constraint weights directly estimated from experimental acceptability values. An implementation of this idea in Praat revealed a frequency distribution that was much more skewed than the empirically observed one; further research will tell whether approaching the frequency-acceptability relationship in this way can be made to work.

6 Relating harmony to grammaticality

This section discusses a question that has been set aside so far: What is the relationship between harmony and grammaticality? Since the notion of grammaticality is used in different ways, several answers to this question are possible. Grammaticality can be understood as the intuition that lets people judge sentences as either grammatical or not – for example, as participants in experiments requiring binary judgments, or as linguists assigning asterisks to some sentences but not to others (see also Luka 2005). Grammaticality in this sense – which is called “perceived grammaticality” by some authors (e. g., Cowart 1997b; Featherston 2005a) – is related to harmony in the same way as it is related to acceptability. As could be seen in Figures 1 and 2, grammaticality and acceptability show the same pattern. Sentences that are rated as highly acceptable are judged as grammatical most of the time, sentences that are rated as highly unacceptable are most of the time rejected as ungrammatical, and sentences of medium acceptability are sometimes judged as grammatical, sometimes as ungrammatical. As pointed out above, by assuming that all constraint weights are negative, constraint weights act as penalties – each violation of a constraint weight reduces the harmony of a candidate

and thereby drives acceptability down under the assumption that harmony is proportional to acceptability. In a similar way, with decreasing harmony, the probability increases that a sentence will be judged as ungrammatical, but there is no natural turning point dividing sentences as either grammatical or ungrammatical. Thus, intuitions of grammaticality are continuous in the same way as intuitions of acceptability, and there is no strict division into grammatical and ungrammatical sentences. In fact, some authors use the terms “acceptability” and “grammaticality” interchangeably (e. g., Schütze 1996; Luka 2005).

According to another use of the term, grammaticality is a property assigned to sentences by a formal grammar. In the simplest case, a grammar divides the set of all sentences (understood as all strings over a given alphabet) into two subsets – the grammatical sentences and the ungrammatical sentences (e. g., Chomsky 1957; Partee et al. 1990). Grammaticality in this sense is a main determinant of sentence acceptability. Ungrammatical sentences in most cases lead to reduced acceptability, although some exceptions known as grammatical illusions exist (Phillips et al. 2011). However, grammaticality is not the only determinant of acceptability. Even for grammatical sentences, acceptability may be reduced, for example because of high complexity, syntactic ambiguity, or semantic implausibility. Furthermore, familiarity has been shown to modulate acceptability (Luka and Barsalou 2005).

Grammaticality as a property of sentences is a central topic of syntactic research. Given the large range of competing syntactic theories (for a recent overview, see Müller 2019), the following discussion can offer no more than some preliminary remarks on how the notion of harmony as used above relates to grammaticality as defined by the grammar. In a standard OT grammar, the division of sentences into grammatical and ungrammatical sentences is achieved by having the evaluator EVAL determine a single optimal candidate, as shown in (13). Harmony can be put to use in the same way. As explained in detail in Pater (2009), the role of the optimal candidate in an OT grammar is taken over by the candidate with the highest harmony – the winning candidate – in Harmonic Grammar. Using harmony to determine a single winning candidate may well be adequate in certain parts of grammar – for example, for mapping input forms to output forms in phonology – but it does not match the role that harmony played in our attempt to link frequency and acceptability. After all, the main reason for invoking harmony was the observation that acceptability was not distributed in a binary way across the possible verb cluster serializations, with one serialization being acceptable and all others being unacceptable. Instead, acceptability declined from highly acceptable orders to highly unacceptable orders, with several intermediate values.

In the analysis presented above, the harmony value of each candidate was directly related to its acceptability score, without invoking the concept of a winning candidate. Making use of harmony in this way is more akin to Maximum Entropy Models (Goldwater and Johnson 2003; Jäger 2007; Jäger and Rosenbach 2006) or the Decathlon Model (Featherston 2005b) than to Harmonic Grammar (Pater 2009). Furthermore, the analysis was based on a set of simple surface constraints that were not intended to compete with current syntactic analyses of verb-cluster formation (e. g. Abels 2016; Barbiers et al. 2018; Salzmann 2013). In principle, the proposed surface constraints are not necessarily incompatible with these analyses (whichever may turn out to be correct), in the sense that such surface constraints may mediate between participants' intuitions and the syntactic structures imposed by the grammar. That is, participants may parse a sentence using the means provided by the grammar and still apply simple surface constraints when asked to rate the sentence's well-formedness. The resulting acceptability rating may then reflect the number and weight of the violated constraints. If so, the surface constraints proposed above would complement a full-fledged syntactic analysis without providing such an analysis by themselves. Future research must show whether an account along these lines provides a valid account of linguists' intuitions.

7 Conclusion

This paper has shown how the notion of harmony can be used to model the relationship between acceptability and corpus frequencies, even in the face of certain mismatches between acceptability and frequency on the level of whole sentence structures. Based on a set of weighted constraints, harmony provides a continuous measure of constraint violations. This measure can be used to link frequency and acceptability, but for this it is crucial that harmony is related to frequency and acceptability in different ways. The relationship between harmony and frequency was assumed to be non-linear whereas a linear relationship was assumed between harmony and acceptability. Due to these assumptions, frequency and acceptability ended up being related to each other in a non-linear way, which made it possible to account for the frequency-acceptability mismatches that have been observed. First, because frequency declines much more steeply than acceptability, a relatively small acceptability difference can translate into a large frequency difference. Second, and relatedly, below a certain value of harmony frequency becomes indistinguishable from zero but substantial acceptability differences are still possible. This allows for graded acceptability distinctions between candidates that all occur with zero frequency.

A striking illustration of the mismatch between frequency and acceptability comes from verb clusters that are even more complex than the clusters considered so far. In contrast to verb clusters of size three or four, verb clusters of size five are extremely rare, as witnessed by the fact that neither a search in the Tiger Corpus nor a search in the deWaC corpus delivered any instances.¹¹ Only a search of the complete internet using Google showed that 5-verb clusters are indeed produced from time to time. Two authentic examples are given below – one with the auxiliary in first position (15) and one with the auxiliary in second position (16).

- (15) ... was alles besser hätte₁ gemacht₅ worden₄ sein₃ können₂
 what all better had made been be can
 ‘... what could have been done better’
 (www.dradio.de/dkulturr/sendungen/fazit/2028303/)
- (16) ... dass der Film auch sehr gut von den Großen des alten Hollywood
 that the movie also very well by the big-ones of-the old Hollywood
 gedreht₅ hätte₁ worden₄ sein₃ können₂.
 made have been be can
 ‘... that the movie could very well have been made by one of Hollywood’s
 big figures.’
 (schwarzmarkt.blog.de/2011/01/02/filmkritik-tourist-10290065/)

Despite the extreme rareness of 5-verb clusters, native speakers of German make sharp distinctions with regard to which orders are allowed. This was shown by Bader et al. (2009) using a timed BGJ procedure. The results of this study, in which the V < Mod Constraint was always observed, show a striking contrast between the four orders that obey the Aux < Mod Constraint (aux positions 1-4) and the one order that violates this constraint (aux position 5). When the auxiliary preceded the modal verb, mean grammaticality ranged from 54 % to 78 %. When the auxiliary followed the modal verb, mean grammaticality dropped to a low 6 %. If grammaticality and frequency were related to each other on a holistic level, these results would be hard to explain given the extreme rarity of 5-verb clusters. On the other hand, this is exactly the pattern that is expected when the constraint weights set on the basis of 3-verb clusters are applied to clusters of size 5. Thus, decomposing

¹¹ At least for size 6, verb clusters can be constructed that are still comprehensible. Whether clusters of this size are ever produced is an open question. An example of this kind is given in (i).

- (i) dass dem Kind hätte abgeholt zu werden angeboten werden müssen
 that the child had picked-up to be offered been must
 ‘that one should have offered the child to be picked up.’

verb clusters in the way proposed above clearly pays off even for very rare verb clusters.

To conclude, this paper has presented a case study exploring different ways how the notion of harmony, which provides a measure of constraint violations in a system of weighted constraints, can be used to model the relationship between acceptability and corpus frequencies. It remains a task for future research to investigate which – if any – of the proposed models can be successfully extended to other domains of grammar, in particular to those domains that give rise to acceptability-frequency mismatches.

Appendix A. Details of corpus analysis

The corpus analyzed in this paper is the deWaC corpus made available by the University of Bologna (see Baroni et al. 2009a and <http://wacky.sslmit.unibo.it>). The deWaC corpus is a huge part-of-speech (POS) tagged and lemmatized corpus of written German built by web crawling. It contains about 1,600,000,000 tokens of text in ca. 92,000,000 sentences. Based on the POS tag associated with each word, the deWaC corpus was searched for complementizer-introduced verb-final clauses ending in a verb cluster with either three or four verbs at least one of which had to be a modal verb. For 3-verb clusters, the analysis was restricted to sentences introduced by the complementizers *dass* ‘that’, *wenn* ‘if’, *ob* ‘whether’, and *nachdem* ‘after’. Because of the lower number of 4-verb clusters, all sentences were included in the analysis. For each corpus hit, the country-code top-level domain of the source website was recorded. The large majority of the websites had the top-level domain “de” for Germany, but websites with “at” for Austrian also occurred in sufficient numbers to yield analyzable results.

Table A1 shows the corpus results for 3-verb clusters and Table A2 the results for 4-verb clusters. Verb orders that were not found are not included in the tables.

Table A1: Frequency and proportions of 3-verb clusters in the deWaC corpus.

Order	All websites		.de websites		.at websites	
	Frequency	Proportion	Frequency	Proportion	Frequency	Proportion
<i>Aux₁-V₃-Mod₂</i>	6885	0.9615	6640	0.9706	228	0.7525
<i>V₃-Aux₁-Mod₂</i>	273	0.0381	198	0.0289	75	0.2475
<i>V₃-Mod₂-Aux₁</i>	3	0.0004	3	0.0004	0	0.0000

Table A2: Frequency and proportions of 4-verb clusters in the deWaC corpus.

Order	All websites		.de websites		.at websites	
	Frequency	Proportion	Frequency	Proportion	Frequency	Proportion
<i>Aux₁-V₄-Pass₃-Mod₂</i>	2909	0.9516	2839	0.9594	46	0.6301
<i>V₄-Aux₁-Pass₃-Mod₂</i>	125	0.0409	103	0.0348	21	0.2877
<i>V₄-Pass₃-Aux₁-Mod₂</i>	23	0.0075	17	0.0057	6	0.0822

References

Abels, Klaus. 2016. The fundamental left–right asymmetry in the Germanic verb cluster. *The Journal of Comparative Germanic Linguistics* 19(3). 179–220.

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.

Bader, Markus. 2020. Focus and auxiliary inversion in German three verb clusters. Ms., Goethe-Universität Frankfurt.

Bader, Markus & Jana Häussler. 2010a. Toward a model of grammaticality judgments. *Journal of Linguistics* 46(2). 273–330.

Bader, Markus & Jana Häussler. 2010b. Word order in German: A corpus study. *Lingua* 120(3). 717–762.

Bader, Markus & Tanja Schmid. 2006. An OT-analysis of do-support in Modern German. Ms., Rutgers Optimality Archive, Nr. 837-0606.

Bader, Markus & Tanja Schmid. 2009a. CAT meets GO: 2-verb clusters in German. In Jeroen van Cranenbroeck (ed.), *Alternatives to cartography*, 203–245. Berlin & New York: Mouton de Gruyter.

Bader, Markus & Tanja Schmid. 2009b. Verb clusters in Colloquial German. *The Journal of Comparative Germanic Linguistics* 12(3). 175–228.

Bader, Markus, Tanja Schmid & Jana Häussler. 2009. Optionality in verb-cluster formation. In Susanne Winkler & Sam Featherston (eds.), *The fruits of empirical linguistics. Volume 2: Product*, 37–58. Berlin & New York: de Gruyter.

Barbiers, Sjef, Hans Bennis & Lotte Dros-Hendriks. 2018. Merging verb cluster variation. *Linguistic Variation* 18(1). 144–196.

Bard, Ellen Gurman, Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1). 32–68.

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

Baroni, Marco, Emiliano Guevera & Roberto Zamparelli. 2009b. The dual nature of Deverbal Nominal Constructions: Evidence from acceptability ratings and corpus analysis. *Corpus Linguistics and Linguistic Theory* 5. 27–60.

Bock, J. Kathryn & William F. Brewer. 1974. Reconstructive recall in sentences with alternative surface structures. *Journal of Experimental Psychology* 103(5). 837.

Boersma, Paul. 2006. Prototypicality judgements as inverted perception. In Gisbert Fanselow,

- Caroline Féry, Ralf Vogel & Matthias Schlesewsky (eds.), *Gradience in grammar: Generative perspectives*, 167–184. New York: Oxford University Press.
- Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32. 45–86.
- Boersma, Paul & Joe Pater. 2008. Convergence properties of a gradual learning algorithm for Harmonic Grammar. Ms., University of Massachusetts at Amherst (available at <http://roa.rutgers.edu/>).
- Boersma, Paul & David Weenink. 2016. Praat: doing phonetics by computer [Computer program]. Version 6.0.14. Retrieved 11 February 2016 from <http://www.praat.org/>.
- Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 75–96. Berlin & New York: Mouton de Gruyter.
- Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.
- Cowart, Wayne. 1997a. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage Publications.
- Cowart, Wayne. 1997b. Review of Schütze, Carson: *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press, 1996. *Language and Speech* 40(1). 91–95.
- Dillon, Brian & Matthew Wagers. to appear. Approaching gradience in acceptability with the tools of signal detection theory. In Grant Goodall (ed.), *Cambridge handbook of experimental syntax*, Cambridge: Cambridge University Press.
- Dudenredaktion. 2009. *Duden – die deutsche Rechtschreibung* (Duden Band 1). Hg. v. d. Dudenredaktion. 25., völlig neu bearb. u. erw. Aufl. Mannheim: Dudenverlag.
- Featherston, Sam. 2005a. *That*-trace in German. *Lingua* 115(9). 1277–1302.
- Featherston, Sam. 2005b. The decathlon model of empirical syntax. In Marga Reis & Stephan Kepser (eds.), *Linguistic evidence. Empirical, theoretical and computational perspectives*, 187–208. Berlin & New York: de Gruyter.
- 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. Stockholm: University of Stockholm.
- Hayes, Bruce. 2000. Gradient well-formedness in optimality theory. In Joost Dekkers, Frank van der Leeuw & Jeroen van de Weijer (eds.), *Optimality Theory: Phonology, syntax, and acquisition*, 88–120. Oxford: Oxford University Press.
- 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.
- 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 Publications.
- Jäger, Gerhard & Anette Rosenbach. 2006. The winner takes it all – almost: Cumulativity in grammatical variation. *Linguistics* 44(5). 937–971.
- Keller, Frank. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Edinburgh: 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.
- Keller, Frank & Ash Asudeh. 2002. Probabilistic learning algorithms and optimality theory. *Linguistic Inquiry* 33. 225–244.
- 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 & New York: de Gruyter.
- Kempen, Gerard & Karin Harbusch. 2008. Comparing linguistic judgments and corpus frequencies as windows on grammatical competence: A study of argument linearization in German clauses. In Anita Steube (ed.), *The discourse potential of underspecified structures*, 179–192. Berlin & New York: de Gruyter.
- Krasselt, Julia. 2013. Zur Serialisierung im Verbalkomplex subordinierter Sätze. Gegenwartssprachliche und frühneuhochdeutsche Variation. *Jahrbuch für germanistische Sprachgeschichte* 4. 128–143.
- Langsford, Steven, Amy Perfors, Andrew T. Hendrickson, Lauren A. Kennedy & Danielle J. Navarro. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: A Journal of General Linguistics* 3(1). 1–34. <https://doi.org/10.5334/gjgl.396>.
- Lenerz, Jürgen. 1977. *Zur Abfolge nominaler Satzglieder im Deutschen*. Tübingen: Narr.
- Linzen, Tal & Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: A Journal of General Linguistics* 3(1). 1–25. <https://doi.org/10.5334/gjgl.528>.
- Lombardi, Linda & Mary C. Potter. 1992. The regeneration of syntax in short term memory. *Journal of Memory and Language* 31. 713–733.
- Luka, Barbara J. 2005. A cognitively plausible model of linguistic intuitions. In Salikoko Mufwene, Elaine Francis & Rebecca Wheeler (eds.), *Polymorphous linguistics: Jim McCawley's legacy*, 479–502. Cambridge, MA: MIT Press.
- Luka, Barbara J. & Lawrence W. Barsalou. 2005. Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language* 52(3). 436–459.
- Manning, Christopher D. 2003. Probabilistic syntax. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 289–341. Cambridge, MA: MIT Press.
- McDonald, Janet L., Kathryn Bock & Michael H. Kelly. 1993. Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology* 25. 188–230.
- Mehler, Jacques. 1963. Some effects of grammatical transformations on the recall of English sentences. *Journal of Verbal Learning and Verbal Behavior* 2(4). 346–351.
- Müller, Stefan. 2019. *Grammatical theory: From transformational grammar to constraint-based approaches*. Berlin: Language Science Press. 10.5281/zenodo.3364215.
- Niehaus, Konstantin. 2014. Kontinuität im Neuhochdeutschen ‘von unten’ und ‘von oben’. Ein variationslinguistisches Nutzungsszenario. *Jahrbuch für germanistische Sprachgeschichte* 5. 299–313.
- Partee, Barbara H., Alice ter Meulen & Robert E. Wall. 1990. *Mathematical methods in linguistics*. Dordrecht: Kluwer.
- Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33. 999–1035.
- Patocka, Franz. 1997. *Satzgliedstellung in den Bairischen Dialekten Österreichs*. Frankfurt a. M.: Peter Lang.

- 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.
- Salzmann, Martin. 2013. New arguments for verb cluster formation at PF and a right-branching VP: Evidence from verb doubling and cluster penetrability. *Linguistic Variation* 13(1). 81–132.
- Sapp, Christopher D. 2011. *The verbal complex in subordinate clauses from Medieval to Modern German*. Amsterdam: Benjamins.
- Schmid, Tanja & Ralf Vogel. 2004. Dialectal variation in German 3-verb clusters. *Journal of Comparative Germanic Linguistics* 7. 235–274.
- Schütze, Carson T. 1996. *The empirical base of linguistics*. Chicago, IL: University of Chicago Press.
- Smolensky, Paul & Géraldine Legendre. 2006. *The harmonic mind: From neural computation to optimality-theoretic grammar (2 Volumes)*. Cambridge, MA: MIT Press.
- Sprouse, Jon & Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: A Journal of General Linguistics* 2(1). 1–32.
<https://doi.org/10.5334/gjgl.236>.
- Stevens, Stanley S. 1975. *Psychophysics. Introduction to its perceptual, neural, and social prospects*. New York: John Wiley.
- Vogel, Ralf, Ruben van de Vijver, Sonja Kotz, Anna Kutscher & Petra Wagner. 2015. Function words in rhythmic optimisation. In Ralf Vogel & Ruben van de Vijver (eds.), *Rhythm in cognition and grammar: A Germanic perspective*, 253–274. Berlin & Boston: de Gruyter.
- Warner, John & Arnold L. Glass. 1987. Context and distance-to-disambiguation effects in ambiguity resolution: Evidence from grammaticality judgements of garden path sentences. *Journal of Memory and Language* 26(6). 714–738.
- Wasow, Thomas & Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115. 1481–1496.
- Weber, Thilo. 2018. An OT analysis of do-support across varieties of German. *The Journal of Comparative Germanic Linguistics* 21(1). 75–129.
- Weiß, Helmut. 1998. *Syntax des Bairischen. Studien zur Grammatik einer natürlichen Sprache*. Tübingen: Niemeyer.
- Weskott, Thomas & Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87(2). 249–273.
- Wurmbrand, Susi. 2004. Syntactic vs. post-syntactic movement. In Sophie Burelle & Stanca Somesfalean (eds.), *Proceedings of the 2003 Annual Meeting of the Canadian Linguistic Association (CLA)*, 284–295. Toronto: Canadian Linguistic Association.
- Wurmbrand, Susi. 2006. Verb clusters, verb raising, and restructuring. In Martin Everaert & Henk van Riemsdijk (eds.), *The Blackwell companion to syntax*, vol. 5, 229–343. Oxford: Blackwell.
- Wurmbrand, Susi. 2017. Verb clusters, verb raising, and restructuring. In Martin Everaert & Henk van Riemsdijk (eds.), *The Blackwell companion to syntax, 2nd edition*, vol. 5, 1–109. Oxford: Blackwell.