Word Order Acceptability and Word Order Choice

Elisabeth Verhoeven & Anne Temme
Humboldt University of Berlin / University of Stuttgart
verhoeve@cms.hu-berlin.de, anne.temme@posteo.de

1 Preliminaries

The relation between acceptability judgments and corpus frequencies has been the subject of a paradigm of empirical studies. That these types of empirical data correlate is an established fact, reported for a range of phenomena, e.g., alternative syntactic constructions (Bresnan 2007), word order permutations (Kempen & Harbusch 2005; Bader & Häussler 2010a; Adli 2010), usage of synonym verbs (Arppe & Järvikivi 2007), object coreference (Featherston 2005), linearization in verb clusters (Bader & Häussler 2010a), etc. The correlation between the two data types opens two relevant research questions that are outlined in (1). The challenge arising from (1) is to specify the function that relates one data type to the other and to figure out which sources of variation are involved in the mapping between them. Beyond the methodological interest of this issue, it promises to contribute to the understanding of the relation between grammatical competence and speakers' choices in speech production.

(1) a. corpus → acceptability
How do frequencies in corpus shape speakers' intuitions of acceptability?

b. acceptability → corpus
How does the acceptability of particular structures influence speakers' decisions in speech production?

The present study deals with the question in (1b). The aim is to understand the function relating the acceptability of an expression with reference to a set of alternative expressions in a particular context to the corresponding probability of this expression to occur in that context, as outlined in Fig. 1. Frequencies in corpus data are the result of repeated choices that native speakers have made in the context of interest. With each choice, the speaker evaluates a *set of alternative expressions* that may be suitable for expressing the intended propositional content. Based on her linguistic competence, the speaker can estimate the *context set*, i.e., the range of potential contexts that license the morphosyntactic features of each expression at issue. The task in speech production is to evaluate the felicity of the alternative expressions with respect to a *given context*. This evaluation involves gradience, since contexts are often complex and the speaker's intentions with regard to prioritizing one or the other contextual property and its consequences for the choice of expression may vary. This process results in the selection of the optimal candidate, i.e., the member of the set of alternative expressions associated with the highest acceptability value in a given context.

If there would be no further sources of variation involved, then the member E_i of the set of alternative expressions that reaches the highest value in the set of acceptability judgments $\alpha_1 \ldots \alpha_n$ in a context type C would always win the competition in this context type, i.e., its conditional probability in this context type would be $p(E_i|C) = 1$. However, this is not what the empirical studies mentioned above report. At least in a subset of the reported phenomena, some variation between alternative expressions is attested (see, e.g., Featherston 2005), i.e., an error term should be included in the model. Note that this does not imply that speakers' competence is probabilistic: a part of the attested variation in the choice of constructions can certainly be explained by various factors not yet understood (including factors such as the linguistic properties of contexts, the lexicalization of particular structures and the variation between speakers). Furthermore, variation may be a stylistic choice (a means for drawing attention to the speech).

acceptability values (a) of expressions $E_1 \ldots E_n$ in context C

probabilities (p) of expressions $E_1 \ldots E_n$ in context C

$$\begin{array}{c} \text{given a context } C \\ \left\{ \begin{array}{l} a(E_1|C) \rightarrow \alpha_1 \\ a(E_2|C) \rightarrow \alpha_2 \\ \ldots \\ a(E_n|C) \rightarrow \alpha_n \end{array} \right\} \end{array} \rightarrow \left\{ \begin{array}{l} p(E_1|C) = p_1 \\ p(E_2|C) = p_2 \\ \ldots \\ p(E_n|C) = p_n \end{array} \right\}$$

Fig. 1: Mapping acceptability and choice

Previous empirical studies established two important properties. Firstly, the relation between acceptability judgments and probabilities in corpus is not linear. The highest ranked alternative expressions predominate in the corpus, while less acceptable alternatives may be not attested at all (Featherston 2005; Kempen & Harbusch 2005; Bader & Häussler 2010a). This asymmetry suggests a power-law distribution, which is the expected outcome of repeated choices and is traced back to the effect of competition (Featherston 2005): the winner candidates are selected consistently, while less optimal candidates – although potentially felicitous choices – only occur scarcely. The comparison between acceptability and forced-choice data confirms that the asymmetry is due to the selection process: contrasts in the intuition of felicity of alternative constructions are strengthened in forced-choice data (see Arppe & Järvikivi 2007).[1]

Secondly, there is an asymmetry between the contextual conditions that determine variation: some contexts involve variation, while other contexts categorically

[1] Further studies such as Bader & Häussler (2010a) or Weskott & Fanselow (2011) compare binary and scalar acceptability (based on magnitude estimation or 7-point-scales) with different results, which are due to the fact that their binary data does not stem from a forced choice between two alternatives but rather involves a binary decision of acceptability (grammaticality) of a given structure.

determine the choice of particular constructions. This observation is obtained by forced-choice data (Rosenbach 2003) and is confirmed in acceptability and corpus data (Bresnan 2007).

The aim of the present study is to focus on the processes that determine the selection between alternative expressions. We assume that at some point of this process, the speaker compares a set of alternative expressions and judges their relative appropriateness in a particular context. The winner candidate is the output-selection. Our aim is to examine the stage of comparison and the stage of selection with the same lexicalizations and the same contextual manipulations. For this purpose, we designed an experiment collecting relative acceptability judgments (split-100 rating) and a forced-choice experiment with the same material.

The phenomenon examined in the studies that will be compared is the choice between SO and OS order in German clauses with canonical and experiencer-object verbs in different contexts, to be outlined in Section 2. Section 3 introduces the empirical design and our predictions for the experimental results both with respect to the grammatical phenomenon at issue as well as regarding the two methods applied. First, we conducted a split-100 rating experiment that aims to reflect the process of a speaker evaluating a set of alternative orders in a given context; see Experiment 1 in Section 4. In order to assess the outcome of the selection process, we used a forced-choice task, i.e. a *forced selection between two competing structures* (e.g., Rosenbach 2003; Bresnan 2007); see Experiment 2 in Section 5. The obtained results confirm the previous observation that differences in preference are strengthened in forced-choice data. Furthermore, they show an exponential relation between the preference for a particular expression in ratings and the likelihood of the selection of this expression in forced choice; see Section 6.

2 German Word Order

German is a language with flexible word order sensitive to information structure. Being a basic OV language it allows for scrambling objects over the subject (see Fanselow 2003; Müller 2004; Frey 2004, 2005; for corpus evidence see Bader & Häussler 2010b). Scrambling can be triggered by interaction of several factors, including definiteness, animacy, focus, case, etc. (e.g., Webelhuth 1995; Müller 1999, 2004). Main declarative clauses are verb-second as a result of an obligatory rule for fronting finite verbs to a higher clausal position (Thiersch 1978; Den Besten 1989). The position preceding the verb in verb-second clauses, that is, the prefield, is obligatorily filled, either as the result of A-bar movement to Spec, CP, indicating a contrastive interpretation of the moved material or – in the absence of a trigger for A-bar movement – by formal movement of the first eligible element in the middlefield. This element is the leftmost one, i.e., the subject constituent or a constituent scrambled past the subject. Assuming that the operation which displaces this constituent to the prefield is purely formal (that is, semantically vacuous), it does not involve any semantic or pragmatic features in addition to those that possibly triggered the scrambling of the highest middlefield constituent (Frey 2006). Thus, OS order may result from A-bar movement of the object or formal movement if the object is scrambled in the middlefield. In both cases, OS order is subject to specific licensing conditions triggering the fronting.

Furthermore, certain verb types are associated with non-canonical argument orders in German, as, e.g., experiencer-object verbs, which, as a number of studies have argued, license object-before-subject order without further contextual licensing. Experimental results on this issue differ depending on the object case of the experiencer. Haupt et al. (2008: 84) – confirming earlier observations by Lenerz (1977), Hoberg (1981), Primus (2004) – show for example on the basis of a single-item-rating study (outbalancing the factors definiteness and animacy) an advantage for ‘dative experiencer ≺ nominative stimulus’ but no overall word order preference for constructions featuring an accusative experiencer and a nominative stimulus.[2] Similarly, a corpus study revealed that in the middlefield dative and accusative experiencers differ in their linearization properties: while dative experiencers precede nominative stimulus arguments independently of animacy more often than not (80% with disharmonic animacy; 50% otherwise), the early realization of accusative experiencers is mainly restricted to a disharmonic animacy configuration, i.e. to cases where the stimulus is inanimate (65% with disharmonic animacy; 4% otherwise) (Verhoeven 2015: 76). If one of the arguments was realized in the prefield, the early occurrence of the experiencer was less frequent: dative experiencers preceded nominative stimulus arguments in 33% of the cases, while accusative experiencers preceded nominative stimulus arguments in 17% of the cases. In a speech production experiment, accusative experiencers preceded the nominative stimulus argument in 10% of the cases (Verhoeven 2014). Thus, all in all, there is evidence for an early linearization of the accusative experiencer which is, however, less consistent than for dative experiencers. Furthermore, the early occurrence of accusative experiencers seems to depend rather strongly on a disharmonic animacy configuration.

The early linearization of experiencers, also termed experiencer-first effect, has been related to the topic-worthiness of the experiencer (see Haspelmath 2001; Bickel 2004, Temme & Verhoeven 2016), i.e., experiencers tend to be aboutness topics, which does not hold for other object arguments such as patients of transitive verbs. Thus, we expect object experiencers, but not patients, to occur in object fronting constructions that can host aboutness topics. Hence, the empirical expectation following from these considerations is that an experiencer-object is more likely than a patient object to occur early in the linearization. We assume that aboutness triggers scrambling (accompanied by formal movement in main declarative clauses). Hence, this syntactic operation is expected to occur with fronted experiencer-objects (unless other contextual triggers apply).

For other predicates, e.g., non-experiential action verbs, a change of the canonical argument order needs to be contextually licensed. Weskott et al. (2011) showed in an experimental study on the licensing of OVS orders that object fronting to the prefield position in main declarative clauses is strongly licensed by contexts involving poset relations as, e.g., a part-whole relation between material in the target sentence and a preceding context sentence.[3] In the present study, we use a poset rela-

[2] In a recent rating study testing unergative dative experiencer verbs (as identified by perfect auxiliary selection, i.e., *haben* ‘have’ for unergatives) vs. unaccusative dative experiencer verbs (auxiliary *sein* ‘be’), Fanselow et al. (2016) found an even finer-grained difference: unaccusatives have an advantage for OS while the unergatives did not show an ordering preference.

[3] Weskott et al. (2011: 7) used examples of the following type: *Peter has washed the car. The side mirror, he left out*, where the referent of the fronted object of the second sentence (*the side mirror*) is in a part-whole relation to the object referent of the first sentence (*the*

tion as a contextual trigger for the licensing of OS orders. The relevant issue for our considerations is that OS order can be achieved through a particular type of contrastive topicalization.

3 Experimental Design, Materials and Predictions

The main aim of the experimental study is to examine the conditions that license OS orders with German accusative experiencer-object verbs. Taking into account the syntactic considerations introduced in Section 2 and the research questions outlined in Section 1, we conducted two experimental studies, a relative judgment task and a forced-choice task each with two alternative expressions. The studies were implemented in parallel and used the same material.[4]

We examine the influence of the factors CONTEXT and VERB TYPE on word order as indicated in (2).

(2) a. Fixed factors
CONTEXT (2 levels): OS licensing vs. neutral (= non-licensing)
VERB TYPE (2 levels): experiencer verb vs. non-experiencer verb

b. Dependent variables
Experiment 1
Acceptability of OS relative to SO: 0-100
Experiment 2
Choice of order: OS vs. SO

This design yields four experimental conditions per experiment:

- experiencer verb & OS licensing context
- experiencer verb & neutral context
- non-experiencer verb & OS licensing context
- non-experiencer verb & neutral context.

The target sentences were constructed in two versions, namely SO and OS, see (3). The factor CONTEXT captures the contextual licensing of these linearizations. We compare the effect of a context licensing object topicalization with an all-new context establishing the baseline. The neutral (non-licensing) context was induced with the question *Was gibt es Neues?* ‘What’s new?’ preceding the target sentence. The object-topic licensing context involves a set-member relationship between the subject of the context sentence and the object of the target sentence and additionally a contrastive reading between the predicates of the two structures (Weskott et al. 2011).[5] The factor VERB TYPE has to disentangle the fronting effect of experiencer-

car).

[4] The results of the forced-choiced study are reported as part of a wider typological study in Temme & Verhoeven (2016).

[5] Similar to the part-whole relations used in Weskott et al. (2011), set-member relations are types of poset relations which have been shown to license topicalization (Ward & Prince 1991).

object verbs from a baseline established by comparable constructions. We established the baseline with canonical transitive verbs governing a patient object.

(3) a. Context C *Die meisten Bürger hatten keine Probleme mit dem Bahnübergang.*
Target A SO: *Die Schranke hat den Pfarrer aufgeregt*$_{\text{EXP}}$/*aufgehalten*$_{\text{CAN}}$.

b. Context C *Die meisten Bürger hatten keine Probleme mit dem Bahnübergang.*
Target B OS: *Den Pfarrer hat die Schranke aufgeregt*$_{\text{EXP}}$/*aufgehalten*$_{\text{CAN}}$.

'Most of the citizens did not have any problems with the railway crossing.'
'The pastor was upset/delayed by the barrier.'

Factors such as definiteness, animacy and agentivity are known to influence argument linearization, hence they were systematically controlled in the material: all target sentences contained an inanimate nominative DP and an animate non-nominative DP, both DPs were definite. In order to meet the contextual conditions for definite descriptions, subject referents are inferable from the context too; see *Schranke* 'barrier' and *Bahnübergang* 'railway crossing' in (3).

We selected sixteen experiencer-object verbs and sixteen canonical transitive verbs by relying on the available literature about the respective verb types (see Table 1 and Temme & Verhoeven 2016 for more information concerning the selection process).

Table 1: Verbs used in the experiment

No.	*experiencer verbs*	*canonical verbs*	No.	*experiencer verbs*	*canonical verbs*
1	plagen 'annoy'	behindern 'hinder'	9	anwidern 'disgust'	vergiften 'poison'
2	erstaunen 'astonish'	schützen 'protect'	10	entzücken 'rapture'	verbessern 'correct'
3	entmutigen 'discourage'	verändern 'change'	11	frustrieren 'worry'	verletzen 'injure'
4	begeistern 'enthuse'	heilen 'heal'	12	wundern 'wonder'	warnen 'warn'
5	verängstigen 'frighten'	wecken 'wake up'	13	beunruhigen 'worry'	blenden 'bedazzle'
6	interessieren 'interest'	abholen 'pick up'	14	erschrecken 'scare'	infizieren 'infect'
7	erfreuen 'delight'	retten 'rescue'	15	aufregen 'upset'	aufhalten 'delay'
8	langweilen 'bore'	zerstören 'destroy'	16	enttäuschen 'disappoint'	blamieren 'disgrace'

Following the discussion in Section 2, we expect for both experiments that the factors CONTEXT and VERB TYPE significantly affect the preference for OS in Experiment 1 and the choice of OS in Experiment 2. In particular, we expect a main effect for CONTEXT, so that contexts licensing object topicalization increase the preference for OS order in the split-100 task (Section 4) and lead to a higher number of OS choices in the forced-choice task (Section 5). For VERB TYPE, correspondingly, we expect that OS order will be preferred (Experiment 1) and more frequently chosen (Experiment 2) with experiencer verbs. Finally, we expect an interaction of the two factors such that the effect of contextual licensing of OS is stronger for canonical than for experiencer verbs.

Regarding the two methods, we expect the typical distribution of the data types described in Section 1: the ranking of the conditions should align for the two alternative measures; the effects in the rating data are expected to be strengthened in the forced-choice data.

4 Experiment 1: Relative Acceptability

4.1 Method

For the examination of scalar intuitions, we collected relative (instead of absolute) judgments in order to observe speakers' intuitions when comparing structures. Such a comparison presumably precedes output-selection in natural speech production.

Based on a latin-square design, we created 16 pseudo-randomized lists, each containing 16 items (8 items of each Verb Type). Each item was accompanied by a context sentence representing one of the levels of Context, so that each list contained four repetitions of each experimental condition. The targets were mixed with 32 filler items also involving a choice between an SO and an OS order. Each item was presented as two context-target pairs (context C with target alternative A and context C with target alternative B), see (3). For any particular context, test subjects were instructed to award points to both alternatives summing up to 100 (i.e., 50/50, 0/100, 81/19, 45/55, etc.) according to the felicity of the target sentences within the presented context. The experiment was run as a web-based study.[6] Each experimental session took approximately 15 minutes and was unpaid. 32 monolingual German native speakers took part in Experiment 1 (27 female, age range 20-36, age average 25.7).

4.2 Results

The obtained results, i.e., the means of the relative acceptability of the OS structure, are presented in Table 2 and plotted in Fig. 2. The descriptive data indicate a main effect of Context (licensed: 56.8; non-licensed: 38.5 across verb types) and a smaller effect of Verb Type, which is driven by the non-licensing context condition (experiencer-object verbs: 50.7; canonical verbs: 44.2 across contexts). Both factors seem to interact in the predicted direction: the OS order has an advantage with experiencer-object verbs even in contexts that do not license object topicalization. However, values below 50 for OS mean that the SO order is still preferred in the non-licensing context.

[6] OnExp Version 1.2; http://onexp.textstrukturen.uni-goettingen.de (last access July 17, 2017)

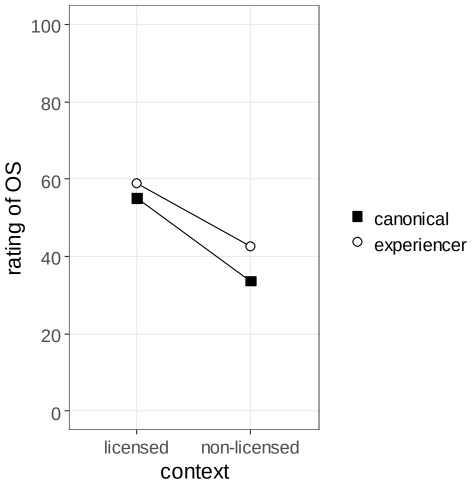

Fig. 2: Split-100 rating: results

Table 2: Split-100 test: results (mean ratings for OS-sentences)

		Context			
		licensing		non-licensing	
		average	SE	average	SE
Verb Type	canonical	54.9	4.84	33.6	3.93
	experiencer	58.8	4.42	42.6	3.68

A linear mixed-effect model was fitted to the data. We applied a backwards-selection procedure selecting the model with the maximal fit based on Log-Likelihood Tests. We started from a model containing a maximal fixed-effects structure and a maximal random-effects structure (intercepts and slopes by Subjects and by Items) (see procedure recommended by Zuur et al. 2009: 121). First, we applied a backwards-selection procedure to the random component of the model. Model selection revealed that the optimal structure is a model only containing random intercepts of Subjects. Second, we applied the backwards selection to the fixed factors with the optimal random-effect structure: the maximal fit was reached by a model that only involves the main effects of the fixed factors. The interaction effect was not significant (the difference between the deviance of the maximal fixed-effects model and the fixed-effects model without interaction is: χ^2 (1) = 1.9). The estimates of the final model are given in Table 3. Verb Type (Log-Likelihood Test: χ^2 (1) = 11.8, $p < .001$) and Context (Log-Likelihood Test: χ^2 (1) = 93.92, $p < .001$) are significant predictors for the results (χ^2 values correspond to the difference between a model with two main effects and a model in which the respective effect is eliminated). The findings confirm the effect of Verb Type and Context on the relative acceptability of OS linearizations in German.

Table 3: Split-100 test: fixed effects

Linear mixed-effect model; Number of obs: 512; groups: item, 32; subject, 32;
Model (of maximal fit): OS.rating ~ verb.type + context + (1|subject)

factors	***β***	**SE**	***t***	***p***[7]
Intercept	34.9	2.7	12.7	< .001
Verb Type (experiencer verb)	6.4	1.9	3.4	< .001
Context (licensing)	18.8	1.9	10.1	< .001

Both main effects are expected and are in line with the assumptions about German word order, as outlined in Section 2. The observed differences are informative for the influence of Verb Type and Context. The effect of Verb Type is evidence for an experiencer-first effect holding for German accusative experiencer verbs. Moreover, the effect of Context is evidence for a topic-first effect, which applies to the same construction. The two effects are cumulated without significant interaction. The results are in line with the view that both contrastive topicalization and scrambling of accusative experiencers as aboutness topics account for OS orders in main clauses with the object realized in the prefield. Furthermore, the experimental results confirm the results in Weskott et al. (2011) about the role of part-whole relations in licensing object fronting. Our results support their claim that these contexts do not display a bias of the contextually non-restricted order (basic SO) but an advantage for the OS linearization.

Finally, note that the baseline in our data may contain an effect of animacy on word order, since the tested items involved inanimate subjects and animate objects; see (3). The disharmonic alignment of the animacy scale with the argument hierarchy may explain the rather high acceptability values for OS (average 33.6) already in the baseline of non-experiential verbs in the non-licensing condition. Regarding the experiencer verbs, the acceptability ratio for OS is similar to those reported in previous studies, which indicate no clear preference for OS vs. SO (e.g., Haupt et al. 2008) and those identifying the role of disharmonic animacy as accounting for a considerable portion of OS with these verbs (e.g., Verhoeven 2015).

4.3 Predictions for Speech Production

The aim of this section is to explore the predictions from the competition model for the behavior of speakers in speech production, i.e., in a situation in which they are forced to choose between the two alternative linearizations. Putting variability aside, speakers are expected to choose the linearization that is judged to be better with a given lexicalization in a given context. In order to assess the predicted outcome, we coded the rating data as indicated in (4). For every item in a certain condition, the choice of OS is predicted if the acceptability of this item in that condition is higher than the corresponding acceptability of SO. Otherwise the choice of OS is expected to be reduced to zero. In the permutations of items and conditions in which both linearizations are judged as equally felicitous, the speaker's choice is unclear. In this case, we expect her to choose at random, i.e., OS is predicted to occur at 50%.

[7] The *p*-values are calculated with the Welch-Satterthwaite equation through the R-package *lmerTest* (Kuznetsova et al. 2016).

(4) For an item *i* in a condition *c*

$$\text{predicted(OS)}_{i,c} \rightarrow \begin{cases} 100, & \text{if } \alpha(\text{OS})_{i,c} > \alpha(\text{SO})_{i,c} \\ 0, & \text{if } \alpha(\text{OS})_{i,c} < \alpha(\text{SO})_{i,c} \\ 50, & \text{if } \alpha(\text{OS})_{i,c} = \alpha(\text{SO})_{i,c} \end{cases}$$

Alternatively, we may expect a stronger bias of the preferred structure in cases of uncertainty about a preference. In this scenario, if both options are equally felicitous, the native speaker is expected to always select the option that is the optimal candidate in a certain condition. This means that for all items where speakers judged both options as equally felicitous in a given condition, selection of OS is predicted when the mean acceptability of OS was higher than that of SO in that condition. In order to calculate the predictions of this scenario, we coded the data as indicated in (5).

(5) For an item *i* in a condition *c*

$$\text{biased(OS)}_{i,c} \rightarrow \begin{cases} 100, & \text{if } \alpha(\text{OS})_{i,c} > \alpha(\text{SO})_{i,c} \\ 0, & \text{if } \alpha(\text{OS})_{i,c} < \alpha(\text{SO})_{i,c} \\ 100, & \text{if } \alpha(\text{OS})_{i,c} = \alpha(\text{SO})_{i,c} \,\&\, \alpha(\text{OS})_{c} > \alpha(\text{SO})_{c} \\ 0, & \text{if } \alpha(\text{OS})_{i,c} = \alpha(\text{SO})_{i,c} \,\&\, \alpha(\text{OS})_{c} < \alpha(\text{SO})_{c} \end{cases}$$

Fig. 3 presents the predicted results according to the simple competition model in (4) in the left panel and the biased competition model in (5). In both cases, the predicted choices move away from the average level (50%) towards extreme values, which is in line with the comparisons between acceptability judgments and frequency data as reported in Section 1. The contrasts are larger if a bias of the winner per condition is added; compare left and right panel.

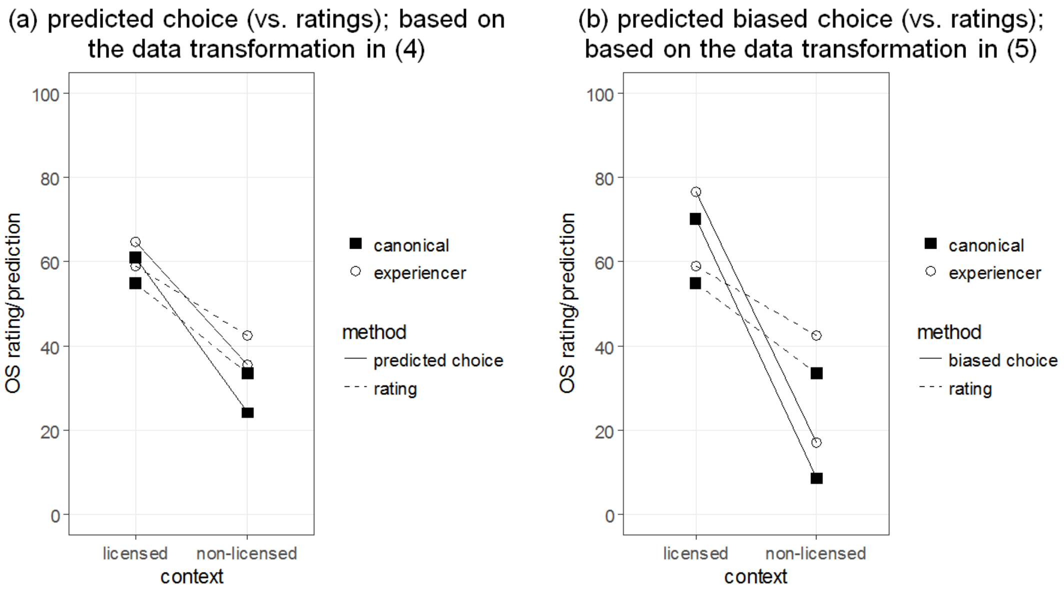

Fig. 3: Expected choice (based on the acceptability judgments)

A simulation of the selection process provides the predictions in Fig. 3 based on the assumption of competition: the speaker (always) selects the winner candidate among the available alternatives. The question is whether these predictions are borne out. In order to answer this question, we performed a forced-choice task with the same material, which is reported in the next section.

5 Experiment 2: Word Order Choice

5.1 Method

As motivated above, we conducted a forced-choice test with the two linearizations SO vs. OS as alternatives, in order to examine the output-selection process; see (3). The factorial design and the material of the forced-choice study were identical to those of Experiment 1. The number of subjects (32 monolingual speakers) was identical as well, however, the speaker sample was different (20 female, age range 23-34, age average 28.3). During the experiment, the subjects were presented with two alternative target sentences, each following a given context as described in Section 3. The speakers were instructed to select the most appropriate sentence in the given context. Experiment 2 was also run web-based (implemented in OnExp). Each experimental session took approximately 15 minutes and was unpaid.

The forced-choice procedure involves a decision between two competing alternatives under the conditions of interest (combinations of VERB TYPE and CONTEXT). The outcome is a choice among the presented alternatives and it simulates the process of selecting a construction in speech production under laboratory conditions. We assume that in naturalistic situations speakers' decisions involve further sources of complexity (e.g., further lexicalization options as well as more complex contextual conditions).

5.2 Results

The frequencies of OS in the forced-choice test are given in Table 4 and plotted in Fig. 4. The descriptive data reveal a main effect of VERB TYPE (experiencer-object verbs: 55.5; canonical verbs: 38.3 across contexts) and a larger effect of CONTEXT (licensed: 63.3; non-licensed: 30.4 across verb types). The advantage of OS frequencies for experiencer-object verbs is higher in the non-licensing contexts than in the licensing contexts, which is in line with the expectation that OS does not require contextual licensing with this verb class.

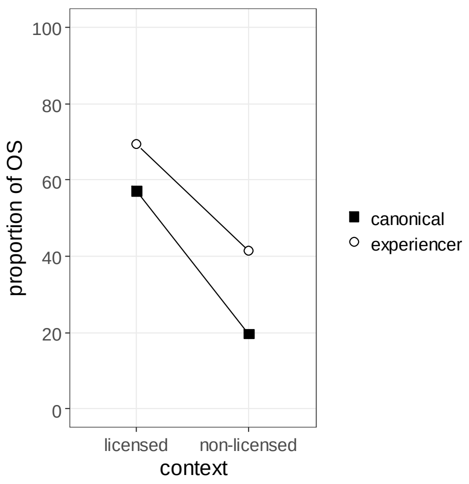

Fig. 4: Forced-choice test: results

Table 4: Forced-choice test: results (absolute frequencies and percentages of OS)

		CONTEXT					
		licensing			non-licensing		
		SO	OS	% OS	SO	OS	% OS
VERB TYPE	canonical	55	73	57.03	103	25	19.53
	experiencer	39	89	69.53	75	53	41.41

The data (i.e., the binary variable choice of OS) was fitted with a generalized linear mixed-effect model. A backwards selection procedure based on Log-Likelihood Tests (same procedure as in 4.2) revealed that the maximal fit is reached by a model with a random intercept for subjects and no interaction effect between the fixed factors (the difference between deviances of the models with and without an interaction effect is 2.2, which does not correspond to a significant *p*-value in the chi-square distribution). VERB TYPE (χ^2 (1) = 21.25) and CONTEXT (χ^2 (1) = 71.22) cannot be eliminated from the model. The estimates of the final model are given in Table 5.

Table 5: Forced-choice test: fixed effects

Generalized linear mixed-effect model; Number of obs: 512; groups: item, 32; subject, 32; Model (of maximal fit): OS.choice ~ verb.type + context + (1|subject)

factors	***β***	**SE**	***z***	***p***
INTERCEPT	–1.5	0.29	–5.8	< .001
VERB TYPE (experiencer verb)	0.9	0.22	4.5	< .001
CONTEXT (licensing)	1.8	0.23	7.8	< .001

The results of the forced-choice study show the same significant effects as the results of the rating study; compare Table 3 and Table 5. However, the contrasts between the obtained means demonstrate differences in magnitude, which we shall examine in detail. This is the aim of Section 6.

6 Comparison Between Experiments

The relation between the ratings and the predicted choice data (based on the competition model in Section 4.3) and the ratings and the data obtained through the forced-choice study (Section 5.2) is shown in Fig. 5. The data points are fitted by a third-order polynomial line. The third-order polynomial function that relates the acceptability data with the predictions of the competition model – without (Fig. 5a) or with (Fig. 5b) a bias of the preferred construction – differs from the polynomial function that relates the acceptability data and the forced-choice data in Fig. 5c. Descriptively, the major difference lies in the fact that the effect of the method interacts with the distance of the data points from the central area of the 0-100 scale. Recall the descriptive data of the two experimental studies. The conditions that involve a single OS-trigger have very similar values in both experiments:

- canonical verb & licensing context:

  54.9 (OS rating), 57% (OS in forced choice);
- experiencer verb & non-licensing context:

  42.6 (OS rating), 41.4% (OS in forced choice).

The conditions that are judged with values having a larger distance from the central value are extrapolated in the forced-choice data:

- canonical verb & non-licensing context:

  33.6 (OS rating), 19.53% (OS in forced choice);
- experiencer verb & licensing context:

  58.8 (OS rating), 69.53% (OS in forced choice).

Intuitively, this relation means the following: the more remote a rating is from the 50-50 split, the greater the bias for the optimal structure in forced choice. This is in line with the observation of previous studies where frequency data display an exponential growth of the advantage in acceptability studies. What our data shows is that the exponential growth correlates with the certainty of the speaker in the selection of a candidate (in other words, with the strength of preference). It starts at the level of 50-50 ratings, i.e., at the level at which the compared expressions are similarly appropriate in the context at issue.

This component of the speakers' behavior is not captured in our predictions from the rating data based on a narrow interpretation of the competition model. In the data transformations in (4) and (5), we predicted that the winner in the rating (for each permutation of item and condition) will be selected in the forced-choice task. Our findings indicate that a further component of variation is involved: the frequencies in forced choice depend on the strength of preference (as reflected in the ratings), i.e., the distance from the 50-50 level. This finding is in line with previous observations in forced-choice data (Rosenbach 2003) or with the comparison between acceptability and corpus data (Bresnan 2007), in which variation differs across contexts such that in particular contexts alternative constructions vary to a significant extent, while in other contexts the choice of construction is almost categorical.

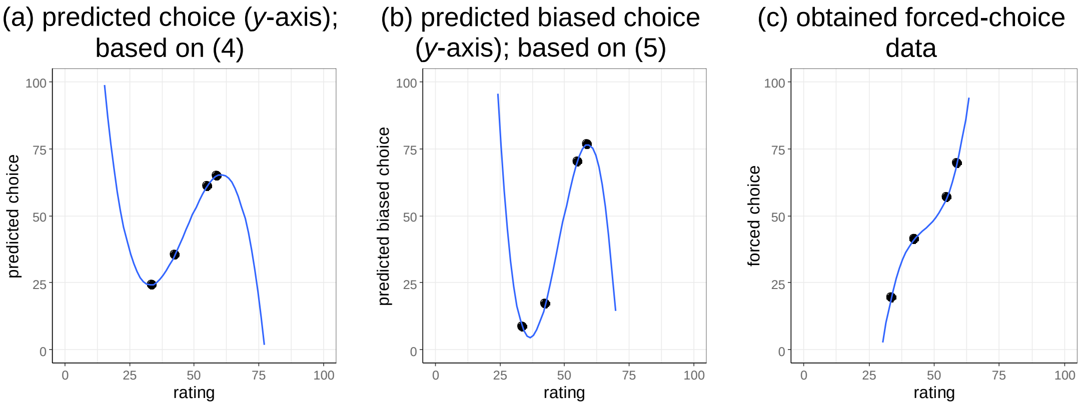

Fig. 5: Comparison of ratings (*x*-axis) with predicted and obtained choice data (*y*-axis)

The descriptive data in Fig. 5 suggests that we can account for a further component of the variation in the forced-choice data, namely the impact of the strength of preference. For this purpose, we added to the generalized linear mixed-effects model in Table 5 the fixed effect of STRENGTH. This fixed effect was calculated *for every trial* in the word-order-choice study by means of the distance of the corresponding ratings from the level at which the alternative orders are equally felicitous: this is the absolute difference of the mean of the ratings that were collected for the *item* and *condition* at issue (i.e., 8 ratings in the relative acceptability study) from the 50-50 level. Hence, STRENGTH is a numeric factor ranging between 0 (when the mean of OS ratings is 50) and 50 (when the mean of OS ratings is either 0 or 100).

The fixed effect of STRENGTH (calculated on the basis of the ratings) has a third-order polynomial relation to the choice of OS (in the forced-choice data). A third-order polynomial relation contains a cubic factor that may capture the fact that the effect of STRENGTH on the choice of OS is not linear: it is correlated with the likelihood of choosing SO below the 50-50 level and with the likelihood of choosing OS above the 50-50 level. The polynomial relation was integrated in the generalized linear-mixed model by means of three components: a linear factor, a quadratic factor, and a cubic factor (cf. Mirman et al. 2008: 486 on growth curve analysis). The backwards selection procedure (same procedure as in 4.2 and 5.2) leads to the model in Table 6. The crucial finding is that the cubic estimate of STRENGTH is significant: a Log-Likelihood Test reveals that the difference between deviances of the models with and without the cubic term yields a χ^2 (1) = 4.3 (p < .05). This finding confirms that the non-linear relation of STRENGTH with the choice of order explains a part of the variation in the data.

Table 6: Forced-choice test and strength of preference: fixed effects

Generalized linear mixed-effect model; Number of obs: 512; groups: item, 32; subject, 32; Model (of maximal fit): OS.choice ~ verb.type + context + poly(strength,3)+(1|subject)

factors		***β***	**SE**	***z***	***p***
INTERCEPT		–1.5	0.31	–5.1	< .001
VERB TYPE (experiencer verb)		0.9	0.22	4.5	< .001
CONTEXT (licensing)		1.7	0.23	7.8	< .001
STRENGTH	LINEAR	–5.0	2.7	–1.8	*
	QUADRATIC	–0.4	2.5	–0.1	*
	CUBIC	5.4	2.5	2.1	< .05

7 Conclusions

The relation between acceptability ratings and frequencies in speech production is known to be exponential, i.e., the likelihood of optimal candidates within a set of alternative expressions grows exponentially. The present study examined the relation between preference ratings and frequencies of choice with a maximally controlled design, using the same material with two experimental procedures (split-100 rating and forced-choice task). This offers the possibility to calculate the exact predictions for the choice of candidates based on the comparison between the ratings for certain contextual conditions and lexicalizations. The results of the forced-choice test show that the behavior of native speakers is not reducible to the choice of the optimal candidate but involves a source of gradience that correlates with the strength of preference. The greater the distance between the acceptability values of the alternative expressions, the stronger the bias for the winner candidate.

References

Adli, A. (2010) On the relation between acceptability and frequency. In E. Rinke & T. Kupisch, eds., *The development of grammar: language acquisition and diachronic change*. Benjamins, Amsterdam: 383-404.

Arppe, A. & J. 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, M. & J. Häussler (2010a) Toward a model of grammaticality judgments. *Journal of Linguistics*, 46: 273-330.

Bader, M. & J. Häussler (2010b) Word order in German: A corpus study. *Lingua*, 120: 717-742.

Den Besten, H. (1989) *Studies in West Germanic syntax*. Atlanta, Amsterdam.

Bickel, B. (2004) The syntax of experiencers in the Himalayas. In P. Bhaskararao & K. V. Subbarao, eds., *Non-nominative subjects*. Benjamins, Amsterdam: 77-112.

Bresnan, J. (2007) Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston & W. Sternefeld, eds., *Roots: Linguistics in search of its evidential base*. Mouton de Gruyter, Berlin: 75-96.

Fanselow, G. (2003) Zur Generierung der Abfolge der Satzglieder im Deutschen. In Japanische Gesellschaft für Germanistik (ed.), *Probleme des Interface zwischen Syntax, Semantik und Pragmatik.* Iudicium, München, 3-47.

Fanselow, G., J. Häussler, & T. Weskott (2016) Constituent order in German multiple questions: Normal order and (apparent) anti-superiority effects. In S. Featherston & Y. Versley, eds., *Quantitative approaches to grammar and grammatical change. Perspectives from Germanic*. Mouton de Gruyter, Berlin: 33-50.

Featherston, S. (2005) The Decathlon Model of empirical syntax. In S. Kepser & M. Reis, eds., *Linguistic evidence: empirical, theoretical, and computational perspectives*. De Gruyter, Berlin: 187-208.

Frey, W. (2006) Contrast and movement to the German prefield. In V. Molnár & S. Winkler, eds., *The architecture of focus*. Mouton de Gruyter, Berlin/New York: 235-264.

Frey, W. (2005) Zur Syntax der linken Peripherie im Deutschen. In F. J. d'Avis, ed., *Deutsche Syntax: Empirie und Theorie*. Acta Universitatis Gothoburgensis, Göteborg: 147-171.

Frey, W. (2004) A medial topic position for German. *Linguistische Berichte*, 198: 153-190.

Haspelmath, M. (2001) Non-canonical marking of core arguments in European languages. In A. Aikhenvald, R. M. W. Dixon, & M. Onishi, eds., *Non-canonical marking of subjects and objects*. Benjamins, Amsterdam/Philadelphia: 53-83.

Haupt, F. S., M. Schlesewsky, D. Roehm, A. D. Friederici, & I. Bornkessel-Schlesewsky (2008) The status of subject-object reanalyses in language comprehension architecture. *Journal of Memory and Language*, 59: 54-96.

Hoberg, U. (1981) *Die Wortstellung in der geschriebenen deutschen Gegenwartssprache*. Hueber, München.

Kempen, G. & 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 & M. Reis, eds., *Linguistic evidence: empirical, theoretical, and computational perspectives*. De Gruyter, Berlin: 329-349.

Kuznetsova, A., P. B. Brockhoff, & R. H. Bojesen Christensen (2016) lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-33. https://CRAN.R-project.org/package=lmerTest (last access July 17, 2017).

Lenerz, J. (1977) Zum Einfluß von "Agens" auf die Wortstellung des Deutschen. In H. W. Viethen, W.-D. Bald, & K. Sprengel, eds., *Grammatik und interdisziplinäre Bereiche der Linguistik. Akten des 11. Linguistischen Kolloquiums Aachen 1976*. Niemeyer, Tübingen: 133-142.

Mirman, D., J. A. Dixon, & J. S. Magnuson (2008) Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59: 475-494.

Müller, G. (2004) Verb-second as vP-first. *Journal of Comparative Germanic Linguistics*, 7(3): 179-234.

Müller, G. (1999) Optimality, markedness, and word order in German. *Linguistics*, 37(5): 777-818.

Primus, B. (2004) Protorollen und Verbtyp: Kasusvariaton bei psychischen Verben. In R. Kailuweit & M. Hummel, eds., *Semantische Rollen*. Narr, Tübingen: 377-401.

Rosenbach, A. (2003) Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In G. Rohdenburg & B. Mondorf, eds., *Determinants of Grammatical Variation in English*. Mouton de Gruyter, Berlin/New York: 379-411.

Temme, A. & E. Verhoeven (2016) Verb class, case, and order: A cross-linguistic experiment on non-nominative experiencers. *Linguistics*, 54(4): 769-814.

Thiersch, C. (1978) *Topics in German syntax*. PhD thesis, MIT.

Verhoeven, E. (2015) Thematic asymmetries do matter! A corpus study of German word order. *Journal of Germanic Linguistics*, 27(1): 45-104.

Verhoeven, E. (2014) Thematic prominence and animacy asymmetries. Evidence from a cross-linguistic production study. *Lingua*, 143: 129-161.

Ward, G. L. & E. F. Prince (1991) On the topicalization of indefinite NPs. *Journal of Pragmatics*, 16(2): 167-177.

Webelhuth, G. (1995) German is configurational. *The Linguistic Review*, 4: 203-246.

Weskott, T. & G. Fanselow (2011) On the informativity of different measures of linguistic acceptability. *Language*, 87(2): 249-273.

Weskott T., R. Hörnig, G. Fanselow, & R. Kliegl (2011) Contextual Licensing of Marked OVS Word Order in German. *Linguistische Berichte*, 225: 3-18.

Zuur, A. F., E. N. Ieno, N. J. Walker, A. A. Saveliev, & G. M. Smith (2009) *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York.