Experimental Evidence for Constraint Competition in Gapping Constructions

Frank Keller

This paper presents the results of two experiments investigating gradient acceptability in gapping constructions. Experiment 1 shows that adjuncts and complements are equally acceptable as remnants in gapping, a fact that has been surrounded by controversy in the literature. It also provides evidence against the claim that gapping must leave behind exactly two remnants, and shows that subject remnants are less acceptable than object remnants. This effect of remnant type can be overridden by context. Experiment 2 confirms the remnant effect and investigates how it interacts with other constraints on gapping to produce a gradient acceptability pattern.

A number of grammar models have been proposed to deal with gradient linguistic data, including the re-ranking model (Keller 1998), which draws on concepts from Optimality Theory. Two assumptions are central to this model: (a) constraint violations are cumulative, i.e., the degree of unacceptability increases with the number of constraints violated; and (b) constraints cluster into two types based on their acceptability profile: hard constraints cause strong unacceptability when violated, while violations of soft constraints cause only mild unacceptability. The experimental data presented in this paper confirm both assumptions and provide additional evidence for the hard/soft distinction by demonstrating that only soft constraints are subject to context effects.

1 Introduction

The aim of this paper is twofold. Firstly, we aim to make a methodological point by showing that experimental techniques can contribute to linguistic theory by settling data disputes that cannot be resolved solely on the basis of intuitive, informal acceptability judgments. More specifically, we apply the experimental paradigm of magnitude estimation to gapping constructions, which allows us to test claims made in the theoretical literature on gapping.

Gereon Müller and Wolfgang Sternefeld, eds., *Competition in Syntax*,
211–248. Berlin: Mouton de Gruyter, 2001.

We provide evidence for constraint competition in gapping and investigate the influence of context on the acceptability of gapped sentences.

The second aim of this paper is to obtain experimental data regarding suboptimal linguistic structures, i.e., structures that attract gradient acceptability judgments. Such gradient data allow us to test aspects of a specific model of gradience in grammar, the re-ranking model. More specifically, the data bear on two central assumptions of this model: the cumulativity of constraint violations and the dichotomy of hard and soft constraints.

In this introduction, we give a brief overview of the theoretical literature on gapping, provide some background on Optimality Theory, and outline the re-ranking model of gradience.

1.1 Gapping Constructions in English

Gapping is a grammatical operation that deletes certain subconstituents of a coordinate structure. As examples consider (1)–(3) below, in which the (a) examples constitute gapped versions of the (b) examples:[1]

(1) a. I ate fish, Bill rice, and Harry roast beef.
    b. I ate fish, Bill ate rice, and Harry ate roast beef.

(2) a. Tom has a pistol, and Dick a sword.
    b. Tom has a pistol, and Dick has a sword.

(3) a. I want to try to begin to write a novel, and Mary

$$\left\{\begin{array}{r}\text{to try to begin to write}\\ \text{to begin to write}\\ \text{to write}\\ \emptyset\end{array}\right\}\text{a play.}$$

    b. I want to try to begin to write a novel, and Mary wants to try to begin to write a play.

These examples indicate that gapping always deletes the matrix verb and leaves behind exactly two constituents as remnants (Kuno 1976: 318). Based on previous work by Hankamer (1973), Jackendoff (1971), and Ross (1970), Kuno (1976) also observes that certain functional principles affect the acceptability of gapping, such as the following restriction on the interpretation of the constituents left behind by gapping:[2]

(4) *The Minimal Distance Principle* [MinDis] (Kuno 1976: 306)
The two constituents left behind by Gapping can be most readily

coupled with the constituents (of the same structures) in the first conjunct that were processed last of all.

The examples in (5) illustrate the Minimal Distance Principle: In (5-a), the remnant *Tom* has to be paired with *Mary*, yielding the interpretation in (5-b). It is not possible to pair *Tom* with the more distant subject *John*, yielding the interpretation in (5-c).

(5) a. John believes Mary to be guilty, and Tom to be innocent.
b. John believes Mary to be guilty, and John believes Tom to be innocent.
c. John believes Mary to be guilty, and Tom believes Mary to be innocent.

A further generalization about gapping constructions is that the gap has to represent contextually given information, while the remnant has to constitute new information. Kuno (1976) captures this using the concept of Functional Sentence Perspective (FSP):

(6) *The FSP Principle of Gapping* [SENTP] (Kuno 1976: 310)
Constituents deleted by Gapping must be contextually known. On the other hand, the two constituents left behind by Gapping necessarily represent new information and, therefore, must be paired with constituents in the first conjunct that represent new information. [...]

Kuno (1976) notes that the FSP Principle seems to be able to override the Minimal Distance Principle. (7-a) is acceptable as a gapped version of (7-b), even though it violates MINDIS. We regard this fact as initial evidence that gapping is subject to constraint competition in an optimality theoretic sense.

(7) a. With what did John and Bill hit Mary? John hit Mary with a stick, and Bill with a belt.
b. With what did John and Bill hit Mary? John hit Mary with a stick, and Bill hit Mary with a belt.

More evidence for constraint competition in gapping comes from Kuno's (1976) observation that the remnants in a gapped sentence tend to be interpreted as a subject and its predicate:

(8) *The Tendency for Subject-Predicate Interpretation*
[SUBJPRED] (Kuno 1976: 311)
When Gapping leaves an NP and a VP behind, the two constituents

are readily interpreted as constituting a sentential pattern, with the NP representing the subject of the VP.

This explains why (9-a) can be interpreted as the gapped version of (9-b) (where *Tom* is the subject of *donate*), but not as the gapped version of (9-c) (where *Tom* is the subject of the object control verb *persuade*). Example (10-a), on the other hand, not only has (10-b) as a possible interpretation, but also (10-c) (or at least (10-c) is considerably better than (9-c)). In (10-c), *Tom* is the subject of *donate*, because the matrix verb *promise* is a subject control verb. Such a subject-predicate interpretation is preferred in gapping constructions. Note that (10-c) violates MINDIS, thus indicating a competition between MINDIS and SUBJPRED.

(9) a. John persuaded Bill to donate $200, and Tom to donate $400.
b. John persuaded Bill to donate $200, and John persuaded Tom to donate $400.
c. John persuaded Bill to donate $200, and Tom persuaded Bill to donate $400.

(10) a. John promised Bill to donate $200, and Tom to donate $400.
b. John promised Bill to donate $200, and John promised Tom to donate $400.
c. John promised Bill to donate $200, and Tom promised Bill to donate $400.

Finally, Kuno (1976) also observes that gapping cannot leave behind remnants that are part of a subordinate clause: (11-a) cannot be understood as a gapped version of (11-b).

(11) a. John persuaded Dr. Thomas to examine Jane and Bill Martha.
b. John persuaded Dr. Thomas to examine Jane and Bill persuaded Dr. Thomas to examine Martha.

This can be formulated as the generalization that the remnants in a gapping construction must be part of a simplex sentence:

(12) *The Requirement for Simplex-Sentential Relationship* [SIMS] (Kuno 1976: 314)
The two constituents left over by Gapping are most readily interpretable as entering into a simplex-sentential relationship. The intelligibility of the gapped sentence declines drastically if there is no such relationship between the two constituents.

According to Kuno (1976: 316), “the Requirement for Simplex-Sentential Relationship is a very strong and nearly inviolable constraint,” and a violation of this constraint leads to strong unacceptability. Kuno (1976) claims that the interaction of this constraint with weaker ones such as MINDIS, SENTP, and SUBJPRED, allows us to derive the degree of acceptability of gapped sentences.

However, Kuno (1976) does not make this interaction explicit; he fails to give an account of how the degree of acceptability of a gapped sentence is computed from the constraint violations it incurs. The present paper aims to overcome this limitation. Using experimental data we investigate how the interaction of constraints on gapping determines the degree of acceptability of a gapped structure. Our investigation is guided by an explicit model of constraint competition that draws on concepts from Optimality Theory, introduced in the next section.

1.2 Optimality Theory

Our model of constraint interaction in gapping constructions relies on the concept of grammatical competition recently introduced into linguistic theory by approaches such as Optimality Theory (OT; Prince and Smolensky 1993, 1997) or the Minimalist Program (MP; Chomsky 1995). In what follows, we focus on Optimality Theory, and briefly introduce its basic mechanisms.

Standard Optimality Theory deviates from more traditional linguistic frameworks in that it assumes grammatical constraints to be (a) universal, (b) violable, and (c) ranked. Assumption (a) means that constraints are maximally general, i.e., they contain no exceptions or disjunctions, and there is no parameterization across languages. Highly general constraints will inevitably conflict; therefore, assumption (b) allows constraints to be violated, even in a grammatical structure, while assumption (c) states that some constraint violations are more serious than others. While, according to (a), the formulation of constraints remains constant across languages, the ranking of the constraints can differ between languages, thus allowing crosslinguistic variation to be accounted for.

In an OT setting, a structure is grammatical if it is the *optimal* structure in a set of candidate structures. Optimality is defined via constraint ranking: The optimal structure violates the least highly ranked constraints compared to its competitors. The number of violations plays a secondary role; if two structures violate a constraint with the same rank, then the number of violations incurred decides the competition. OT therefore deviates from traditional

Table 1. Constraint profile for direct object extraction (simplified from Legendre et al. 1995: (22-a))

	[Q$_j$ [think$_{\text{CP}}$ [x$_j$]]]	SubCat	Bar4	Bar3	Bar2	*t
a.	what$_j$ do [you [think [he [said t$_j$]]]]	*		*		*
b.	what$_j$ do [you [think [t$_j$ that [he [said t$_j$]]]]]				**	**
c.	what$_j$ do [you [think [that [he [said t$_j$]]]]]		*			*

grammatical frameworks in that the grammaticality of a sentence is not determined in isolation, but in comparison with other possible structures. Note that there is no inherent restriction on the number of optimal candidates for a given candidate set; more than one candidate may be optimal if several candidates share the same constraint profile, i.e., if they incur the same constraint violations.

We will illustrate how OT works with a simple example taken from an account of *wh*-extraction by Legendre et al. (1995). Our example deals with extraction from direct objects in English. Legendre et al. (1995) assume that the following constraints govern extraction: SubCat, which states that the subcategorization requirements of the verb have to be met; *t, which disallows traces (i.e., movement); and Bar*n*, which rules out movement that crosses more than *n* barriers (for a definition of barrier, see Legendre et al. 1995). For English, the assumption is that these constraints are ranked as follows:

(13) SubCat ≫ Bar4 ≫ Bar3 ≫ Bar2 ≫ *t

This means that a violation of SubCat is more serious than a violation of Bar4, which in turn is more serious than a violation of Bar3, etc.

A crucial assumption in OT is that all candidate structures (syntactic representations) that take part in a grammatical competition are generated from a common input, assumed to be a predicate argument structure by Legendre et al. (1995). The input structure specifies the verb and the arguments of the verbs, plus operators and scope relations that might be present. As an example, consider the first line of Table 1: This input contains the verb *think* (subcategorizing for a CP complement) and specifies that its argument has to contain a syntactic variable x$_j$ which is in the scope of a question operator Q$_j$. Such an input has to be realized by a *wh*-question.

Possible realizations of this input are the candidates (a)–(c) in Table 1.

These candidates violate different constraints, as indicated by the asterisks in Table 1. For example, candidate (a) violates SUBCAT (as the verb takes an IP complement, instead of a CP complement), *t (due to the moved *wh*-element it contains), and BAR3 (because the movement crosses three barriers).

The *optimal* structure in a candidate set is computed as the structure that violates the least highly ranked constraints. As an example, consider the competition between candidates (a) and (c): (a) violates SUBCAT, while (c) violates BAR4. According to the constraint hierarchy in (13), SUBCAT is ranked higher than BAR4, which means that candidate (c) wins the competition. Note that all the other constraints that are violated by either of the candidates are not taken into account in determining the winner. Only the most highly ranked constraint on which the two candidates differ matters for the constraint competition (*strict domination* of constraints). Two candidates differ on a constraint if one candidate violates that constraint more often than the other one (e.g., (a) violates SUBCAT once, while (b) violates it zero times).

In Table 1 the optimal candidate is (b): It wins against (c), as it violates BAR2 instead of BAR4. The additional trace that (b) contains allows it to avoid crossing four barriers at once. This means that (b) incurs two violations of *t (instead of just one). However, this is not relevant to the competition with (c), due to strict domination. (Note that (a) would win if the input contained *think* subcategorizing for an IP.)

Another important aspect of OT can also be illustrated using the extraction example: In OT, crosslinguistic variation can be accounted for by *constraint re-ranking*. Assume that there is an additional constraint *Q, which disallows empty question operators. For English, the ranking *Q $\gg$ *t holds. This means that questions are formed by movement of *wh*-elements, while in-situ *wh*-elements, which have to be bound by the Q operator, are ungrammatical. Chinese, on the other hand, exhibits the opposite ranking *t $\gg$ *Q, i.e., the use of an empty question operator is preferred to the use of a trace. This explains why in Chinese, *wh*-elements remain in situ in direct object extractions, where the *wh*-element is bound by the Q operator. English, on the other hand, requires *wh*-movement in such configurations, as illustrated by the example in Table 1.

1.3 Suboptimal Candidates

Standard OT assumes that all non-optimal candidates are equally ungrammatical, which leads to a binary notion of grammaticality. We propose dropping this assumption and argue for an extended version of OT that not only computes the optimal candidate for a given candidate set, but also makes predictions about the relative grammaticality of *suboptimal candidates*. More specifically, we adopt the following hypothesis (see Keller and Alexopoulou 2000 for details):

(14) **Suboptimality Hypothesis**

a. Suboptimal candidates differ in grammaticality.

b. The relative grammaticality of suboptimal candidates can be used as evidence for constraint rankings.

Note that (14-b) follows from (14-a): If suboptimal candidates differ in grammaticality, then the comparison between two suboptimal candidates can be used as evidence for constraint rankings in the same way as the comparison between a grammatical candidate and an ungrammatical candidate is used to determine rankings in standard OT.

There are several ways of implementing the suboptimality hypothesis, i.e., of extending OT to make predictions about suboptimal structures; the most straightforward one is based on the assumption that the relative grammaticality of a candidate corresponds to its relative optimality in the candidate set (Keller 1997). Such a model will make predictions of the form: Candidate S_1 is more optimal (i.e., more grammatical) than candidate S_2, where both S_1 and S_2 may be suboptimal candidates. This prediction can be tested empirically by showing that S_1 is more acceptable than S_2.

This "naive" model of suboptimality (which simply equates relative optimality with relative grammaticality) has been criticized for a number of reasons (Keller 1998, Müller 1999). One problem is that it predicts grammaticality differences *only* for structures in the same candidate set; relative grammaticality cannot be compared across candidate sets. Another problem is that grammaticality differences are predicted between *all* structures in a candidate set. A typical OT grammar assumes a richly structured constraint hierarchy, therefore all or most structures in a given candidate set will differ in optimality. The naive model predicts that there is a grammaticality difference whenever there is a difference in optimality. This means it will probably overgenerate, i.e., predict far more degrees of grammaticality than we can reasonably expect to find in the data.

1.4 The Re-Ranking Model

A number of suboptimality-based models of gradience have been proposed that avoid the problems with the naive model (Hayes 2000, Hayes and MacEachern 1998, Keller 1998, Müller 1999). The present paper takes as its starting point the re-ranking model put forward by Keller (1998), which is based on experimental research on gradient acceptability in extraction from picture NPs (Cowart 1989, 1997, Keller 1996, 1997). We summarize the relevant experimental findings:

- **Soft and Hard Constraints:** constraints cluster into two types based on their acceptability profile: Hard constraints cause strong unacceptability when violated (e.g., constraints on phrase structure, agreement, and subcategorization), while violations of soft constraints cause only mild unacceptability (e.g., constraints on referentiality and definiteness). Violations of hard constraints are significantly less acceptable than violations of soft constraints.[3]
- **Cumulativity:** constraint violations are cumulative, i.e., the degree of unacceptability increases with the number of constraints violated. This finding holds both for soft and for hard constraints.

Apart from lending a certain plausibility to OT's notions of constraint ranking and constraint interaction (see Keller 1998 for details), these results also provide evidence against a naive model of gradience. The naive model fails to accommodate the distinction between hard and soft constraints and cannot explain the cumulativity effect.

Keller (1998) suggests an alternative model of gradience that draws on concepts from OT learnability theory (Tesar and Smolensky 1998). The central idea of this model is to compute which constraint re-rankings are required to make a suboptimal structure optimal. This information can then be used to compare structures with respect to their degree of grammaticality: The assumption is that the degree of grammaticality of a candidate structure *S* depends on the number and type of re-rankings required to make *S* optimal. Such a re-ranking model offers the necessary flexibility to accommodate the experimental findings on constraint ranking and constraint interaction in OT:

- The re-ranking model allows us to determine the relative grammaticality of arbitrary structures by comparing the number and type of re-rankings required to make them optimal. Comparisons of grammaticality are not confined to structures in the same candidate set, which accounts for the fact that subjects can judge the relative grammaticality of arbitrary sentence pairs.

- It seems plausible to assume that some constraint re-rankings are more serious than others, and hence cause a higher degree of ungrammaticality in the target structure. This assumption allows us to model the experimental findings that some constraint violations are more serious than others. The experimental data justify two types of re-rankings, corresponding to the soft and hard constraint violations discussed above.
- Another assumption is that the degree of grammaticality of a structure depends on the number of re-rankings necessary to make it optimal: The more re-rankings a structure requires, the more ungrammatical it becomes. This predicts the cumulativity of violations that was found experimentally both for soft and for hard constraints.[4]

The work presented in this paper aims to provide additional evidence for two assumptions underlying the re-ranking model: (a) the dichotomy of hard and soft constraints and (b) the cumulativity of constraint violations. An additional aim is to investigate how context effects interact with the soft/hard distinction and the cumulativity effect.

1.5 Magnitude Estimation

The present study relies on very subtle linguistic intuitions, viz., on judgments about the relative acceptability of information structurally different realizations of a sentence. Such intuitions about relative acceptability should be measured experimentally, since the informal elicitation technique traditionally used in linguistics is unlikely to be reliable here (Cowart 1997, Schütze 1996, Sorace 1992). A suitable experimental paradigm is magnitude estimation, a technique standardly applied in psychophysics to measure judgments of sensory stimuli (Stevens 1975). The magnitude estimation procedure requires subjects to estimate the magnitude of physical stimuli by assigning numerical values proportional to the stimulus magnitude they perceive. Highly reliable judgments can be achieved for a whole range of sensory modalities, such as brightness, loudness, or tactile stimulation.

The magnitude estimation paradigm has been extended successfully to the psychosocial domain (Lodge 1981), and recently Bard et al. (1996) and Cowart (1997) have shown that linguistic judgments can be elicited in the same way as judgments of sensory or social stimuli. In contrast to the five

Table 2. Factors in Experiment 1

verb frame (*Frame*)		remnant (*Remn*)	context (*Con*)
trans.	NP V NP	—	felicitous context
	NP V PP		null context (control)
	NP V VP		
	NP V PP-adj		
ditrans.	NP V NP NP	NP _ XP XP	felicitous context
	NP V NP PP	_ _ XP XP	null context (control)
	NP V NP VP	NP _ _ XP	
		NP _ XP _	

or seven point scale conventionally used to measure human intuitions, magnitude estimation employs a continuous numerical scale. It provides fine-grained measurements of linguistic acceptability, which are robust enough to yield statistically significant results, while being highly replicable both within and across speakers. Since magnitude estimation provides data on an interval scale, parametric statistics can be used for evaluation.

Magnitude estimation requires subjects to assign numbers to a series of linguistic stimuli proportional to the acceptability they perceive. First, subjects are exposed to a modulus item, to which they assign an arbitrary number. Then, all other stimuli are rated proportional to the modulus, i.e., if a sentence is three times as acceptable as the modulus, it gets three times the modulus number, etc.

2 Experiment 1: Verb Frame, Remnant, and Context

2.1 Introduction

Experiment 1 was designed to investigate whether the following constraints on gapping that have been proposed in the literature have a gradient effect on the acceptability of gapped sentences: (a) the verb frame of the gapped verb, (b) whether the remnant left behind by gapping is a complement or an adjunct, (c) the structure of the remnant, and (d) the context preceding the gapped sentence. Table 2 gives an overview of the factors included in this experiment and their levels.

The factor verb frame (*Frame*) included both transitive and ditransitive verbs. The transitive case included verbs with NP, PP, and VP complements.

PP adjuncts were also included in order to test the claim that adjunct remnants are more acceptable than complement remnants (Hankamer 1973). The following examples illustrate the levels of the factor *Frame* for transitive verbs:

(15) a. **NP V NP:** She repeated the question, and he the answer.
b. **NP V PP:** She negotiated with the manager, and he with the secretary.
c. **NP V VP:** She expected to win, and he to lose.
d. **NP V PP-adj:** She read in the bedroom, and he in the lounge.

For ditransitive verbs, the factor *Frame* included verbs that have an NP as their first complement, and another NP, a PP, or a VP as their second complement, such as the examples in (16).

(16) a. **NP V NP NP:** She charged the client 50 pounds, and he the manufacturer 100 pounds.
b. **NP V NP PP:** She accompanied the boy to school, and he the girl to university.
c. **NP V NP VP:** She authorized the manager to leave, and he the secretary to stay.

Transitive verbs allow only one type of remnant (where the subject and the object are left behind, while the verb is gapped). Ditransitive verbs, on the other hand, allow more complicated remnants, which we took into account by including the additional factor remnant type (*Remn*) for ditransitive verbs. The levels of *Remn* can be exemplified by the following sentences:

(17) a. **NP _ XP XP:** She charged the client 50 pounds, and he the manufacturer 100 pounds.
b. **_ _ XP XP:** She charged the client 50 pounds, and the manufacturer 100 pounds.
c. **NP _ _ XP:** She charged the client 50 pounds, and he 100 pounds.
d. **NP _ XP _:** She charged the client 50 pounds, and he the manufacturer.

Note that we use pronouns in (17-c) and (17-d) to make sure that the remnant is interpreted as the subject NP.

Context (*Con*), the third factor in the experiment, was meant to test the influence of context on the acceptability of gapping. A felicitous context for gapping (according to Kuno's 1976 SentP constraint) is one in which the gapped constituent contains given information, while the remnants constitute new information. Such a given-new partition can be realized using a question context: new constituents in the answer are realized as *wh*-phrases in the question, while given constituents in the answer are realized as full NPs in the question. This is illustrated by the questions in (18), which constitute felicitous contexts for the transitive sentences in (15):

(18) a. What did Hanna and Michael repeat?
b. Who did Emily and Matthew negotiate with?
c. What did Rachel and Andrew expect to do?
d. Where did Rebecca and Mark read?

The factor *Con* was the same for the ditransitive condition. Here are the felicitous contexts for the examples in (17):

(19) a. Who did Hanna and Michael charge what?
b. Who did Hanna charge what?
c. What did Hanna and Michael charge the client?
d. Who did Hanna and Michael charge 50 pounds?

A null context condition was included as a control condition, allowing us to determine how subjects behave in the absence of contextual information.

2.2 Predictions

The predictions for the present experiment can be summarized as follows:

1. As far as the factor *Frame* is concerned, no clear predictions can be derived from the literature as to the effect of complement type (NP, PP, or VP) or arity (transitive or ditransitive) of the verb. As for the complement/adjunct status of the remnant, our experiment allows us to verify Hankamer's (1973) claims that PP adjuncts are more acceptable than PP complements.[5]

2. For the factor *Remn*, the constraint MinDis predicts that the remnant _ _ XP XP is more acceptable than the remnants NP _ _ XP and NP _ XP _. Another relevant prediction is that the remnant NP _ XP XP is unacceptable, based on the claim of Kuno (1976: 318) that gapping has to leave behind exactly two constituents.

3. As for the effect of *Con*, Kuno's (1976) constraint SentP predicts that the acceptability of a gapped sentence should be increased in a felicitous context, compared to the control condition (the null context).

Furthermore, we predict an interaction between the factors *Remn* and *Con*, based on Kuno's (1976) observation that the satisfaction of SENTP seems to override a violation of MINDIS (see Section 1.1).

2.3 Method

2.3.1 *Subjects*

Fifty-five native speakers of English participated in the experiment. The subjects were recruited over the Internet by postings to newsgroups and mailing lists. Participation was voluntary and unpaid. Subjects had to be naive, i.e., neither linguists nor students of linguistics were allowed to participate.

The data of two subjects were excluded as they turned out to be non-native speakers. The data of a further two subjects were excluded because they were linguists. Finally, the data of two subjects were eliminated after an inspection of their response times showed that they had not completed the experiment adequately (uniform response pattern or response times $<$ 1s). This left 49 subjects for analysis. Of these, 29 subjects were male, 20 female; eight subjects were left-handed, 41 right-handed. The age of the subjects ranged from 14 to 52 years; the mean was 30.6 years.

2.3.2 *Materials*

Training Materials

The experiment included a set of training materials that were designed to familiarize subjects with the magnitude estimation task. The training set contained six horizontal lines. The range of largest to smallest item was 1:6.7. The items were distributed evenly over this range, with the largest item covering the maximal window width of the web browser. A modulus item in the middle of the range was provided.

Practice Materials

A set of practice items was used to familiarize subjects with applying magnitude estimation to linguistic stimuli. The practice set consisted of six sentences that were representative of the test materials. A wide spectrum of acceptability was covered, ranging from fully acceptable to severely unacceptable. A modulus item in the middle of the range was provided.

Test Materials

The experiment included two subdesigns, as illustrated in Table 2. For the transitive items, a full factorial design was used with verb frame (*Frame*) and context (*Con*) as the two factors, yielding a total of $\mathit{Frame} \times \mathit{Con} = 4 \times 2 = 8$ cells. For the ditransitive items, the additional factor remnant type (*Remn*) was included, yielding $\mathit{Frame} \times \mathit{Remn} \times \mathit{Con} = 3 \times 4 \times 2 = 24$ cells. Four lexicalizations were used for each of the cells, which resulted in a total of 128 stimuli. A set of 32 fillers was used, designed to cover the whole acceptability range.

To control for possible effects from lexical frequency, the stimuli in both subdesigns were matched for frequency. Verb and noun frequencies were obtained from a lemmatized version of the British National corpus (100 million words) and average frequencies were computed for the verb, the head noun of the subject, and the head noun of the object for each frame. An ANOVA confirmed that the average verb, subject, and object frequencies did not differ significantly between frames.

2.3.3 *Procedure*

The method used was magnitude estimation as proposed by Lodge (1981) and extended to linguistic stimuli by Bard et al. (1996). Each subject took part in an experimental session that lasted approximately 15 minutes and consisted of a training phase, a practice phase, and an experimental phase. The experiment was self-paced, though response times were recorded to allow the data to be screened for anomalies.

The experiment was conducted remotely over the Internet. The subject accessed the experiment using his or her web browser. The browser established an Internet connection to the experimental server, which was running WebExp 2.1 (Keller et al. 1998), an interactive software package for administering web-based psychological experiments.

Instructions

Before the actual experiment started, a set of instructions were presented. The instructions first explained the concept of numerical magnitude estimation of line length. Subjects were instructed to make estimates of line length relative to the first line they would see, the reference line. Subjects were told to give the reference line an arbitrary number, and then assign a number to each following line so that it represented how long the line was in proportion to the reference line. Several example lines and corresponding numerical estimates were provided to illustrate the concept of proportionality.

Then subjects were told that linguistic acceptability could be judged in the same way as line length. The concept of linguistic acceptability was not defined; instead, examples of acceptable and unacceptable sentences were provided, together with examples of numerical estimates.

Subjects were told that they could use any range of positive numbers for their judgments, including decimals. It was stressed that there was no upper or lower limit to the numbers that could be used (exceptions being zero or negative numbers). Subjects were urged to use a wide range of numbers and to distinguish as many degrees of acceptability as possible. It was also emphasized that there were no "correct" answers, and that subjects should base their judgments on first impressions, and not to spend too much time thinking about any one sentence.

Demographic Questionnaire

After the instructions, a short demographic questionnaire was administered. The questionnaire included name, email address, age, sex, handedness, academic subject or occupation, and language region. Handedness was defined as "the hand you prefer to use for writing", while language region was defined as "the place (city, region/state/province, country) where you learned your first language". The results of the questionnaire were reported above.

Training Phase

The training phase was meant to familiarize subjects with the concept of numeric magnitude estimation using line lengths. Items were presented as horizontal lines centered in the window of the subject's web browser. After viewing an item, the subject had to provide a numerical judgment over the computer keyboard. After pressing Return, the current item disappeared and the next item was displayed. There was no possibility of revisiting previous items or change responses once Return had been pressed. No time limit was set for either the item presentation or for the response.

Subjects first judged the modulus item, and then all the items in the training set. The modulus remained on the screen all the time to facilitate comparison. Items were presented in random order, with a new randomization being generated for each subject.

Practice Phase

This phase allowed subjects to practice magnitude estimation of linguistic acceptability. The presentation and response procedures were the same in the training phase, with linguistic stimuli being displayed instead of lines. Each subject judged the whole set of practice items.

Experimental Phase

The presentation and response procedures in the experimental phase were the same as in the practice phase. A between subjects design was used to administer the factor *Con*: Subjects in Group A judged non-contextualized stimuli, while subjects in Group B judged contextualized stimuli. The factors *Frame* and *Remn* were administered within subjects. There were 64 stimuli per group, which were placed in a Latin square design, generating four lexicalizations at 16 items for each of the groups.

Each subject saw one of the lexicalizations and 16 fillers, i.e., a total of 32 items. Each subject was randomly assigned to a group and a lexicalization: 26 subjects were assigned to Group A, and 23 to Group B. Instructions, examples, training items, and fillers were adapted for Group B to take context into account.

2.4 Results

The data were normalized by dividing each numerical judgment by the modulus value that the subject had assigned to the reference sentence. This operation creates a common scale for all subjects. All analyses were carried out on the geometric means of the normalized judgments. The use of geometric means is standard practice for magnitude estimation data (Bard et al. 1996, Lodge 1981).

Separate analyses of variance (ANOVAs) were performed for the transitive and ditransitive verb frames. The analysis of the transitive frames failed to find a significant main effect of verb frame. The main effect of context was significant only by items ($F_1(1,47) = .326$, $p = .571$; $F_2(1,6) = 29.720$, $p = .002$), and the interaction of frame and context was non-significant. The average judgments for the transitive condition are graphed in Figure 1.

For the ditransitive frames, a marginal main effect of verb frame was found ($F_1(2,94) = 2.727$, $p = .071$; $F_2(2,12) = 6.037$, $p = .015$). Furthermore, the ANOVA showed a highly significant main effect of remnant type ($F_1(3,141) =$

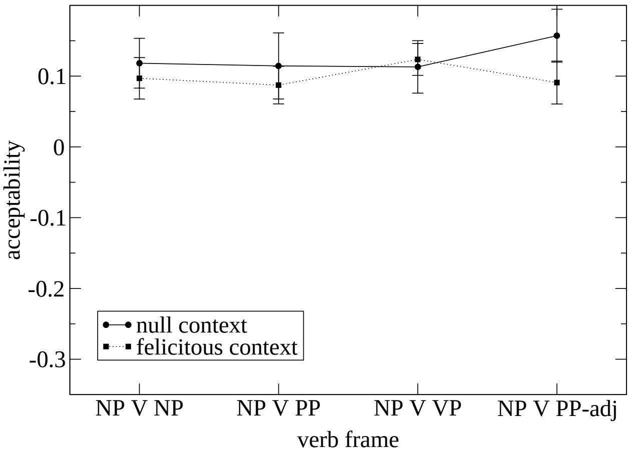

Figure 1. Effect of verb frame on gapping (transitive frames)

18.936, $p < .0005$; $F_2(3,18) = 6.564$, $p = .003$), and an interaction of verb frame and context ($F_1(2,94) = 5.661$, $p = .005$; $F_2(2,12) = 5.096$, $p = .025$). The interaction of remnant type and context was significant only by subjects ($F_1(3,141) = 5.483$, $p = .001$; $F_2(3,18) = 1.847$, $p = .175$). No main effect of context was found, and all the remaining interactions were non-significant.

To further investigate the interactions context/verb frame and context/remnant type, separate ANOVAs were performed for the context condition and the null context condition. In the null context condition, remnant type was significant ($F_1(3,75) = 15.066$, $p < .0005$; $F_2(3,9) = 5.766$, $p = .018$), while verb frame, as well as all interactions, failed to reach significance. The mean judgments for the null context conditions are graphed in Figure 2. This graph shows that the _ _ XP XP remnant is more acceptable than the other remnants. This effect is consistent across all frame types.

In the ANOVA for the context condition, remnant type ($F_1(3,66) = 4.092$, $p = .010$; $F_2(3,9) = 1.112$, $p = .394$) and the interaction between verb frame and remnant type ($F_1(6,131) = 3.256$, $p = .005$; $F_2(6,18) = 1.240$, $p = .332$) produced weak effects that were significant only by subjects. The mean judgments for the felicitous context conditions are depicted in Figure 3. This graph shows that the remnant effect disappears in a felicitous context: The _ _ XP XP remnant is not significantly more acceptable than the other remnants. This is compatible with Kuno's (1976) account of the interaction of the constraints MINDIS and SENTP.

The ANOVA for the context condition also revealed a significant main effect

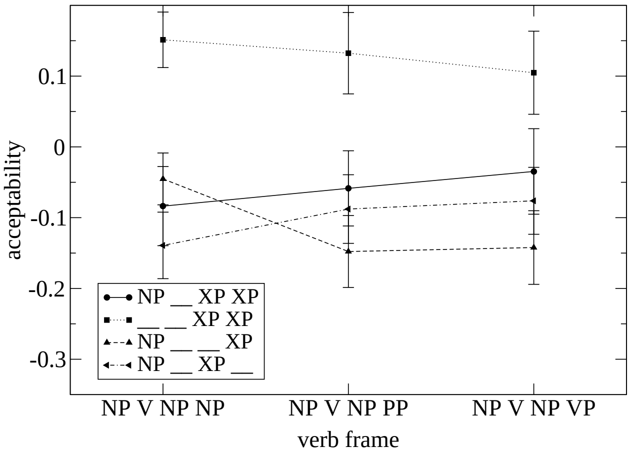

Figure 2. Effect of verb frame and remnant type on gapping (ditransitive frames, null context)

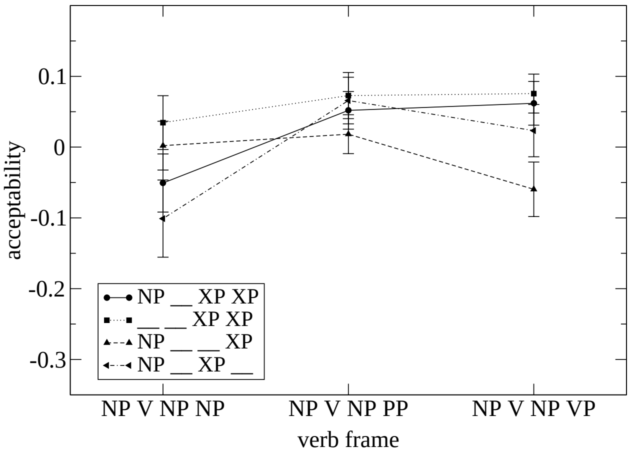

Figure 3. Effect of verb frame and remnant type on gapping (ditransitive frames, felicitous context)

of verb frame ($F_1(2,44) = 7.677$, $p = .001$; $F_2(2,6) = 15.919$, $p = .004$). A post-hoc Tukey test showed that the NP V NP NP verb frame was significantly less acceptable than both the NP V NP PP and the NP V NP VP frames ($\alpha < .05$).

2.5 Discussion

For transitive verbs, we found that gapping is equally acceptable for all types of verbal complements tested (NP, PP, VP). We also failed to find a difference between PP complements and PP adjuncts. This result settles the controversy on the status of complements and adjuncts in gapping: Hankamer (1973) claims that PP adjuncts are more acceptable than PP complements, a claim that is disputed by Jackendoff (1971) and Kuno (1976). These negative results are also important for our next experiment, as they allow us to disregard the distinction between different verb frames, and between adjuncts and complements, thus enabling us to use a more compact experimental design.

In contrast to transitive verbs, ditransitive verbs showed an effect of *Frame*: in a felicitous context, the NP V NP NP frame was less acceptable than the other frames. Note, however, that this effect, for which the literature on gapping fails to offer an explanation, is rather small (see Figure 3).

The main finding of Experiment 1 is the effect of remnant type and its interaction with context. We showed that the _ _ XP XP remnant is more acceptable than all the other remnants, an effect that is very strong in a null context, but disappears completely in a felicitous context. This provides strong evidence for Kuno's (1976) Minimal Distance Principle, and in particular for his observation that a violation of MinDis can be overridden by a satisfaction of the context requirements on gapping (his constraint SentP).

On the other hand, we found that the NP _ XP XP remnant is not significantly less acceptable than NP _ _ XP and NP _ XP _, contrary to Kuno's (1976) claim that gapping must leave behind exactly two remnants.

Now let us briefly consider an alternative explanation for the interaction of remnant type and context. One could argue that this effect is actually due to the contexts used, rather than to the stimulus sentences proper. Some initial plausibility for this view derives from the fact that two of the remnants (NP _ XP XP and _ _ XP XP) used double *wh*-questions as contexts (see (19-a) and (19-b)), while the other two remnants (NP _ _ XP and NP _ XP _) had single *wh*-questions as contexts (see (19-c) and (19-d)). It seems plausible to assume that multiple *wh*-questions are less acceptable than single ones, and maybe subjects actually took the acceptability of the context into account when they judged the acceptability of the stimulus sentences.

To test this hypothesis, an ANOVA was conducted on the contextualized data with question type as the only factor. This yielded an effect of question type which was significant by subjects ($F_1(1,2) = 8.982$, $p = .007$; $F_2(1,3) = 1.257$, $p = .344$). However, this effect went the other way than was expected: Single questions (mean $= -.0085$) were less acceptable than double questions (mean $= .0410$). This result allows us to rule out the hypothesis that the effect of *Remn* is due to the type of question used, rather than to the remnant itself.

Another alternative explanation for the remnant is that _ _ XP XP is more acceptable because it does not contain a subject pronoun. This pronoun is present in the other three remnants and might reduce acceptability in the null context condition, as it cannot be anchored to an NP in the context. This would explain why the remnant effect disappears in context, where such an antecedent is provided (see (15) and (18)). This alternative explanation for the remnant effect cannot be ruled out on the basis of Experiment 1. We will address this issue in the next experiment, which will investigate the behavior of gapping in non-felicitous contexts. A non-felicitous context provides an antecedent for the subject pronoun, but differs from a felicitous context in that it violates SENTP.

3 Experiment 2: Minimal Distance, Subject-Predicate Interpretation, Simplex Sentence, and Context

3.1 Introduction

The aim of this experiment was to replicate and extend the findings of Experiment 1. It was designed to investigate how the remnant effect found in Experiment 1 interacts with other constraints on gapping, and how it behaves in a neutral and non-felicitous context. Table 3 gives an overview of the factors included in Experiment 2. The constraints are the ones detailed in Section 1.1, either violated or not: Minimal Distance (MINDIS), Functional Sentence Perspective (SENTP), Subject-Predicate Interpretation (SUBJPRED), and Simplex-Sentential Relationship (SIMS).

The constraint MINDIS (see (4)) is satisfied if the distance between the remnants and their antecedents is minimal, as in (20-a), where *the thief* can be paired with *the criminal* and *for robbing the bank* can be paired with *for burgling the house.* (20-b), on the other hand, is in violation of MINDIS, as *she* cannot be paired with *the neighbor,* but has to be paired with the subject *he.*

(20) a. He punished the criminal for robbing the bank and the thief for burgling the house.

Table 3. Factors in Experiment 2

MinDis (*Dis*)	SubjPred (*Pred*)	SimS (*Sim*)
not violated (_ _ XP XP)	not violated	not violated
violated (NP _ _ XP)	violated	violated

SentP (*Con*)
not violated (fel. context)
violated (non-fel. context)
neutral context (control)
null context (control)

b. He helped the neighbor by doing the shopping and she by washing the dishes.
c. He punished the criminal for robbing the bank and the thief the house.
d. He helped the neighbor by doing the shopping and the friend by washing the dishes.

Another constraint on gapping postulated by Kuno (1976) is SubjPred (see (8)), which requires that the remnants left behind by gapping be interpreted as a subject and its predicate. This constraint is met in (20-a), where *the thief* is the subject of *for burgling the house*, but it is violated in (20-d), where the subject of *washing the dishes* is not the remnant *the friend*, but the main clause subject *he*.

The constraint SimS (see (12)) requires that the constituents left behind by gapping have to be part of a simplex sentence, i.e., gapping out of subordinate clauses is disallowed. This constraint is met in (20-a), where the gapped clause is interpreted as *he punished the thief for burgling the house*, while it is violated in (20-c), where the interpretation of the gapped clause is *he punished the thief for robbing the house*.

Finally, the experiment included the constraint SentP (see (6)), which governs the context required for gapping. Extending the results of Experiment 1, we included not only a felicitous context condition, in which the remnants are new while the gap is given (i.e., SentP is satisfied), but also a non-felicitous context, in which the remnants are given while the gap is new (i.e., SentP is violated). The contexts were formulated as questions, on a par with Experiment 1. In addition to the felicitous and non-felicitous contexts, we included two control conditions: a null context condition and a neutral context condition. In the null context condition, the stimuli were presented in isolation. In the neutral context condition, the stimuli were prefixed by the question *What happened?*, which indicates an all focus information structure.

The examples in (21) show the felicitous contexts that belong to the stimuli in (20), while (22) gives the corresponding non-felicitous contexts.

(21) a. Who did Michael punish, and why?
b. How did David and Hanna help the neighbor?
c. Who did Michael punish, and why?
d. Who did David help, and how?

(22) a. Why did Michael punish the criminal and the thief?
b. Who did David and Hanna help, and how?
c. Why did Michael punish the criminal and the thief?
d. How did David help the neighbor and the friend?

3.2 Predictions

3.2.1 *Constraints*

Based on the results of Experiment 1 and on the claims in the theoretical literature on gapping, we can arrive at a set of predictions regarding the constraints investigated in the present experiment.

We expect strong unacceptability for a violation of SIMS, i.e., for sentences in which the remnants are not in a simplex-sentential relationship. Intuitively, a violation of SIMS is so serious that it cannot be remedied by the satisfaction of other constraints such as MINDIS, SUBJPRED, or SENTP.

An effect of MINDIS is also predicted, i.e., structures with subject remnants (see (20-b)) are expected to be reduced in acceptability. In line with the findings of Experiment 1 this effect should disappear in a felicitous context (see (21-b)).

We also expect a significant effect of SUBJPRED; gapped sentences that do not allow a subject-predicate interpretation of the remnants (see (20-d)) are predicted to be dispreferred. In line with Kuno's (1976) observations, we expect this effect to interact with MINDIS, and possibly with SENTP, i.e., with context (even though Kuno (1976) does not explicitly mention this possibility).

Finally, Kuno's (1976) account also predicts an effect of SENTP, i.e., a felicitous context should improve the overall acceptability of a gapped sentence.

3.2.2 *Constraint Ranking*

The present experiment also allows us to test the validity of Keller's (1998) model of gradient grammaticality: We predict that the constraints tested in this experiment cluster into hard and soft constraints. Hard constraints are expected to receive a high ranking, i.e., trigger a high degree of unacceptability, while soft constraints will receive a low ranking, i.e., cause only mild unacceptability when violated.

Intuitively, SIMS is a good candidate for a hard constraint, while SUBJPRED and MINDIS are probably soft constraints. A particularly interesting question is how context interacts with soft and hard constraints. It seems plausible to expect soft constraints to be more susceptible to context effect than hard ones.

3.2.3 *Constraint Interaction*

Another prediction is that constraint violations are cumulative, i.e., that the degree of unacceptability of a sentence increases with the number of constraint violations it incurs. This finding underpins the re-ranking model of gradience. Note that Keller (1998) found that the cumulativity effect holds for both soft and hard constraint violations.

3.3 Method

3.3.1 *Subjects*

Sixty native speakers of English from the same population as in Experiment 1 participated in the experiment. None of them had previously participated in Experiment 1.

The data of two subjects had to be excluded because they were linguists. The data of another three subjects were eliminated after an inspection of their response times showed that they had not completed the experiment adequately (response times < 1s or > 100s). This left 55 subjects for analysis. Of these, 32 subjects were male, 23 female; eight subjects were left-handed, 47 right-handed. The age of the subjects ranged from 17 to 72 years; the mean was 31.6 years.

3.3.2 *Materials*

Training and Practice Materials

These were the same as in Experiment 1.

Test Materials

A full factorial design was used which included the factors *Dis*, *Sim*, *Pred*, and *Con*, representing the constraints MINDIS, SIMS, SUBJPRED, and SENTP, respectively (see Table 3 for an overview of the experimental design). The factors *Dis*, *Sim*, and *Pred* had two levels (constraint violated or not violated), while the factor *Con* had four levels: constraint violated (non-felicitous context), not violated (felicitous context), plus the two control conditions (null context and neutral context). This yielded a total of $Dis \times Sim \times Pred \times Con = 2 \times 2 \times 2 \times 4 = 32$ cells. Eight lexicalizations were used for each of the cells, which resulted in a total of 256 stimuli. A set of 24 fillers was used, designed to cover the whole acceptability range.

3.3.3 *Procedure*

Instructions, Demographic Questionnaire, Training and Practice Phase

These were the same as in Experiment 1.

Experimental Phase

The presentation and response procedures in the experimental phase were the same as in Experiment 1. A between subjects design was used to administer the experimental stimuli: Subjects in Group A judged non-contextualized stimuli, while subjects in Group B judged contextualized stimuli.

For Group A, four test sets were used: Each set contained two lexicalizations for each of the cells in the design $Dis \times Sim \times Pred$, i.e., a total of 16 items. The items were distributed over the test sets in a Latin square design. For Group B, eight test sets were used, each containing the design in one lexicalization and three contextualizations. This yielded 24 items per test set, which again were placed in a Latin square.

In Group A, each subject saw 32 items: 16 experimental items and 16 fillers. In Group B, each subject saw 40 items: 24 experimental items and 16 fillers.

Each subject was randomly assigned to a group and a lexicalization; 25 subjects were assigned to Group A, and 30 to Group B. Instructions, examples, training items, and fillers were adapted for Group B to take context into account.

3.4 Results

As in Experiment 1, all analyses were carried out on the geometric means of the normalized judgments. Separate ANOVAs were performed for the null context condition and the context condition.

3.4.1 *Constraints*

Simplex Sentence

In the null context condition, a highly significant main effect of *Sim* was found ($F_1(1,24) = 23.415$, $p < .0005$; $F_2(1,7) = 18.918$, $p = .003$). The same effect of *Sim* was present in the context condition ($F_1(1,29) = 97.310$, $p < .0005$; $F_2(1,7) = 15.548$, $p = .006$). The interaction between *Sim* and context was non-significant.

Figure 4 depicts the mean judgments for a violation of SIMS in all contexts. It indicates that SIMS violations have a strong effect on acceptability and illustrates the absence of a context effect: A violation of SIMS results in the same decrease in acceptability in all contexts (including the null context and the neutral context).

Minimal Distance

In the null context condition, a highly significant main effect of *Dis* was found ($F_1(1,24) = 25.997$, $p < .0005$; $F_2(1,7) = 14.612$, $p = .007$). *Dis* was also significant in the context condition ($F_1(1,29) = 23.315$, $p < .0005$; $F_2(1,7) = 11.421$, $p = .012$), where an interaction of DIS and SIM was also present, significant by subjects only ($F_1(1,29) = 4.568$, $p = .001$; $F_2(1,7) = 2.111$, $p = .190$).

The ANOVA also revealed a significant interaction of *Dis* and context ($F_1(2,58) = 4.568$, $p = .014$; $F_2(2,14) = 6.553$, $p = .010$). We investigated this interaction by conducting separate ANOVAs for the three context conditions. In the neutral context condition, we found a main effect of *Dis*

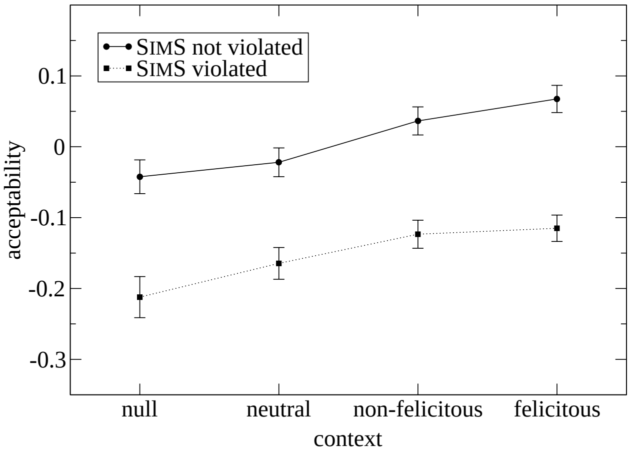

Figure 4. Context effects for SIMS

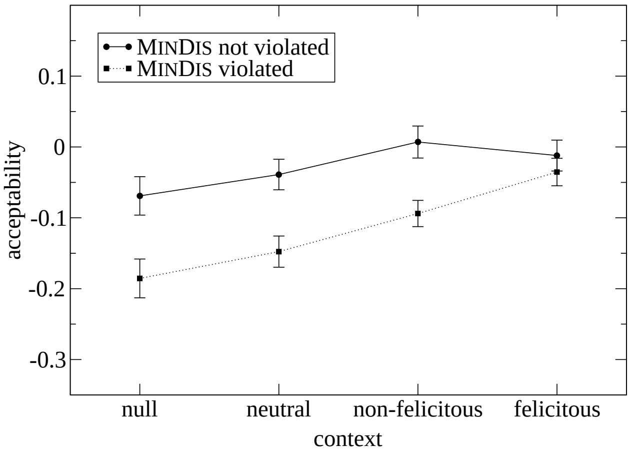

Figure 5. Context effects for MINDIS

($F_1(1,29) = 15.282$, $p = .001$; $F_2(1,7) = 11.207$, $p = .012$). Also in the non-felicitous context condition, a highly significant effect of *Dis* was obtained ($F_1(1,29) = 20.747$, $p < .0005$; $F_2(1,7) = 16.904$, $p = .005$). However, the ANOVA for the felicitous context failed to detect an effect of *Dis*. Figure 5 depicts the interaction of context with *Dis*. It shows that the effect of *Dis* disappears in the felicitous context, in line with our predictions.

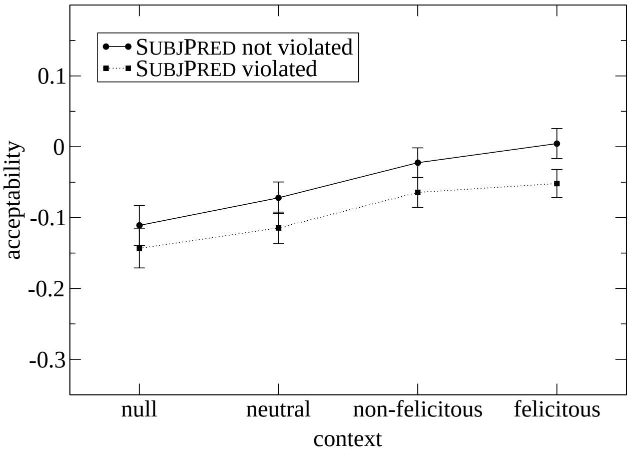

Figure 6. Context effects for SUBJPRED

Subject-Predicate Interpretation

The main effect of *Pred* failed to reach significance in the null context condition. In the context condition, a main effect of *Pred* was found ($F_1(1,29) = 19.377$, $p < .0005$; $F_2(1,7) = 9.891$, $p = .016$). The interaction of *Pred* and context failed to be significant. There was, however, an interaction of *Pred* and *Sim* that was significant by subjects only ($F_1(1,29) = 11.453$, $p = .002$; $F_2(1,7) = 2.524$, $p = .156$).

Figure 6 depicts the interaction of context with *Pred*. Note the absence of a context effect, contrary to our expectation that SUBJPRED is a context dependent constraint. However, the presence of a *Pred*/*Sim* interaction might indicate that the effect of *Sim* blocks out the context effect of *Pred*. Recall that a violation of SIMS leads to a high degree of unacceptability, while SUBJPRED only has a small effect on acceptability. It is therefore appropriate to factor out violations of SIMS (and other constraints), and to look at the effect of context on single violations of SUBJPRED. The mean judgments for single violations of SUBJPRED are depicted in Figure 7, which indicates that the effect of *Pred* in the neutral context is stronger than in the other contexts.

To confirm this observation, we conducted separate ANOVAs for single violations of SUBJPRED for the four context conditions. In the null context, the felicitous context, and the non-felicitous context, no significant effect of a single SUBJPRED violation was found. In the neutral context, however, a

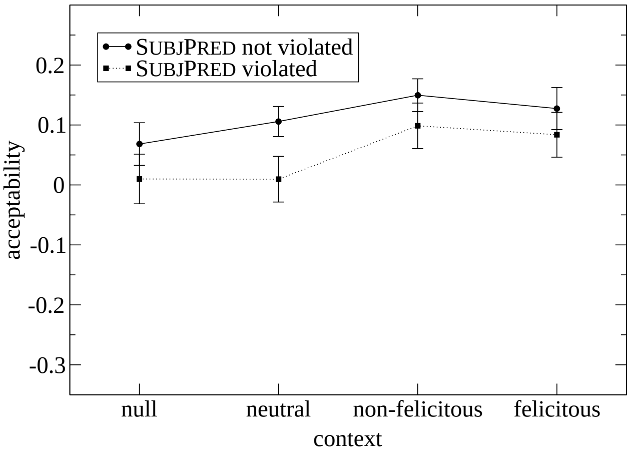

Figure 7. Context effects for SUBJPRED (single violations)

single violation of SUBJPRED led to a significant reduction in acceptability ($F_1(1,29) = 8.327$, $p = .007$; $F_2(1,7) = 5.610$, $p = .050$).

Functional Sentence Perspective

The ANOVA on the context condition showed a significant main effect *Con* ($F_1(1,29) = 10.209$, $p < .0005$; $F_2(1,7) = 13.082$, $p = .001$). A post-hoc Tukey test was conducted to investigate the locus of the *Con* effect. It was found that the neutral context was significantly less acceptable than both the felicitous and the non-felicitous context ($\alpha < .01$ in both cases). However, there was no difference between the felicitous and the non-felicitous context.

3.4.2 *Constraint Ranking*

Figure 8 compares the degree of unacceptability caused in each context by single violations of the constraints SIMS, MINDIS, and SUBJPRED. The graph indicates that a violation of SUBJPRED only has a small effect on acceptability. A violation of SIMS leads to serious unacceptability, while a violation of MINDIS is somewhere in-between.

To test if these differences in unacceptability were significant, we conducted a separate ANOVA on the subset of the data that only contained single violations. In the null context, a significant effect of constraint type was found ($F_1(2,48) = 6.817$, $p = .002$; $F_2(2,14) = 5.509$, $p = .017$). A post-hoc

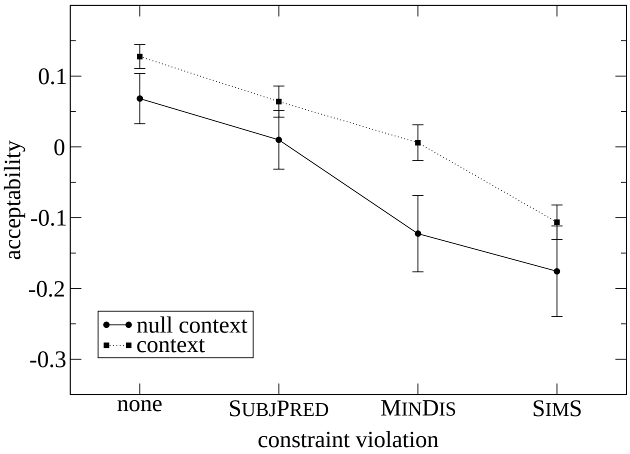

Figure 8. Effect of type of violation (single violations)

Tukey test showed that the degree of unacceptability caused by a violation of SIMS was higher than the degree of unacceptability caused by a violation of SUBJPRED (by subjects, $\alpha < .01$, and by items, $\alpha < .05$). Also, the degree of unacceptability associated with a MINDIS violation was higher than that associated with a SUBJPRED violation (by subjects only, $\alpha < .05$).

We also found a significant effect of constraint type in the context condition ($F_1(2,58) = 19.251$, $p < .0005$; $F_2(2,14) = 3.693$, $p = .052$). A Tukey test showed that a violation of SIMS caused a higher degree of unacceptability than either a violation of SUBJPRED ($\alpha < .05$) or a violation of MINDIS (by subjects only, $\alpha < .01$). The difference between MINDIS and SUBJPRED failed to reach significance in the context condition.

3.4.3 *Constraint Interaction*

To test the hypothesis that constraint violations are cumulative, we recoded the data such that the number of constraint violations was the independent variable. In the null context condition, an ANOVA on the recoded data revealed a significant effect of number of violations ($F_1(3,72) = 21.817$, $p < .0005$; $F_2(3,21) = 19.217$, $p < .0005$). Also in the context condition, a highly significant effect of number of violations was obtained ($F_1(3,87) = 65.062$, $p < .0005$; $F_2(3,21) = 24.993$, $p < .0005$).

The effect of number of violations is graphed in Figure 9. This graph shows

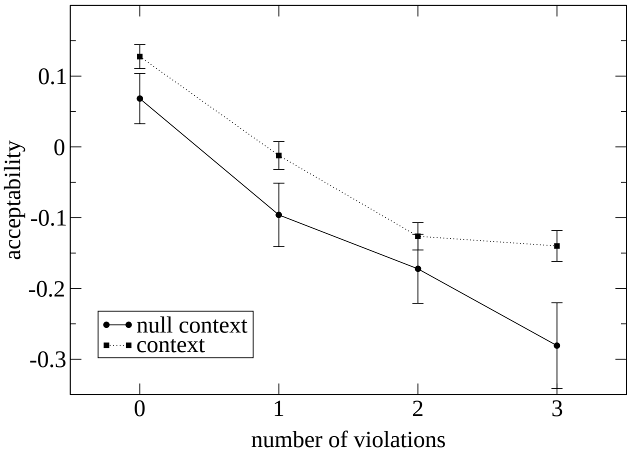

Figure 9. Effect of number of violations

a consistent cumulativity effect for both the null context and the context condition. A post-hoc Tukey test was conducted to locate the effect of number of violations. For the null context condition, it was found that a single violation was significantly less acceptable than zero violations (by subjects, $\alpha < .01$, and by items, $\alpha < .05$). The difference between one and two violations failed to be significant, but two violations were significantly less acceptable than zero violations ($\alpha < .01$). The difference between two and three violations was again not significant, but three violations were significantly less acceptable than one violation ($\alpha < .01$).

The same post-hoc test was conducted for the context condition. Again, it was found that one violation was less acceptable than zero violations ($\alpha <$ .01), while two violations were less acceptable than one violation ($\alpha < .01$). The difference between two and three violations was again too small to reach significance, but the three violations were significantly less acceptable than one violation ($\alpha < .01$).

3.5 Discussion

3.5.1 *Constraints*

Experiment 2 found main effects of *Sim*, *Dis*, and *Pred*. This demonstrated that violations of the constraints MINDIS, SUBJPRED, and SIMS significantly reduce the acceptability of gapped sentences, as predicted by Kuno's

(1976) account of gapping. A main effect of *Con* was also present, but contrary to predictions, no difference between the acceptability of gapping in a felicitous and a non-felicitous context was found. However, the acceptability of gapping in the felicitous and the non-felicitous context was significantly higher than in the neutral context. This seems to indicate that even a non-felicitous context provides an information structure that is partially compatible with the requirements of the constraint SentP.

We also found that SentP interacts with other constraints on gapping. A significant interaction of *Con* and *Dis* was obtained: A violation of MinDis leads to reduced acceptability in the null context, the neutral context, and the non-felicitous context. In the felicitous context (that satisfies the information structure constraint SentP), the effect of *Dis* disappeared. Note that the null context and the neutral context behaved in the same fashion with respect to MinDis violations; this is expected based on the hypothesis that even a null context carries implicit information structural assumptions, and is interpreted by subjects on a par with a neutral (all new) context.

Similar to the *Dis* effect, the effect of *Pred* was also found to be context dependent. Considering stimuli that incur a single violation of SubjPred, we found a significant effect of *Pred* only in the neutral context; in the felicitous and non-felicitous contexts, the effect of *Pred* was too small to be significant. Also, in the null context, no effect of *Pred* was found, even though this would be expected under the assumption that the null context behaves like a neutral (all new) context.

In contrast to MinDis and SubjPred, the Simplex S constraint SimS was found to be immune to context effects: It caused consistently strong unacceptability, independent of which context was presented. This is in line with our predictions regarding the behavior of SimS.

Another one of Kuno's (1976) observations can be tested against the data from Experiment 2. Examples like (9) and (10) seem to indicate that a satisfaction of SubjPred can override a violation of MinDis. However, we failed to find an interaction of *Dis* and *Pred* in either the null context condition or the context condition. This might indicate that the interaction of SubjPred and MinDis that Kuno (1976) observes is limited to examples like the ones in (9) and (10), and does not generalize to our experimental stimuli.

Finally, the results of the present experiment allow us to evaluate the alternative explanation for the *Dis* effect we discussed in Section 2.5: The _ _ XP

XP remnant is more acceptable than the XP _ _ XP remnant because the latter contains a subject pronoun, which reduces acceptability if it is not contextually anchored (in a null or neutral context). This explanation can be ruled out on the basis of Experiment 2, which demonstrated a *Dis* effect for the non-felicitous context condition, i.e., even if the subject pronoun can be anchored to a contextually given NP.

3.5.2 *Constraint Ranking*

A second set of predictions for Experiment 2 was based on Keller's (1998) model of gradient grammaticality as constraint re-ranking. This model rests on the assumption that constraints cluster into hard constraints (that lead to serious unacceptability) and soft constraints (that cause only mild unacceptability). Consider Figure 8, which graphs the unacceptability incurred by single violations of the three constraints SimS, MinDis, and SubjPred. We found that a SimS violation was significantly more serious than a violation of MinDis, which in turn was significantly more serious than a violation of SubjPred, leading to the overall ranking of SimS $\gg$ MinDis $\gg$ SubjPred. We conclude that SimS qualifies as a hard constraint, as it leads to strong unacceptability, while SubjPred induces only mild unacceptability and thus should be classified as soft. The status of MinDis is less clear, as it falls in-between these two extremes.

Note, however, that we also observed that the soft constraint SubjPred was subject to contextual variation (consider the increased effect of a SubjPred violation in the neutral context). On the other hand, SimS, a hard constraint, was immune to context effects. This leads to the more general hypothesis that soft constraints are subject to context effects, while hard constraints are immune to contextual variation. If correct, this hypothesis would provide us with a new diagnostic for the hard/soft distinction, in addition to constraint strength (proposed in Keller 1998). Based on this hypothesis, we can classify MinDis as a soft constraint, as it is clearly subject to context effects, even though its constraint strength is relatively close to that of SimS, a hard constraint.

3.5.3 *Constraint Interaction*

The findings of Experiment 2 confirm another assumption on which the reranking model rests: Constraint violations are cumulative, i.e., the degree of unacceptability increases with the number of violations. A clear cumulativity effect was obtained for both the null context condition and the context condition (see Figure 9).

4 General Discussion

4.1 Implications for Linguistic Methodology

This paper is part of a line of research that draws on the experimental paradigm of magnitude estimation to obtain linguistic judgment data that are reliable and maximally delicate. This line of research, which was initiated by Bard et al. (1996) and Cowart (1997), has contributed to linguistic theory by settling data disputes that could not be resolved solely on the basis of intuitive, informal acceptability judgments. Relevant experimental findings have been obtained in studies on extraction (Cowart 1989, 1997, Keller 1996, 1997), binding theory (Cowart 1997), unaccusativity (Sorace 1993*a*,*b*, 2000), and word order (Keller and Alexopoulou 2000, Keller 2000*a*).

The results of Experiments 1 and 2 confirm the usefulness of an experimental approach to linguistic data by applying magnitude estimation to gapping constructions. Experiment 1 showed that PP adjuncts and PP complements are equally acceptable as remnants in gapping, a fact that has been surrounded by controversy in the theoretical literature. It also provided evidence against the claim that gapping must leave behind exactly two remnants (Kuno 1976). Another theoretically interesting result is that subject remnants are less acceptable than object remnants, an effect that turned out to be context dependent. Experiment 2 confirmed this result and provided evidence for another context dependent constraint on gapping (Subject-Predicate Interpretation), but also discovered a constraint that is immune to context effects (Simplex S). More importantly, Experiment 2 provided data on how the constraints on gapping interact, i.e., on what happens if more than one constraint is violated. Such interaction data, which cannot easily be obtained with the traditional intuitive approach, allows us to make observations on how constraints compete, and thus can inform an optimality theoretic model that deals with gradient linguistic data (Hayes 2000, Hayes and MacEachern 1998, Keller 1998, Müller 1999).

4.2 Implications for Optimality Theory

The interaction of *Dis*, *Pred*, and *Con* demonstrated in Experiment 2 can be regarded as evidence that gapping is subject to constraint competition, a fact that was already noted by Kuno (1976) (who, however, did not have the conceptual tools of modern Optimality Theory at his disposal).

Offering a detailed analysis of the experimental data based on an optimality theoretic model is beyond the scope of the present paper. The reader is referred Keller (2000*b*), who presents both an explicit model of gradience in Optimality Theory, and a detailed account of the gapping data from Experiments 1 and 2.

4.3 Implications for the Re-Ranking Model

The work presented in this paper provided additional evidence for two central assumptions underlying the re-ranking model of gradience (Keller 1998): First, the experimental data confirmed the cumulativity of constraint violations assumed by the re-ranking model. In addition, the results support the soft/hard distinction of constraint violations previously demonstrated for extraction. Context effects on gapping were also investigated, and we arrived at the hypothesis that soft constraints are subject to context effects, while hard constraints are immune to contextual influences. If correct, this hypothesis would provide us with an additional diagnostic for the hard/soft distinction. This has to be validated in further experimental work.

Also, the present results allow us to speculate on the theoretical status of hard and soft constraints, and its implications for grammar architecture. One possible line of argumentation is that soft constraints are limited to the interface level of the grammar (syntax-semantics, syntax-pragmatics, syntax-lexicon), while hard constraints are internal to syntax. This would explain why soft constraints cause only weak acceptability effects and can be overridden by context, while hard violations cause strong unacceptability and are immune to context effects.

The constraints identified as soft in the present study belong to the syntax-semantics or syntax-pragmatics interface (Minimal Distance, Subject-Predicate Interpretation), while the hard constraint (Simplex S) seems to be syntactic in nature. This observation squares well with previous results on extraction, where constraints on phrase structure, agreement, and subcategorization were found to be hard, while soft constraints included referentiality and definiteness, i.e., constraints located at the syntax-semantics interface.

Notes

Thanks to Mark Steedman for important advice on the work reported here. Comments on earlier stages of this paper were provided by Maria Lapata and by the audience of the Workshop on Competition in Syntax, Constance, March 1999. The support of an ESRC Postgraduate Research Studentship is also acknowledged.

[1] All examples in this section are taken from Kuno (1976).

[2] We supply constraint names for notational convenience.

[3] This terminology should not be taken to imply that hard constraints are inviolable, while soft constraints are violable in an optimality theoretic sense. The soft/hard distinction is an empirical one, based on the acceptability profile of a constraint.

[4] Note however, that a re-ranking model can only explain cumulative violations of *different* constraints. Cumulative violations of the *same* constraint are not predicted to lead to an increase of unacceptability, as they can be dealt with by a single re-ranking (see Keller 1998 for details).

[5] Consider the following examples from Hankamer (1973), which are analogous to our sentences (15-b) and (15-d) (the acceptability judgments are his):

(i) a. *Max wanted to put the eggplant on the table, and Harvey in the sink.
    b. ?Max writes plays in the bedroom, and Harvey in the basement.

References

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

Chomsky, Noam
1995 *The Minimalist Program*. Cambridge, MA: MIT Press.

Cowart, Wayne
1989 Illicit acceptability in *picture* NPs. In: Caroline Wiltshire, Randolph Graczyk and Bradley Music (eds.) *Papers from the 25th Meeting of the Chicago Linguistic Society*, vol. 1: The General Session, 27–40. Chicago.

Cowart, Wayne
1997 *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage Publications.

Hankamer, Jorge
1973 Unacceptable ambiguity. *Linguistic Inquiry* 5: 17–68.

Hayes, Bruce P.
2000 Gradient well-formedness in Optimality Theory. In: Joost Dekkers, Frank

van der Leeuw and Jeroen van de Weijer (eds.) *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford: Clarendon Press.

Hayes, Bruce P. — Margaret MacEachern
1998 Folk verse form in English. *Language* 74(3): 473–507.

Jackendoff, Ray S.
1971 Gapping and related rules. *Linguistic Inquiry* 2: 21–35.

Keller, Frank
1996 How do humans deal with ungrammatical input? Experimental evidence and computational modelling. In: Dafydd Gibbon (ed.) *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielefeld, October 1996*, 27–34. Berlin: Mouton de Gruyter.

Keller, Frank
1997 Extraction, Gradedness, and Optimality. In Alexis Dimitriadis, Laura Siegel, Clarissa Surek-Clark, and Alexander Williams, eds., *Proceedings of the 21st Annual Penn Linguistics Colloquium*, 169–186. (Penn Working Papers in Linguistics, no. 4.2.) Department of Linguistics, University of Pennsylvania.

Keller, Frank
1998 Gradient grammaticality as an effect of selective constraint re-ranking. In: M. Catherine Gruber, Derrick Higgins, Kenneth S. Olson and Tamra Wysocki (eds.) *Papers from the 34th Meeting of the Chicago Linguistic Society*, vol. 2: The Panels, 95–109. Chicago.

Keller, Frank
2000a Evaluating competition-based models of word order. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Philadelphia, PA.

Keller, Frank
2000b Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality. PhD thesis, University of Edinburgh.

Keller, Frank — Theodora Alexopoulou
2000 Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition*, to appear.

Keller, Frank — M. Corley — S. Corley — L. Konieczny — A. Todirascu
1998 *WebExp: A Java Toolbox for Web-Based Psychological Experiments*. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh.

Kuno, Susumo
1976 Gapping: A functional analysis. *Linguistic Inquiry* 7: 300–318.

Legendre, Géraldine — C. Wilson — P. Smolensky — K. Homer — W. Raymond

1995 Optimality and *wh*-extraction. In: Jill Beckman, Laura Walsh Dickey and Suzanne Urbanczyk (eds.) *Papers in Optimality Theory*, 607–636. (University of Massachusetts Occasional Papers in Linguistics 18) University of Massachusetts, Amherst.

Lodge, Milton
1981 *Magnitude Scaling: Quantitative Measurement of Opinions*. Beverley Hills, CA: Sage Publications.

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

Prince, Allan — Paul Smolensky
1993 *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report 2, Center for Cognitive Science, Rutgers University.

Prince, Allan — Paul Smolensky
1997 Optimality: From neural networks to universal grammar. *Science* 275: 1604–1610.

Ross, John R.
1970 Gapping and the order of constituents. In: Manfred Bierwisch and Karl Erich Heidolph (eds.) *Progress in Linguistics: A Collection of Papers*, 249–259. The Hague: Mouton.

Schütze, Carson T.
1996 *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Sorace, Antonella
1992 Lexical conditions on syntactic knowledge: Auxiliary selection in native and non-native grammars of Italian. Ph.D. dissertation, University of Edinburgh.

Sorace, Antonella
1993a Incomplete vs. divergent representations of unaccusativity in non-native grammars of Italian. *Second Language Research* 9: 22–47.

Sorace, Antonella
1993b Unaccusativity and auxiliary choice in non-native grammars of Italian and French: Asymmetries and predictable indeterminacy. *Journal of French Language Studies* 3: 71–93.

Sorace, Antonella
2000 Gradients in split intransitivity: Auxiliary selection in Western European languages. *Language*, to appear.

Stevens, Stephen S.
1975 *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New York: John Wiley.

Tesar, Bruce — Paul Smolensky
1998 Learnability in Optimality Theory. *Linguistic Inquiry* 29(2): 229–268.