



ELSEVIER

Available online at www.sciencedirect.com

Lingua xxx (2004) xxx–xxx

Lingua

www.elsevier.com/locate/lingua

Magnitude estimation and what it can do for your syntax: some wh-constraints in German

Sam Featherston*

Universität Tübingen, SFB441 Linguistic Data Structures, Nauklerstr. 35, D-72074 Tübingen, Germany

Abstract

In this paper, we explore some of the insights into the grammar that become available with the use of a more strictly controlled judgement elicitation method, magnitude estimation (Bard et al, 1996; Cowart, 1997). In particular, we focus on wh-movement in German, and show how a range of assumptions, both specific to German grammar and more generally in syntactic study are made questionable. We apply this methodology to show that German respects superiority and discourse-linking (*sensu* Pesetsky, 1987), in contrast to the standard view in the literature, but in line with the predictions of generative grammar. But we further argue that this data type, and the gradient grammaticality that it reveals (Keller, 2000), permits us further insights into the nature of the grammar. The empirical results suggest that both the superiority effect and the discourse linking effect have rather different features to those generally assumed. On this basis, we advance novel suggestions as to the precise nature of these constraints. But more generally, we argue that this data type supports the hypotheses that grammatical constraints are violable, and that grammatical constraint violations are not necessarily fatal. We argue that the acceptance of these non-standard assumptions about the structure of the grammar is necessary since the abstractions underlying the standard assumptions about grammaticality are obscuring relevant information and distorting the data base of grammatical theory.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Discourse-linking; Empirical syntax; Magnitude estimation; Superiority

Generative grammar has had a remarkably good four decades. Researchers have developed and refined generative analyses of an ever-widening range of syntactic

* Tel.: +44 70712977152; fax: +44 7071295830.

E-mail address: sam.featherston@uni-tuebingen.de.

phenomena, and the range of languages studied continues to rise. This has led to significant new insights: perspectives on syntactic structure have matured to an extent that views held in previous decades sometimes appear quaint. In some ways however, progress has been less rapid than might have been expected. In spite of the larger data base and descriptive coverage of generative syntax, many core syntactic issues have still not been settled finally, and the analysis of many constructions still remains controversial. For example, there are still sharply differing competing suggestions about the nature of binding: the Binding Theory (e.g. Chomsky, 1981) has undergone many revisions and reformulations, but still faces stiff competition from alternatives such as the binding theory based on the obliqueness hierarchy (e.g. Pollard and Sag, 1994), the reflexivity account of Reinhart and Reuland (1993) and pragmatic approaches such as Levinson (1991). It would appear that, while the dominant approaches used within generative syntax are very successful in stimulating hypothesis generation and innovative analyses, they do not so readily permit the filtering of these hypotheses and their evaluation. It should go without saying that the relatively restricted capacity of the current research paradigm to produce firm answers to long-standing questions such as the nature of binding must be considered a serious failing and cause for concern. One factor contributing to this is the restricted data basis of much work in syntax, and it is no doubt in the light of this state of affairs that the current interest in new data sources for syntactic theory has developed.

One approach to this is essentially comparative: researchers compare data across languages and across constructions and thus broaden the sample of the data underlying theory. An alternative approach is to apply evidence from different data types or produce more precise data points within specific fields. Many syntacticians are making increased use of linguistic corpus data, but another recent approach is to use experimental techniques from psychology and psycholinguistics to produce more finely grained introspective data (Coward, 1997; Keller, 2000; Sorace and Keller, this issue). By doing this, linguists hope to obtain not only more detail about the descriptive issues concerned but also insight into the nature of the judgements involved and the factors affecting grammaticality which are reflected in them. It is this strategy which we apply here. Let us note straight away that this approach requires the jettisoning of a cherished convention of generative syntax. It is generally accepted within generative grammar that its data base should be idealized by abstracting from all confounding factors (“... ideal speaker-listener, ...” etc. Chomsky, 1965:3), but it is also clear that syntacticians also standardly utilize an idealized model of grammaticality: a given example sentence is generally evaluated as either grammatical or ungrammatical. Notice that we shall refer to this assumption in the text as the *binary model* for the sake of conciseness, and because it incorporates the idea that on this view the vast majority of sentences are either grammatical or ungrammatical, and that those which do not fit easily into either of these categories are in some way peripheral. Chomsky (1965, 11; see also 1957:36; 1964 *passim*) illustrates this well. He is at once well aware that grammaticality is “a matter of degree” but develops a model in which this fact is abstracted over, a model whose aim is “to separate the *grammatical* sequences (...) from the *ungrammatical* sequences” (1957:13).

One way that linguists have dealt with recalcitrant data is to introduce one, or occasionally two or more intermediate values (for some extreme cases see Lakoff,

1973; Müller, 1995; Wurmbrand, 2001). Example (1)a below is clearly grammatical, while example (1)b is equally plainly not a well-formed sentence of standard English.

- (1) a. Tony says he loves cherries.
b. *Tony says he love cherries.
c. ?Tony groans he loves cherries.
d. ??Cherries, Tony groans he loves.

The next pair of examples (1)c and (1)d are intermediate cases which are increasingly odd, though not as obviously ill-formed as (1)b. That this is a simplification is generally recognized (e.g. Schütze, 1996) and indeed immediately clear from such example sets as (1), since even here clear distinctions are difficult to draw. Occasional attempts have been made to exploit the detail which is generally abstracted over (e.g. the influential Ross, 1972; Lakoff, 1973, for recent overviews see Keller, 2000; Manning, 2003), but we may generalize that the binary model remains the standard assumption, since it is still generally accepted that there is such a thing as absolute (un)grammaticality, and that marginal data are merely in a grey zone between the two fixed points. The idea that grammaticality is a true continuum, as we shall assume here, is non-standard. We discuss in more detail below what we understand by *grammaticality* in this context.

In this article, we provide an overview of some studies that we have undertaken in connection with this, and discuss some of the more interesting findings. We shall report three studies, concentrating on the more general implications of the findings rather than on finer points of the experimental design and implementation and details of the results. We have discussed some of the more German-specific aspects of two of these studies in Featherston (forthcoming), where we used the data to decide between competing generative analyses of the structure of German. Here, we take a step back and consider the broader implications for the grammar and for grammaticality.

1. The methodology

One major reason for the ubiquity of the simplification of the scale of grammaticality has been that the individual linguist's judgements, the data type most frequently used, do not reliably permit much greater detail. The scale of grammaticality adopted was thus suited to the data available, but more finely grained data is now obtainable by the adaptation of psychological testing procedures to linguistic data elicitation. In the studies reported here, we employed the *magnitude estimation* method (Bard et al., 1996; Stevens, 1975, see also Sorace and Keller, this issue), which allows the elicitation of grammaticality judgements with only a minimum of restriction on the informant from the nature of the scale itself. The method is discussed at length in Bard et al. (1996) and Cowart (1997) so we shall only sketch it briefly here. In this procedure, subjects are asked to provide grammaticality judgements about the well-formedness of example sentences, and express these judgements in numerical form. In the studies reported here, subjects were asked whether the experimental sentences “sounded natural”, to avoid triggering any feeling that they were being asked for normative judgements. They were instructed that the object of

interest was the spoken language, rather than the written form, and that they should complete the task briskly, without thinking too deeply about any given sentence since it was their first impressions which were of interest.¹

A procedure such as magnitude estimation allows more finely differentiated grammaticality judgements to be gathered because of three factors. First, multiple lexicalizations of each structure are used, which greatly reduces irrelevant lexical variation between syntactic conditions. Second, these lexicalizations are closely matched, which minimizes background noise within conditions. Third, idiomatic variation is largely eliminated by obtaining judgements from multiple informants. Last, and specific to magnitude estimation, informants are enabled to express their intuitions without any restrictions of the judgement scale. Subjects are asked to provide purely comparative judgements: these are relative both to a reference item and the individual subject's own previous judgements, but at no point is an absolute criterion of grammaticality applied. Also, all judgements are proportional; i.e. subjects are asked to state not only if sentence A is better or worse than sentence B, but how many times better or worse A is than B. Next, the subjects themselves fix the value of the reference item relative to which subsequent judgements are made. The scale on which judgements are made is open-ended and has no minimum division: subjects can always add a further highest score or produce an additional intermediate rating. The net result is that subjects are able to produce judgements which distinguish all and only the differences they perceive. When the limitation to a scale selected by the linguist is removed, the results obtained exhibit more differentiation than conventional judgements are assumed to contain. But this additional information is an inherent part of grammaticality judgements, which previous collection methods were insufficiently sensitive to reveal, but which were nevertheless potentially present. This new information need therefore not be regarded as a separate, new data type but rather as a "closer look" at standard introspective judgements. We feel a need to underline this point since it is frequently raised in discussion.

The validity of this can be supported in several ways. Firstly, subjects show a strong tendency to produce parallel responses. The mean responses obtained are thus not merely averages of independent samples but the reflection of robust trends across subjects. This is reflected in the statistical tests which can be applied to magnitude estimation results. Secondly, the mean judgements obtained can be crossed-checked with conventional introspective judgements. Informants shown magnitude estimation results generally agree that clear differences between mean scores correspond to perceptible differences between the structures. The method produces no surprises therefore: it merely confirms independently what we might have suspected anyway, but could not demonstrate. Additionally, the judgements obtained are remarkably stable even across studies. For example, grammatical variants *A*, *B* and *C* of a structure *S*, which exhibit a pattern of scores in the proportion *a:b:c* will generally exhibit the same pattern of scores if tested in a parallel structure *S'* which has additionally the feature *F*. *F* will of course produce an overall shift in the pattern, and make

¹ Judging example sentences requires more time when they are ambiguous, or when the availability of different readings is to be judged, but these more sophisticated judgement tasks were not required here – only the single most accessible reading was to be judged. There was therefore no reason to expect an improvement in quality over time. Another reason for specifying first reactions as of interest is to reassure participants that they should trust their *Sprachgefühl*.

the data from the two studies non-aligned, but interactions between grammatical factors seem not to occur: the relative proportions typically remain constant. This stability supports the proposition that the variation between *A*, *B*, and *C* is based upon some essential features of their constructions, not merely their surface plausibility, which is also visible (and as far as possible controlled for) in the data. Let us also note here that the scalar pattern of judgements obtained with magnitude estimation is not exclusive to this one methodology – other methods produce it too. For example, when multiple subjects are asked for judgements on a simple five-point scale a continuum inevitably arises (e.g. Crain and Fodor, 1987). A continuum is produced even when multiple informants are instructed to judge on a binary scale, since the proportion of subjects accepting and rejecting marginal structures itself forms a continuum, which may be assumed to be proportional to their well-formedness. This is the basis of the speeded grammaticality judgement methodology (e.g. Frazier, 1985; Bader, 1996).² The prevalence of the linguist using just their own introspective judgements has no doubt tended to obscure the fact that the binary model is an abstraction, and that the primary data of judgements has a scalar character. Since this data type behaves like conventional grammaticality judgements and there is no good reason to suspect that it is measuring anything other than these do, we shall refer to the judgements we elicit as *grammaticality judgements* and the construct measured as *grammaticality* as a convenient shorthand, whilst readily accepting that factors which are not “narrowly grammatical” can affect them, just as these factors affect conventional judgements.³

In the remainder of the article, we shall show some examples of the ways in which magnitude estimation can fruitfully be used and highlight some implications which can be drawn from the results of these studies. One application of magnitude estimation data is in testing hypotheses about the grammar in cases where conventional introspective judgements fail to provide sufficient clarity for decisions to be made with any confidence. In this way, the approach can usefully augment the methods more usually applied in syntax, which, as we have noted, provide only a weak basis for hypothesis evaluation. We shall see an example of this in our testing for superiority and discourse linking effects in German – both of these can be shown to exist in German in spite of claims in the literature to the

² In fact the only way of producing exclusively binary judgements is reduce the number of informants to one. Two informants judging on a binary scale are sufficient to produce a three-valued scale of well-formedness ({OK,OK}, {OK, *}, {*,*}).

³ For clarity, let us specify what we mean by “grammaticality” here: “well-formedness as elicited in introspective judgements, but which is due neither to lexical factors, plausibility, processing constraints, contextual felicity, nor any other performance factor”. There are two possible differences to Chomsky’s “grammaticalness” (1965): first, our definition is explicitly linked to a data type, but Chomsky (1957: 13) chooses to “assume intuitive knowledge of the sentences of English” as his “behavioural criterion for grammaticalness”, so we are in good company here. The second possible difference is dependent on our non-adoption of the idealization to a binary model of grammaticality. This position necessitates that we consider any consistent, structure-dependent contrast in judgements to be a diagnostic for a grammatical feature, even if the contrast is not sufficiently severe to render a sentence fully unacceptable, unless of course it can be attributed to the performance-related factors mentioned above. Note that this specification of the word “grammaticality” is in line with its use in a loose sense in the syntactic literature to refer to the status of example structures (for example, in the recent Handbook of Contemporary Syntactic Theory (Baltin and Collins, 2001), 23 of the 30 authors use this term in this sense). The only difference between our use and the standard loose use in the literature is that necessitated by our non-adoption of the binary model.

contrary. In this way, syntactic theory building can be supported by greater precision in the data set. Grammatical models can be more successful when the descriptive facts are that they are aiming to capture are clearer.

But the finer definition of magnitude estimation data can also provide new insights into structure and support new syntactic analyses. Here, we shall proceed rather more speculatively and provocatively, but hope to demonstrate that the new quality of data may require new thinking in syntax. Our approach here is explicitly data-driven: we consider just the evidence of the introspective judgements and consider atheoretically what conclusions it supports for the structures we consider. This requires some reassessment of standard assumptions about grammars and grammaticality – but the aim is to evaluate the evidence theory-independently, following the empirical facts wherever they may lead us, bearing in mind that the evidence we have is the primary linguistic data that theories of grammar are supposed to account for. We do not make any strong claims for the correctness of the accounts we offer – this would require further empirical testing – but we think that syntacticians should benefit from exposure to the alternative explanatory perspectives.

2. Superiority in German

One of our chief focuses of interest has been wh-movement in German. German behaves very similarly to English in many facets of wh-movement; for example, one and only one wh-item raises compulsorily to sentence-initial position, there is a verb movement absent from embedded wh-questions which occurs in matrix wh-questions, and both preposition stranding and pied piping are possible, although more restrictedly in German. However, there are also a number of contrasts between the two languages which have stimulated considerable interest. This has had considerable influence in particular on views of the structure of the German clause, since they have often had implications for the role of subjects and thus for the empirical basis of an IP in German. We shall briefly outline the relevant constraints here and then discuss the evidence for their existence in German.

The first of the movement constraints we address is superiority, which seems a fairly straightforward phenomenon in English. In a pair of sentences such as (2)ab below, the first is a possible sentence, but the second is not.

- (2) a. Mary asked who read what.
 b. *Mary asked what who read.

The generalization seems to be that there is a correlation between the canonical position of a grammatical function and the extractability of its wh-form in multiple wh-questions. This is most obvious in a subject–object asymmetry; while a wh-subject can readily raise into the clause-initial position no matter what other in-situ wh-item there may be – (2)a, an object cannot. If non-subject wh-items are raised to spec-CP leaving a subject wh-item in situ, the sentence is markedly degraded – (2)b.

In German on the other hand, the existence of superiority is generally denied (Grewendorf, 1988; Müller, 1991; Haider, 1993; Lutz, 1996; Fanselow, 2001). The reason for this is that it is possible to generate grammatical sentences which violate the superiority

constraint as formulated for English. For instance, examples (3)a and (3)b show nothing like the clear grammaticality difference we find in their English equivalents (2).

- (3) a. Maria fragt wer was gelesen hat.
 Maria asks who what read has
 b. Maria fragt was wer gelesen hat
 Maria asks what who read has

Since counter-examples can be found, the standard view is that superiority does not apply or does so only in more restricted circumstances than in English (Wiltschko, 1997). Syntacticians tend to attribute this to some difference in structure between English and German, most commonly a divergence in clause structure, but it is fair to say that there has been no consensus on exactly what this difference might be, no doubt in part because the motor of superiority has not been uniquely identified even in languages such as English where its application is uncontroversial.

The two aims of our experiment were to test whether the superiority effect existed in German, and, if so, to gather evidence to which would permit us to understand it better. Several different accounts of the motor of superiority have been put forward, such as the Empty Category Principle (e.g. Lasnik and Saito, 1984) and Shortest Move (Chomsky, 1993), but as yet these have been largely empirically indistinguishable. We do not intend to discuss this aspect of the experiment or its results here (see Featherston (forthcoming) for more detailed presentation and analysis).

3. Superiority experiment

In this study, we tested all 26 feasible multiple wh-questions made up of the following wh-items: subject *wer* (“who”), direct object *was* (“what”), discourse-linked direct object *welches X* (“which X”), indirect object *wem* (“to whom”), discourse-linked indirect object *welchem X* (“to which X”), and temporal adjunct *wann* (“when”), hoping thus to establish whether German has such an effect, and if so, which combinations of grammatical functions as wh-items would trigger it.

These syntactic conditions were realized in 26 lexical variants. The structures and lexis in the experimental materials were strictly controlled to minimize background variation. All subjects and indirect objects were animate, and all direct objects inanimate (since these are the unmarked values in these positions). NPs were matched for length in letters and lemma frequency from the CELEX lexical database (MPI Nijmegen).⁴ The 26 target structures are shown in Table 1.

We exemplify these materials below in (4), showing the first five structures as variants of the lexical base *Der Zahnarzt hat dem Patienten die Zahnpasta empfohlen* (“the dentist has recommended the toothpaste to the patient”).

⁴ NP length range 7–11 letters, mean lengths: subjects 8.6, indirect objects 8.6, direct objects 8.6. Frequencies (mean per million figures): subjects 30.4, indirect objects 41.6, direct objects 35.5. These linguistic materials are at <http://www.sfb441.uni-tuebingen.de/~sam/db/soup.maz1.html>.

Table 1
Superiority experiment design

Moved wh-items	In-situ wh-items					
	wh-subj:	wh-DO:	wx-DO:	wh-IO:	wx-IO:	wh-adj:
wh-subj:		wer ... was	wer ... welches X	wer ... wem	wer ... welchem X	wer ... wann
wh-DO:	was ... wer			was ... wem	was ... welchem X	was ... wann
wx-DO:	welches X ... wer			welches X ... wem	welches X ... welchem X	welches X ... wann
wh-IO:	wem ... wer	wem ... was	wem ... welches X			wem ... wann
wx-IO:	welchem X ... wer	welchem X ... was	welchem X ... welches X			welchem X ... wann
wh-adj:	wann ... wer	wann ... was	wann ... welches X	wann ... wem	wann ... welchem X	

Note that “wh-DO” indicates non-discourse-linked direct object wh-item, i.e. *was* “what”, while “wx-IO” indicates discourse-linked indirect object wh-item, i.e. *welchem X* “to which X”.

- (4) a. wh-subj wh-DO
 Wer hat dem Patienten was empfohlen?
 who has to.the patient what recommended
 “Who has recommended what to the patient?”
- b. wh-subj wx-DO
 Wer hat dem Patienten welche Zahnpaste empfohlen?
 who has to.the patient which toothpaste recommended
 “Who has recommended which toothpaste to the patient?”
- c. wh-subj wh-IO
 Wer hat wem die Zahnpaste empfohlen?
 who has to.whom the toothpaste recommended
 “Who has recommended the toothpaste to whom?”
- d. wh-subj wx-IO
 Wer hat welchem Patienten die Zahnpaste empfohlen?
 who has to.which patient the toothpaste recommended
 “Who has recommended the toothpaste to which patient?”
- e. wh-subj wh-adj
 Wer hat dem Patienten die Zahnpaste wann empfohlen?
 who has to.the patient the toothpaste when recommended
 “Who recommended the toothpaste to the patient when?”

Thirty-eight subjects took part in this experiment. Each saw a version of the materials such that each syntactic condition and each lexical variant appeared once, randomly mixed among another eighteen items functioning as fillers. Participants were asked to supply their name, age, sex, occupation and dialect background.⁵ The participants were recruited by flier in the student dining hall, and offered a financial incentive to take part. The experiment was performed remotely using the package WebExp (Keller et al., 1998, see <http://www.language-experiments.org/>). The data was normalized by transformation to z-scores. This effectively unifies the different scales that the individual subjects adopted for themselves, and allows us to inspect the results visually.

4. Results

The most significant result for our present purpose is presented in Fig. 1, which shows the mean normalized grammaticality judgement score and 95% confidence intervals as error bars. Higher scores indicate greater perceived naturalness, but note that these scores are purely relative: there is no point which indicates absolute (un)grammaticality. Along the horizontal axis, the structures are grouped by the in-situ wh-element but not distinguished by raised wh-element. Note that we use the abbreviations from the table above, e.g.

⁵ Age: mean age 29.1, range 19–52; sex: 16 females, 22 males; occupations: all but three students or graduates; dialect background: 16 from the southern areas of Bavaria or Baden-Württemberg, 5 from the central areas, 5 from the north, 12 claimed no dialect background.

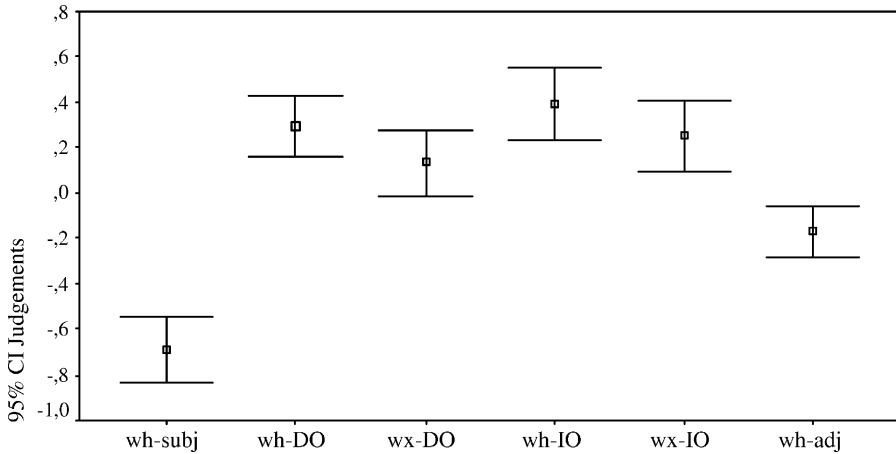


Fig. 1. Results of superiority experiment by in-situ wh-item.

“bare wh-direct object” appears as *wh-DO*, and “discourse-linked direct object” is abbreviated as *wx-DO* (*wx* indicates “which X”).

For full detail and extensive discussion of these results the reader is referred to Featherston (forthcoming); we do not intend to repeat this detail here, but merely summarize the main conclusion, which we wish to develop further. Visual inspection of the graph reveals that structures with in-situ subjects are scored much lower than structures with any other in-situ wh-item. There is also some variation among the other conditions, both by in-situ wh-item and by raised wh-item, but these are much weaker than the dispreference for in-situ wh-subjects and never cancel it. In a repeated measures anova, the effect of the in-situ wh-item type is highly significant both by subjects ($F(5) = 30.73$, $p < 0.001$) and by items ($F(5) = 45.86$, $p < 0.001$). In pairwise Tukey HSD tests, the in-situ subjects (all $p < 0.001$) and the in-situ adjuncts (all $p < 0.03$) were shown to differ from all other conditions, while these others did not differ from each other (all $p > 0.1$). There is no relevant interaction with the raised wh-item type.⁶

Since a dispreference for in-situ subjects in multiple wh-questions is the core empirical content of the superiority constraint, we argue in Featherston (forthcoming) that German too has a superiority effect. It is plainly the case that in a German multiple wh-question there is a constraint on which wh-item is raised to initial position. Note that the full set of data is very conclusive: all the five structures with an in-situ wh-subject are judged worse than all of the other 21 multiple wh-question structures. Nor should the relative weakness of the violation cost be regarded as

⁶ None of the data discussed in this paper was log-transformed since its distribution neither motivated nor required it. Tests of normality of distribution (Kolmogorov–Smirnov procedure with the Lilliefors correction) and homogeneity of variance (Levene test) were carried out prior to analysis of variance. Only one condition of one analysis of one data set failed this testing: the in-situ subjects condition of the by items analysis of the superiority experiment data are not fully normally distributed. In the light of the robustness of the effects observed, this cannot undermine the basic empirical facts from which we argue.

evidence that the effect we find here has a different syntactic nature to superiority in English: this pattern can be observed with other structures too. For example, we demonstrate (Featherston, forthcoming) that this is true of the *that*-trace effect too. Normally held to be absent from German, we show it is unambiguously present, but with a weaker violation cost than in German. The use of the magnitude estimation procedure for data collection can thus shed light on an outstanding problem of German syntax.

5. Discussion

Since, as we noted, the standard view in German linguistics has been that superiority is not operative in German, our result raises the question why competent and informed linguists should have reached such a conclusion. The first step towards the answer is fairly clear: linguists denying superiority are using different criteria to judge the existence of a constraint. The general assumption appears to be that a grammatical constraint such as superiority must exclude any structure violating it by rendering it fully ungrammatical. Since there are examples of legitimate structures of German which violate superiority, there can therefore be no superiority constraint in German (Grewendorf, 1988; Müller, 1991; Haider, 1993; Lutz, 1996; Fanselow, 2001). Even authors who accept that superiority is not wholly absent from German accept this counterexample-driven model of argumentation: if one can identify an acceptable sentence which violates superiority, then superiority cannot apply to that configuration (Wiltschko, 1997). This approach presupposes an essentially binary model of grammaticality and violation cost.

This conception of the nature of a grammatical constraint differs from that which is more or less forced upon us by the nature of the data from this study. There is no evidence of a binary division between grammatical and ungrammatical nor any sign that the violation of constraints carries a single violation cost. While Fig. 1 above shows that structures with in-situ *wh*-subjects are systematically judged worse than all the others, this should not be interpreted as showing that just these structures are to be regarded as ungrammatical and all the others as grammatical – merely that they are a bit worse. In fact all of these structures are judged as grammatical in the literature; only the perspective of grammaticality as a continuum permits us to see them as nevertheless different.

Now this has implications for our ideas about the nature of (un)grammaticality and the cost of constraint violations. It seems fairly clear that for this syntactic phenomenon, a constraint may exist which is syntactically active and bears a measurable violation cost, but whose violation cost is insufficient to render a sentence bad enough to make it perceived to be ungrammatical. Linguists who deny superiority in German are implicitly ruling this possibility out – Haider (1993:159) does so explicitly. But our magnitude estimation data indicates otherwise – the results are sufficiently clear-cut to allow us some certainty. So it seems that the reason that our own findings and the claims of many previous authors diverge is exactly to do with the issue of the abstraction of the scale of grammaticality. It appears that this abstraction to a binary system has hindered rather than helped linguists, at least in this descriptive field.

6. Experiment on superiority effects in topicalization structures

What therefore is the nature and cause of this difference between English and German? Why should a constraint which exists in both languages produce an ungrammatical output if violated in the one language but merely a dispreferred output in the other? In order to find out in more detail we conducted a further experiment, testing a wider range of structures to see if we would find the subject–object asymmetry which is the core descriptive fact of superiority manifested in them too. Our own expectation (based on previous unpublished studies on different structures with in-situ wh-items) was that we would find a similar dispreference for in-situ subjects in multiple wh-questions and in structures with only just one embedded wh-item, but that this would not be replicated in simple German declarative topicalization structures. If confirmed, this would provide evidence that the motor of the superiority effect is the embedded wh-item and would tend to confirm the syntactic similarity between topicalization and wh-movement in the sense of Chomsky (1977).

We retested the multiple wh-questions, and added standard declaratives (which in German involve a topicalized initial constituent), and declaratives with in-situ wh-items. We illustrate these structures in (5), based upon the sentence *Der Lehrer hat den Schüler ausgeschimpft* (“the teacher has told off the pupil”). Sentence types (5)a and (5)b are the classic cases where we might expect to find a superiority effect, if there is one; (5)c and (5)d have the same in-situ wh-items but no raised wh-item, only a topicalized constituent; (5)e and (5)f are a pair of control conditions designed to quantify the background word order preference.

- (5) a. whS whDO
 Wer hat wen ausgeschimpft?
 Who has whom told.off
- b. whDO whS
 Wen hat wer ausgeschimpft?
 Whom has who told.off
- c. topS whDO
 Der Lehrer hat wen ausgeschimpft?
 The teacher.NOM has whom told.off
- d. topDO whS
 Den Schüler hat wer ausgeschimpft?
 The pupil.ACC has who told.off
- e. topS DO
 Der Lehrer hat den Schüler ausgeschimpft.
 The teacher.NOM has the pupil told.off
- f. topDO S
 Den Schüler hat der Lehrer ausgeschimpft
 The pupil.ACC has the teacher told.off

This experiment was carried out using the same procedures as detailed in the superiority experiment above. Twelve sets of sentences of the form of (5) were constructed, and each subject saw two lexical variants of each syntactic condition. The structures and lexis were

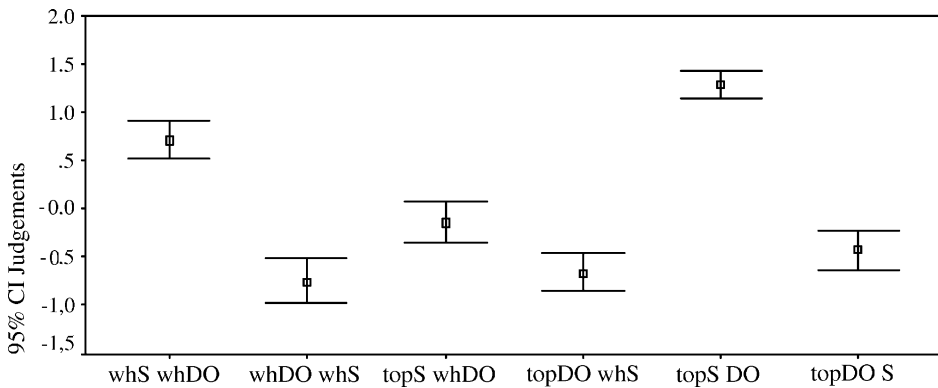


Fig. 2. Results of experiment testing the subject–object asymmetry in multiple wh-questions, declaratives with embedded wh-items, and declarative topicalization structures.

strictly controlled: all subjects and direct objects were animate. NPs were matched for length and lemma frequency from the CELEX lexical database (MPI Nijmegen).⁷ Twenty-one subjects took part in this experiment but one subject’s data was discarded due to doubts about its quality.⁸

7. Results

Fig. 2 illustrates the results of this experiment. As before the vertical axis shows mean judgements, with higher scores indicating “more natural”, and the horizontal axis distinguishes the six different conditions tested. The error bars show the 95% confidence intervals for the means by conditions. Starting from the left, the first pair of error bars show the superiority effect as before. The leftmost *whS whDO* condition corresponds to sentence type (5)a in which the wh-subject is raised and the wh-object in situ, while *whDO whS* has these two wh-items reversed as in sentence type (5)b. The difference between the two is clear: our finding from the previous experiment is replicated. The next pair correspond to sentence types (5)c and (5)d, that is, declarative topicalization structures with in-situ wh-items. Here too there is a clear difference between *topS whDO* with an in-situ wh-object and *topDO whS* with an in-situ subject: the condition with the in-situ wh-subject is judged clearly worse (Fig. 2).

Let us now consider the third pair of error bars. These represent standard declarative topicalization structures with no wh-items. Here too there is a clear difference between the

⁷ Length range: 6–9 letters, mean lengths: subjects 7.5, direct objects 7.2; lemma frequency (mean per million figures) subjects 28.6, direct objects 34.5. The full list of the materials can be inspected at <http://www.sfb441.uni-tuebingen.de/~sam/db/topsup.maz.html>.

⁸ Participants were instructed to supply their names, ages (mean age 25.8, range 19–47), sex (6 females, 13 males), occupations (all except one students or employees of the university) and dialect backgrounds (8 from the southern areas of Bavaria or Baden–Württemberg, 2 from other dialect areas, 9 claimed no dialect or responded “deutsch”).

SVO word order illustrated in (5)e and the OVS order as in (5)f; in fact the grammaticality contrast is as large here as in the pair of conditions illustrating the classic superiority effect (5)a and (5)b. Before we discuss the implications of this result, let us note that none of these structures would be standardly regarded as fully ungrammatical in German, although the *topS whDO* and *topDO whS* conditions are only feasible as echo questions. We may summarize this result by saying that in each pair of structures, the one with the in-situ subject is judged worse than the one with the in-situ object, and that this contrast is very similar in the multiple wh-question pair and the topicalization pair, but smaller in the single in-situ wh-item pair.

8. Discussion

The aim of this experiment was to explore whether an in-situ wh-subject in an otherwise standard declarative clause would trigger similar degradation in judgements to an in-situ subject in a multiple wh-question. If the subject–object contrast in the multiple wh-question pair (i.e. the superiority effect) had been replicated in the single in-situ wh-item pair, but not in the topicalization control condition pair, there would have been reason to pursue the hypothesis further.⁹ Neither of these conditions was fulfilled, in the event, and so the data does not support this hypothesis. However, a different parallelism is visible: the behaviour of the superiority pair of structures and the topicalization pair of control conditions is remarkably similar. To quantify this similarity we carried out an analysis of variance on just these four conditions. This reveals a significant effect for Structure (multiple wh-question versus declarative topicalization) both by subjects ($F(1) = 16.92, p = 0.001$) and by items ($F(1) = 20.50, p = 0.001$), as well as a highly significant main effect for Argument Order (Subject > Object, Object > Subject) ($F(1) = 377.64, p < 0.001$) and ($F(1) = 241.36, p < 0.001$), but no interaction of the two ($F(1,1) = 1.02, p = 0.325$) and ($F(1,1) = 0.017, p = 0.898$). This confirms the similarity of these effects.

This close correspondence between the superiority conditions and the topicalization conditions provides the basis for a different account of the cause of the superiority effect. It would appear possible that the superiority effect, far from being a free-standing filter on structures as its formulation as the Superiority Condition would imply, is instead simply one symptom of a more general effect which constrains the ordering and relative position of grammatical functions. We have regularly observed its effects across a range of structures in our own studies, and it is well attested in the literature. For example, the extensive work on word order preferences of Pechmann et al. (1994) showed robust evidence from several methodologies of a preference for subjects to linearly precede indirect objects, and indirect objects to precede direct objects, although this order is not a precondition for a grammatical structure in German. Our finding that non-wh-constituents in German show a parallel dispreference for in-situ subjects offers a clue how we might account for the difference in visibility of the superiority effect in German and English.

⁹ Let us note here that this was a pilot experiment. An echo question is a particular use of a structure, and this therefore necessarily very context-dependent. It goes without saying that the context would have to have been manipulated as a variable in a full-size study on echo questions.

It has often been noted that German word order is relatively free, indeed Hale (1982) went so far as to argue that German was non-configurational. Now this claim has been effectively laid to rest, but it is nevertheless true that variation of argument and adjunct location in the German clause is fairly readily possible, both in order to achieve particular focus effects, but also often without any effect on the semantic content of the sentence at all – (6)ab.

- (6) a. weil niemand sich dabei wohl fühlt
 since nobody REFL about.it well feels
 “since nobody feels happy about the situation”
 b. weil sich niemand dabei wohl fühlt

Lenerz (1977) and Uszkoreit (1987) provide extensive discussion of the factors which constrain relative position of grammatical functions, work which Keller (2000) has shown to be empirically well supported in its predictions.

Now in English too constituents can be located elsewhere than in their canonical positions for a similar variety of effects, as a glance at any book of poetry will show, but as can be seen in standard spoken language too (Radford, 1997: chapter 10). Semantically empty variation is also possible in English – (7).

- (7) a. Nobody told off the teacher.
 b. Nobody told the teacher off.

But it is also clear that the circumstances under which this can felicitously be done in English are more restricted than in German: put differently, the factors constraining word order in English have higher violation costs than in German. The same applies to what we shall follow Quirk et al. (1985:1377ff) in referring to as *inversion* (i.e. structures with a non-subject in the sentence-initial canonical subject position, and the subject in a post-verbal position). This inversion of object and subject is readily feasible in German, but far more limited in English – (8).¹⁰

- (8) a. Auf dem Tisch stand ein Blumentopf.
 a'. On the table stood a flower pot.
 b. “Komm, schnell!” rief Anna.
 b'. “Come quick!” cried Anna.
 c. Ein großer König bin ich.
 c'. A great king am I. (*Yertle the Turtle* Dr Seuss, 1950, ISBN 0394800877)
 d. Neunzehn Pasteten aß der erste Kandidat, dreiundzwanzig der zweite and ganze fünfundzwanzig der glückliche Sieger.
 d'. MNineteen pies ate the first contestant, twenty-three the second, and no less than twenty-five the lucky winner.

¹⁰ One might argue that *fronting* (i.e. XP SUBJ V ...) (e.g. *Really good meals they serve at that hotel*) or *negative preposing* (i.e. XP_(NEG) AUX SUBJ V ...) (e.g. *Never again would he glimpse the shores of his homeland*) should be regarded as the English equivalent of the German inversion structure, but these too are far more tightly constrained than the equivalent German construction. The force of the argument is therefore unaffected.

While German allows such structures, the English equivalents must be carefully selected; presentational PPs allow it, direct speech as an object does too, copular verbs permit it fairly easily, but canonical transitive verbs and direct objects permit it only marginally – (8)d'. Generalizing this to the location of *wh*-items, we see that this closely parallels the contrast in superiority effects between German and English too. Our hypothesis therefore is that the superiority effect is not specific to *wh*-items but merely a sub-part of the more general constraint on the surface realization of word order in the clause which regulates inversion, and that this has a higher violation cost in English than in German. Since inversion is strictly constrained in English, a *wh*-object raised over a *wh*-subject triggers a constraint violation sufficiently strong for the sentence to be regarded as ungrammatical on a two-valued scale of grammaticality. The equivalent structure in German triggers the same constraint violation, but the cost of this violation is less, which means that the structure is not sufficiently ungrammatical for it to be classed among the group of fully ungrammaticals on a binary scale.

Let us sum up. We demonstrated in Featherston (forthcoming) that German does indeed have a superiority effect, in that multiple *wh*-questions with an in-situ *wh*-subjects are judged markedly worse than all other types. In a further experiment, we discovered that a remarkably similar effect can also be observed in otherwise identical sentences without any *wh*-items. On the basis of this data, we hypothesize that the superiority effect is merely one particular sub-case of a more general constraint on inversion structures. Three observations speak in favour of this. First, the difference in violation cost of inversion constraints in English and German is closely paralleled by the difference in superiority constraint violation costs. Second, the costs of subject–object inversion is remarkably consistent with *wh*-items and non-*wh*-items. Third, this account is theoretically economical.

A note of caution is in order at this point. We are speculating on the basis of one experimental result; more needs to be done to test this. Moreover, this result merely shows that the structures are introspectively judged alike; it provides no direct evidence about the motor of the effects seen. There are data problems too: the multiplicity of English inversion structures (OVS, OSV, $O_{NEG}V_{AUX}SV$) makes it difficult to determine which is the precise structure for comparison with German OVS.¹¹ Nevertheless, the suggestion is much to recommend it: it has the advantage of requiring no additional mechanisms within the grammar, but relying entirely on well-known descriptive facts about German (cf. Behaghel, 1932: 50ff; Paul, 1919: 50, see also references in Lee, 1979) and English (Quirk et al., 1985: 1379ff). It is theory-neutral, and thus depends on no specific assumptions about the nature of the grammar. The cross-linguistic, perhaps universal existence of this linguistic constraint, as well as its apparent variation in violation cost are testified to in Greenberg's universal number 1 (Greenberg, 1966)¹² and are no doubt related to the cross-linguistic variation in the importance of word order found by MacWhinney (1989). In fact this effect is not well understood, but if and when it can be captured in syntactic terms, the superiority

¹¹ It is perhaps this difficulty in identifying one inversion structure in English which is at the origin of the idea that a separate constraint should be required to account for the superiority effect. It seems unlikely that this would have occurred to a linguist working in a language where the parallelism of +*wh* and –*wh* inversion is more transparent.

¹² Thanks to an anonymous reviewer who pointed this out to us.

effect, we suggest, will be captured with it. Let us note too that this hypothesis is eminently falsifiable; there is a clear prediction that the strength of the superiority effect will covary with the strength of the dispreference for inversion, both cross-linguistically and intra-linguistically across structures, all other things being equal. If this were to turn out to be false, and no interfering factor can be identified, the hypothesis would be falsified. For example, in our superiority experiment we noted no contrast of direct and indirect objects. That is to say, subjects found structures of the form *Wem. . . was* (“to whom . . . what”) just as good as *Was. . . wem* (“what . . . to whom”). It is thus predicted that the same balance of preference will apply in topicalization inversion. This can readily be tested. Similarly, both were equally bad raised over an in-situ wh-subject. This too can be tested on topicalization inversion, as can equivalent structures in English. This is work in progress.

This account, if confirmed, has rich implications for grammars. In particular, they can be made more economical by the removal of whatever mechanism is in place to account for superiority, unless it is required for other purposes. This is particularly pertinent to superiority since it is an island constraint, and island effects are theoretically interesting because they seem to be exceptions to what applies more generally in linguistic structure, which has caused them to play a more central role in theory building than their frequency might be held to motivate. If superiority can be derived from more general constraints on the form of clauses, the explanatory problem can be removed from the descriptive burden of syntactic theory. Let us note too that all of the structures we examined here would be standardly judged to be grammatical within the binary model: the differentiation between them would be effectively invisible. But before discussing this further, we report one more experimental study which we consider to have interesting implications for the analysis of superiority and to support this interpretation of our findings.

9. Discourse linking experiment

A special case of superiority is discourse linking. Pesetsky (1987) notes that there are cases in which the superiority effect does not appear, even though the syntactic conditions would predict that it should. When an in-situ subject is *discourse linked* (or *d-linked*), i.e. it clearly refers to a referent already within the universe of discourse, then this in-situ subject does not trigger a superiority effect. The standard method of grammatically marking d-linking is to make the wh-item a *which-X* form rather than a bare wh-pro-form (e.g. *what*). Thus, unlike the standard superiority examples in (2)a and (2)b above, (9)a and (9)b show no very apparent grammaticality contrast.

- (9) a. Mary asked which man read which book.
 b. Mary asked which book which man read.

Pesetsky claims that discourse linked items do not need to move at LF for interpretation, as they are “unselectively bound” by a Q morpheme. Let us note here that Pesetsky’s own examples (9)ab have both raised and in-situ wh-items in d-linked form, while his account requires only the d-linking of the in-situ item. It thus remains unclear to what extent he intends the d-linking of the raised wh-item to play a role.

It will be clear that linguists denying the existence of superiority in German will not recognize the existence of Pesetsky's d-linking effect in German, since this latter is an exception to the former. We illustrate the constructions in which d-linking might apply in German in (10). If superiority applies, then (10)b should be worse than (10)a, but if d-linking applies too, then (10)c should be better than (10)b.

- (10) a. Wer hat was gelesen?
 who has what read
 b. Was hat wer gelesen?
 what has who read
 c. Welches Buch hat welcher Mann gelesen?
 which book has which man read

In the light of our finding of a superiority effect in German, in this experiment, we aimed to test whether German would pattern like English in this special case of superiority. This experiment was a follow-up experiment from the superiority experiment reported above and used a subset of its materials, testing only for d-linking effects with wh-subjects and wh-objects, where we previously found a superiority effect. We examined the different types of multiple wh-questions obtained by manipulating the variables of argument order (subject > object, object > subject) and discourse linking of in-situ wh-item (bare wh-item, *which-X* wh-phrase). Note that we also examined the effect of discourse linking of raised wh-items (bare wh-item, *which-X* wh-phrase), but we shall simplify our exposition here to the four conditions most relevant to the current discussion (for full details see Featherston, forthcoming). We exemplify the experimental structures in (11).

- (11) a. wh-subj wh-DO
 Wer hat dem Patienten WAS empfohlen?
 who has to.the patient what recommended
 b. wh-subj wx-DO
 Wer hat dem Patienten welche Zahnpasta empfohlen?
 who has to.the patient which toothpaste recommended
 c. wh-DO wh-subj
 Was hat WER dem Patienten empfohlen?
 what has who to.the patient recommended
 d. wh-DO wx-subj
 Was hat welcher Zahnarzt dem Patienten empfohlen?
 what has which dentist to.the patient recommended

The methodology adopted was as in the previous experiment. Thirty subjects took part (mean age 26.1, range 20–42; 20 males, 10 females; 15 from the south, 2 from the centre, 5 from the north, 8 claim no dialect influence). Eight lexical variants were used. Unlike in the superiority experiment, bare in-situ wh-items were capitalised (*WER, WAS*). This written form was intended to represent the word stress necessary on such items in the spoken language, and remind subjects that these should be interpreted as wh-items, not as weak pronominals (*was = etwas* “something”). In the instruction phase, subjects were informed

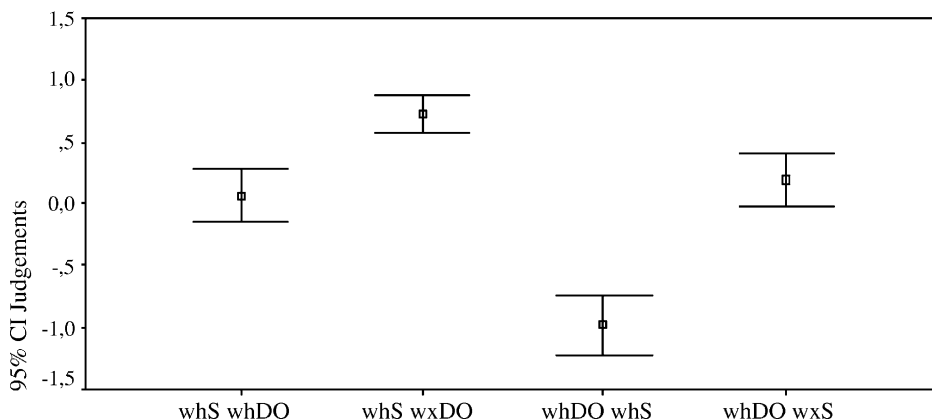


Fig. 3. Results of experiment on discourse linking in German.

about this practice and its aim. The results of this experiment are illustrated in Fig. 3. We use the same abbreviation system for conditions here as above: so for example *wh-subj wxDO* indicates a raised bare *wh*-subject over an in-situ *which-X* direct object.

Fig. 3 illustrates the results of this experiment using the same conventions as the previous error bar result graphs. The first thing to check for is a replication of the superiority effect we found in the first experiment. This is clearly present: for example *wh-subj wh-DO* condition (corresponding to (11)a above) is judged clearly better than the *wh-DO wh-subj* – (11)c. The precondition for testing for a discourse linking effect is thus fulfilled. And Pesetsky's discourse linking effect is indeed visible: the *wh-DO wh-subj* condition with the in-situ bare *wh*-subject scores clearly worst, which we may attribute to superiority, but the *wh-DO wx-subj* condition with a discourse linked *which-X*-type in-situ subject is much better, in spite of the superiority violation: it is in fact as good as the *wh-subj wh-DO* condition which does not violate superiority. These effects are clear in the statistical analysis too. This reveals a significant effect for Argument order (Subject > Object, Object > Subject) both by subjects ($F(1) = 46.2, p < 0.001$) and by items ($F(1) = 68.9, p < 0.001$), as well as for in-situ *wh*-item type (bare *wh*-item, d-linked *wh*-item) ($F(1) = 81.8, p < 0.001$) and ($F(1) = 60.1, p < 0.001$). Thus far the results confirm that German not only knows the superiority effect just as English does, but exhibits the same puzzling exception to it which Pesetsky discusses. We discuss this finding and its implications for the universality of *wh*-constraints in Featherston (forthcoming) but here we should like to propose an alternative explanation of the same data set.

Let us consider the four conditions in Fig. 3 again. Superiority accounts for the fact that the *wh-subj wh-DO* condition (11)a is better than the *wh-DO wh-subj* (11)c. The discourse linking phenomenon accounts for the fact that the *wh-DO wx-subj* (11)d is better than the *wh-DO wh-subj* (11)c. These two effects make up the standard view of what is going on in this set of structures. What is however not accounted for is the behaviour of the *wh-subj wx-DO* condition (11)b. That this is among the more grammatical group of structures is no surprise, but in fact it scores higher than the others. Now on the standard binary view of grammaticality this fact is invisible – the *wh-subj wx-DO* is simply grammatical in the

same way that the *wh-subj wh-DO* and *wh-DO wx-subj* conditions are grammatical. Magnitude estimation reveals more detail however; we see that the *wh-subj wx-DO* condition is actually better than these other two. This is confirmed by the statistical analysis. In pairwise Tukey HSD tests, the *wh-subj wx-DO* condition was significantly different from all other conditions (all $p < 0.001$). In fact the systematicity is apparent: the effect of the ordering of the arguments is independent of d-linking status, and the effect of d-linking is independent of the order of arguments. It looks very much as if there is a superiority effect as discussed above, but as if the discourse linking effect and the interaction of the two effects are rather different from what is generally thought. The discourse linking effect is not merely a cancellation of a superiority effect, rather it is a constraint on the form of an in-situ wh-item independent of superiority: an in-situ wh-item not in d-linked form incurs a violation cost. As such we might term it the *in-situ wh-item constraint* – (12). Note that this makes reference to the concept of discourse linking, but not to superiority, from which it is completely independent.

- (12) *In-situ wh-item constraint*
 In-situ wh-items must be discourse-linked

Let us note too that this constraint interacts cumulatively with superiority; as Keller (2000) has convincingly shown, this interaction type is common among grammatical constraints. By this we mean that the *wh-DO wx-subj*, which violates superiority, is indeed worse than the *wh-subj wx-DO*, even though both are fully acceptable structures of German. Since indeed all of these structure types are judged grammatical in German, from the perspective of a linguist applying the binary model of grammaticality there is nothing to explain here: since German superiority has an insufficiently strong violation cost to bring about ungrammaticality even in combination with an in-situ wh-item constraint violation, both superiority and the effect of discourse linking are invisible. In English on the other hand, the *wh-DO wh-subj* structure is ungrammatical on the binary scale, being penalized both by its superiority violation and its *in-situ wh-item constraint* violation, the cumulative effect of these being sufficient to push it into the group of the ungrammaticals. But we have seen that pattern of data in German in fact corresponds quite closely to the acknowledged facts of English. Here too the key to the perceived difference between English and German seems to reduce to the strength of violation cost of superiority violations in the two languages.

What therefore is the nature of the discourse linking of in-situ wh-items effect? We have redefined its sphere of application, but so far offered no account of what might cause it. Here we are in the realm of speculation, but the discovery that it has no necessary connection with superiority does provide a useful clue. We have noted that the feature of discourse linking does indeed seem to be the trigger of the effect, and it may well be indeed this which is at its root. Structures with in-situ wh-items perform mostly, if not always, the function of echo questions (as Sobin, 1990, notes). Multiple wh-questions which are not strictly echo questions are possible, but the most readily available interpretation of them is as echo questions. It is a defining feature of echo questions that they are linked into the discourse: without a preceding model in the discourse no echo question can exist. We will scarcely be surprised then to find that in an utterance type which is necessarily discourse-

linked, the key element of the sentence is also most natural in discourse-linked form. We may therefore attribute the preference for discourse linked wh-items in-situ to the echo question function of the containing sentence. This is likely to apply even to multiple wh-questions which are not strictly echo questions, since they too are strongly linked to the discourse environment. It is infelicitous to ask *Who bought what?* outside a discourse context in which someone bought something. This possibility of course relegates the discourse linking effect from being a grammatical constraint to being a mere functional preference, and as such it allows no interesting predictions and does not enrich the syntax in any way, but it accounts for the discourse-linking phenomenon in a way which is both economical and testable.

10. Magnitude estimation data: assumptions and implications

Let us pull together the assumptions that we are making and the conclusions which they enable us to draw on the basis of our experimentally obtained data. In fact, it is necessary to make considerable alterations to the standard assumptions about grammaticality, but we believe these to be justified by the data. But even before this we must first accept that the finer differences between structures, which are revealed by magnitude estimation but which have not previously been noted, are grammatical in nature. This point is often doubted by linguists who are loath to see changes to the status quo of the data base underlying work in syntax; they often attribute the differences found to factors irrelevant to grammar such as “acceptability”, “pragmatic felicity” and “markedness”. But we regard this critique as unfounded. We do not claim that these factors are not playing a role in the judgements we measure, on the contrary we readily admit that they are included. But this is an argument against the use of any introspective judgements in syntax, not against those gathered experimentally. All such judgements are sensitive to plausibility, frequency, lexical preference and so on (e.g. Schütze, 1996). There is however no reason to suspect that these factors play a greater role in experimentally obtained data than in standard linguists’ grammaticality judgements, on the contrary, there are strong reasons for suspecting experimentally obtained data is less affected by them. In our data, the variables of context and lexis are scrupulously controlled for: all structures to be tested are presented in identical contexts and in multiple lexical forms, and only effects which are consistent across lexical variants and speakers will appear in the results. Our own results are thus less, not more, vulnerable to these reproaches than is the standard linguist’s grammaticality judgement as generally used in work on syntax. Our data is as free of irrelevant effects as can be practically achieved within our current understanding.

It is true that we have not presented a clear theoretical basis for distinguishing between grammatical and extra-grammatical factors, but the studies discussed here are in part a contribution towards developing such a distinction with some empirical grounding (cf. Chomsky, 1965: 11). Within the binary model of grammaticality, the distinction has usually been implicitly drawn by implementing a cut-off point. If the cost of a violation is sufficiently severe to render any violating structure ill-formed, then the corresponding constraint is attributed to the “narrow” grammar. If the violation cost is less severe, so that violating examples may yet squeeze their way into the class of grammaticals, then the

constraint is thought of as not narrowly grammatical. Now as an approach to the problem of syntactic description, this is not misguided: essentially it implements the prioritized consideration of stronger effects. But the availability of better data must motivate its reconsideration, a process which we attempt to develop here with our own data-driven approach.

This is made up of two steps: when analyzing a set of results we first discount from being grammatical any effect for which we can identify a non-grammatical explanation in terms of plausibility, lexical preference, ease of processing or any other performance-type factor. We then regard the remainder of the variation attested as grammatical in nature, even if this results in the adoption of grammatical constraints whose violation cost is not alone sufficient to exclude a violating structure absolutely. This remainder is that which the grammar must account for. The effect of this may be to reduce the proportion of variation attributed to the grammar: in our studies of binding we demonstrate that some variation which is generally assumed to be strictly related to binding, and thus narrowly grammatical, was in fact due to a range of surface factors interacting cumulatively. The patterns of binding remaining after the discounting of non-grammatical factors provided a much more transparent picture of the nature of anaphoric binding, and one which generally confirmed the standard generative approach in terms of Binding Conditions (Featherston, 2002; Sternefeld and Featherston, 2002).

Acceptance of the validity of the data makes acceptance of our syntactic assumptions much easier, since they are more or less forced by the data. The first assumption is that grammaticality should be regarded as a true continuum and not just as a binary opposition of grammatical and ungrammatical, or even a continuum with fixed end points. This is counter to standard practice in syntax, but we are prepared to defend it stoutly, for two reasons. The first is that linguists have always known that the binary model was an abstraction and that the real pattern of grammaticality was more complex (Chomsky, 1963; Bard et al., 1996; Schütze, 1996; Keller, 2000, and references therein). One may sometimes see as many as seven different values identified in the literature, a practice which tacitly admits that the binary model is no more than a convenient abstraction.¹³ Let us note here that we do not doubt the validity or usefulness of abstracting from the empirically observed pattern of grammaticality judgements, in fact we consider it a useful heuristic practice which has served the field well, but the fact that it is possible to make advances in linguistics using a grammaticality model simplified in this way does not mean that this simplification is necessary or obligatory, or that no relevant data is being obscured.

The second reason for our assumption that grammaticality can be treated as a continuum is quite simply that this is what the data reveals. Given a free choice of judgement scale, informants choose to assign a wide range of degrees of grammaticality and do not distinguish a group of “good” structures from a group of “bad” structures. Our use of a grammaticality continuum is not an assumption required for a theory to go through, but a conclusion forced upon the observer by the data. Magnitude estimation data consistently displays a continuum of grammaticality, as do many other data types which use multiple

¹³ Müller (1995) has five: \emptyset [i.e. no mark = fully grammatical], ?, ??, *?, and *. Lakoff (1973) has six: \emptyset , ?, ??, ?*, * and **. Wurmbbrand (2001) uses seven: OK, #, %, ?, ??, *.

informants, including those which prescribe a binary choice. It would seem to be a reasonable procedure to accept this feature at least on a trial basis and explore what implications it has for the grammar.

A further assumption, which is a necessary corollary of our discussion above, is that grammatical constraints have a violation cost, but are violable. Let us state immediately that this *violability* is some way from the idea of violability in optimality theory (OT) (Prince and Smolensky, 1993). There the violability of constraints merely means that a given constraint in a given constraint ordering may fail to apply in one particular set of circumstances, namely if it can have no effect upon the selection of optimal candidate. If all the remaining candidate structures violate a certain constraint, then this constraint causes no candidate to be excluded, since it fails to distinguish between them. Similarly, if a constraint is so low ranked that only one candidate remains before it is applied, then too, it cannot affect the selection. The term *violability* in OT therefore merely represents the possibility that a violation does not necessarily exclude a candidate. It essentially means that a constraint can fail to have any effect on the output, it can thus *fail to apply*.

In our usage *violability* is more akin to *survivability*. When a structure violates a constraint it incurs a violation cost, but this violation cost in perceived grammaticality may not by itself be sufficient to exclude the structure as a possible expression of the language. In German, a sentence violating superiority is degraded in perceived grammaticality; our notion of violability reflects the fact that this single violation does not necessarily exclude the sentence from being a legitimate sentence of German. Note also that our conception of violability does not simply reduce to a failure to apply: every violation has an effect upon the output, but this effect is not necessarily fatal. In this sense, therefore, constraints are *survivable*. Keller (2000) has explored interactions of non-fatal violations in some detail, showing that their effect is cumulative.

These two assumptions accepted, our results have rich implications for the syntax. On the one hand, we are able to remove from the list of problems the unexplained absence of a superiority constraint from German, (see Featherston, in press), and that theory-neutrally. We can also offer an explanation of why so far no fully satisfactory account of the effect has been found, in spite of the considerable attention paid to the phenomenon: syntacticians have been looking in the wrong place. Since the assumption has been that the effect was specific to wh-items, was absent from some languages, and interacted with discourse linking, researchers have been looking for an account with these features. The data suggests that the effect does none of these, but is rather a specific case of a more general constraint on inversion.

We are thus able to advance novel accounts of superiority and discourse linking which are firmly based in our empirical results. This in itself is perhaps no giant leap for syntax, but it can be seen as merely one example of what will no doubt be a whole series of revisions to standard assumptions that the approach will make possible, indeed demand. In our own work on German we have found a number of problematic descriptive areas which, on closer inspection with the tool of magnitude estimation, have been resolvable, and we are not alone in this approach (Keller, 2000). The addition to the syntactician's toolkit of a powerful instrument which allows the testing of hypotheses must be regarded as welcome and long overdue. One only has to think of the difficulties that linguists have had in trying to find definitions of binding domains, the scope of the ECP, and movement bounding

domains, to see that the exclusive use of traditional data types and the standard assumptions about grammaticality (i.e. firstly, a binary grammaticality scale and secondly, an “instant death” constraint violation cost) are actually blocking progress in syntax.

Here then are some of the implications of our chosen data type, introspective grammaticality judgements elicited by use of magnitude estimation. The unsurpassed definition attainable with this approach makes it the methodology of choice in empirical approaches to syntax, since it offers sufficient detail to allow meaningful statistical analysis whilst still operating on informants’ introspective judgements of structures, rather than merely on frequency of occurrence or speed of processing. Since the construct measured is that which has been traditionally used in generative syntax, it is particularly well-suited to work in this paradigm. In the first instance this data allows the syntactician to test hypotheses which were previously untestable, since the detail offered by conventional introspective judgements was simply insufficiently sharp. In this paper, we noted that German does indeed have a superiority effect, thus effectively resolving a long-standing problem in German syntax. That the effect noted is indeed of the same nature as the superiority effect observed in English cannot be in doubt, since we also saw that the effect in German, just like its counterpart in English, contains an apparent exception: when the in-situ *wh*-item is discourse-linked the superiority effect is neutralized. Since both the main effect and the exception to it are additionally robustly statistically significant, the case is very firm indeed. This effectively demonstrates the ability of the methodology to provide firm answers to old questions. Additionally, data of this quality permits us to adjudicate between competing theoretical accounts of these phenomena. Lack of space has prevented us from discussing this aspect of the superiority results here, but we have shown elsewhere (Featherston, *in press*) that this data goes some way towards discriminating previously empirically indistinguishable generative accounts of superiority, such as the ECP and Shortest Move.

The possibilities of hypothesis evaluation which the approach permits are clear, and we are sure that they will be welcomed by all syntacticians who wish to advance syntactic knowledge by testing the predictions made by competing accounts of phenomena. But this approach to data gathering does not only permit existing accounts to be tested, it also permits rich new perspectives and good prospects of real descriptive advances to those linguists who embrace the fairly radical changes to assumptions about the nature of grammaticality. The additional descriptive detail offered by the increase in differentiation power in this approach readily provides the basis for new theoretical analyses. It enables linguists to develop new accounts therefore, and not just determine which old ones should be discarded. Provocatively, we have tried to illustrate here just how radical the reassessments of syntactic structure might be, by the device of staying close to the data and making all and only the assumptions about the nature of the grammar justified by the data, following the pattern of the data without presupposition. The aim has been to show the breadth of possibilities for new insights.

We do not deny that the results we report here could be interpreted differently, perhaps more in line with standard assumptions, but would claim that they would still make interesting new insights available. We hope that it will be clear also to syntacticians with other preferences, that the techniques we have used here are available to them for their own purposes.

Acknowledgements

This work was carried out within subproject A3 *Suboptimal Syntactic Structures* of the SFB 441 *Linguistic Data Structures*, funded by the Deutsche Forschungsgemeinschaft. Many thanks to project leader Wolfgang Sternefeld, but also to Frank Keller, Bob Borsley and two reviewers. The errors and failings remain my own.

References

- Baltin, M., Collins, C., 2001. *The Handbook of Contemporary Syntactic Theory*, Blackwell, Malden, MA/Oxford.
- Bader, M., 1996. On reanalysis: evidence from German. MS, University of Jena.
- Bard, E., Robertson, D., Sorace, A., 1996. Magnitude estimation of linguistic acceptability. *Language* 72 (1), 32–68.
- Behaghel, O., 1932. *Deutsche Syntax: Eine geschichtliche Darstellung, Band IV: Wortstellung, Periodenbau*, Carl Winters, Heidelberg.
- Chomsky, N., 1957. *Syntactic Structures*, Mouton, The Hague.
- Chomsky, N., 1964. Degrees of grammaticalness. In: Fodor, J.A., Katz, J. (Eds.), *The Structure of Language: Readings in the Philosophy of Language*, Prentice-Hall, Eaglewood Cliffs, NJ, pp. 384–389.
- Chomsky, N., 1965. *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.
- Chomsky, N., 1977. On Wh-movement. In: Culicover, P., Wasow, T., Akmajian, A. (Eds.), *Formal Syntax*, Academic Press, New York.
- Chomsky, N., 1981. *Lectures on Government and Binding: The Pisa Lectures*, Mouton de Gruyter, Berlin.
- Chomsky, N., 1993. A minimalist program for linguistic theory. In: Hale, K., Keyser, S. (Eds.), *The View from Building 20*, MIT Press, Cambridge, MA.
- Cowart, W., 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgements*, Sage, Thousand Oaks, CA.
- Crain, S., Fodor, J., 1987. Sentence matching and overgeneration. *Cognition* 26, 123–169.
- Fanselow, G., 2001. Features θ -roles and free constituent order. *Linguist. Inquiry* 32, 405–437.
- Featherston S. Universals and grammaticality: Wh-constraints in German and English. *Linguistics* 43(4), in press.
- Featherston, S., 2002. Coreferential objects in German: experimental evidence on reflexivity. *Linguistische Berichte* 192, 457–484.
- Frazier, L., 1985. Modularity and the representational hypothesis. In: *Proceedings of NELS 15*, 131–145.
- Greenberg, J., 1966. *Universals of Language*, second ed. MIT Press, Cambridge, MA.
- Grewendorf, G., 1988. *Aspekte der deutschen Syntax*, Narr, Tübingen.
- Haider, H., 1993. *Deutsche Syntax: Generativ*, Narr, Tübingen.
- Hale, K., 1982. Preliminary remarks on configurationality. In: Pustejovsky, J., Sells, P. (Eds.), 1982. *NELS*, vol. 12, pp. 86–96.
- Keller, F., 2000. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. Ph.D. thesis, University of Edinburgh.
- Keller, F., Corley, M., Corley, S., Konieczny, L., Todirascu, A., 1998. *WebExp: A Java Toolbox for Web-Based Psychological Experiments*. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh.
- Lakoff, G., 1973. Fuzzy grammar and the performance/competence terminology game. *Chicago Linguist. Soc.* 9, 271–291.
- Lasnik, H., Saito, M., 1984. On the nature of proper government. *Linguist. Inquiry* 15, 235–289.
- Lee, D.-H., 1979. *Aspekte der deutschen Syntax: Untersuchungen zur deutschen Syntax, mit besonderer Berücksichtigung der Wortstellung*, Tuduv Verlag, Munich.
- Lerner, J., 1977. *Zur Abfolge nominaler Satzglieder im Deutschen*, Narr, Tübingen.
- Levinson, S., 1991. Pragmatic reduction of the Binding Conditions revisited. *J. Linguist.* 27, 107–161.
- Lutz, U., 1996. Some notes on extraction theory. In: Lutz, U., Pafel, J. (Eds.), *On Extraction and Extraposition in German*. *Linguistik Aktuell* 11, Benjamins, Amsterdam.

- MacWhinney, B., 1989. *The Cross-Linguistic Study of Sentence Processing*, CUP, Cambridge.
- Manning, C., 2003. Probabilistic syntax. In: Bod, R., Hay, J., Jannedy, S. (Eds.), *Probabilistic Linguistics*, MIT Press, Cambridge, MA, pp. 289–341.
- Müller, G., 1991. Beschränkungen für W-in-situ. *Groninger Arbeiten zur Germanistischen Linguistik* 34, 106–154.
- Müller, G., 1995. A-bar syntax. A study in movement types, *Studies in Generative Grammar* 42, de Gruyter, Berlin, New York.
- Paul, H., 1919. *Deutsche Grammatik*. Band III, Teil IV, Niemeyer, Halle.
- Pechmann, T., Uszkoreit, H., Engelkamp, J., Zerbst, D., 1994. Word Order in the German Middle Field. *Claus Report No. 43*, University of Saarland.
- Pesetsky, D., 1987. Wh-in-situ: movement and unselective binding. In: Reuland, E., ter Meulen, A. (Eds.), *The Representation of (In)Definiteness*, MIT Press, Cambridge, MA, pp. 98–129.
- Pollard, C., Sag, I., 1994. *Head-Driven Phrase Structure Grammar*, University of Chicago Press, Chicago.
- Prince, A., Smolensky, P., 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report No. 2, Center for Cognitive Science, Rutgers University.
- Quirk, R., Greenbaum, S., Leech, G., Svartvik, J., 1985. *A Comprehensive Grammar of English*, Longman, London.
- Radford, A., 1997. *Syntactic Theory and the Structure of English*, CUP, Cambridge.
- Ross, J., 1972. The category squish: Endstation Hauptwort. *Chicago Linguist. Soc.* 8, 316–328.
- Reinhart, T., Reuland, E., 1993. Reflexivity Linguist. *Inquiry* 24 (4) 657–720.
- Schütze, C., 1996. *The Empirical Base of Linguistics: Grammaticality Judgements and Linguistic Methodology*, University of Chicago Press, Chicago.
- Sobin, N., 1990. On the syntax of English echo questions. *Lingua* 81, 141–167.
- Sternefeld, W., Featherston, S., 2002. Reciprocal reference with *einander* in German. In: Gunkel, L., Müller, G., Zifonun, G. (Eds.), 2002. *Arbeiten zur Reflexivierung*. *Linguistische Arbeiten*, vol. 481. Niemeyer, Tübingen, pp. 239–266.
- Stevens, S. (Ed.), 1975. *Psychophysics: Introduction to its Perceptual, Neural and Social Prospects*, Wiley, New York.
- Uszkoreit, H., 1987. Word Order and Constituent Structure in German. *CLSI Lecture Notes No. 8*, CSLI, Stanford.
- Wiltschko, M., 1997. D-linking, scrambling and superiority in German. *Groninger Arbeiten zur Germanischen Linguistik* 41, 107–142.
- Wurmbrand, S., 2001. *Infinitives: Restructuring and Clause Structure*, Mouton de Gruyter, Berlin/New York.