Universals and grammaticality: wh-constraints in German and English*

SAM FEATHERSTON

Abstract

In this article we investigate the movement constraints superiority and discourse linking. These have played an important role in the tradition of generative syntax and might therefore be expected to be universal, but they are usually argued to be absent from German. We looked for evidence of them in German data using the methodology of magnitude estimation of well-formedness, and compared this data with parallel results from English. The results showed these effects to be robustly active in the grammar of German, and revealed few differences between the two languages. We suggest that the reason why linguists have denied their existence in German is that they have been assuming a binary and categorical concept of grammaticality, forgetting that this is merely a simplifying abstraction from the primary linguistic data. We demonstrate that the admittedly convenient assumption of categorical grammaticality is obscuring our view of the syntax, and that studies using our own more empirically adequate assumptions of grammaticality can be productive. In particular, we hope that our conceptions of constraint survivability *and definition of* syntax relevance *may permit insights into the size of the grammar, crosslinguistic variation, and syntactic universals.*

1. Introduction

An important issue for linguists is the generalizability of the mechanisms they hypothesize: in particular the aspiration to universality lends linguistic analyses additional significance as evidence of certain aspects of the structure and functioning of the mind. It can therefore be disappointing for a linguist to find parts of what they consider to be the core grammar absent from certain languages. An example of this can be found in German, a language in which the existence of many constraints on wh-

Linguistics 43–4 (2005), 667–711
0024–3949/05/0043–0667

movement is frequently denied, since it is possible to find acceptable counterexamples in which these constraints appear not to hold. This is of theoretical significance, since movement and constraints upon movement are one of the chief components of the distinctiveness of generative approaches to syntax. This is also an important descriptive issue, since many authors have made key decisions about such questions as the structure of the German clause on the basis of data from movement constraints, since these are seen as providing key evidence about the position of the subject in the clause (e.g. Haider 1993). The concept of the parameter is one response to this: it attempts to account in universal terms for linguistic features that apply to some languages but not to others.

We applied the technique of magnitude estimation of grammaticality (Bard et al. 1996) to carefully matched and counterbalanced sets of materials. This experimental methodology reveals grammatical regularities much more finely than do conventional grammaticality judgements. In this article we report a series of experiments that test for superiority and discourse linking (in the sense of Pesetsky 1987) in German and compare them to equivalent results from English, where the existence of these constraints is not in question. Contrary to the general assumption, these constraints are demonstrated to be fully active and robustly measurable in German, and vary in only minor ways from their reflexes in English. We argue that this supports a model of grammar and grammaticality in which syntactic constraints have a violation cost, but are "survivable," that is, a violation does not automatically trigger absolute ungrammaticality, but nevertheless always affects the status of the violating structure. This contrasts with the concept of "violability," familiar from optimality theory (Prince and Smolensky 1993), where violated constraints simply fail to apply and have no effect upon the output. We finally discuss some implications of these findings for syntactic theory building and the extent of the phenomena space relevant to syntax.

2. Superiority

The term "superiority" seems to have originated in Chomsky (1973), where it is defined in terms which we might today think of as asymmetric m-command. The phenomenon of superiority is easily illustrated: in English, for example, (1a) is licit, but (1b) is not, except as a echo question in response to *Mary asked what zyxvft read.*

(1) a. Mary asked who read what.
b. *Mary asked what who read.

The basic generalization appears to be that there are (at least) subject–object asymmetries in extraction possibilities in multiple wh-questions: while a wh-subject can take the clause-initial position no matter what other in-situ wh-item there may be, the reverse is not the case. If non-subject wh-items are raised to clause-initial position leaving a subject wh-item in situ, the sentence is markedly degraded.

There is no consensus on the syntactic cause of this effect, as Ginzburg and Sag in their thoughtful discussion note (Ginzburg and Sag 2000). In this article we shall distinguish two schools of thought about the causal factor. The first type of account relates to the position of generation of a wh-item, and relates to the inability of certain wh-item types to be in-situ wh-items when any wh-item is superior to them. We might term this group of accounts "categorical superiority," since they would exclude certain wh-items from the in-situ position in a multiple wh-structure absolutely. The main representative of this approach is the empty category principle (ECP, e.g., Lasnik and Saito 1984), which explains the extractability asymmetry in terms of the proper government of the base position. Simplifying drastically, we may say that direct objects are properly governed because they are theta-governed, and can thus always be moved. Subjects and (on some accounts) adjuncts are not theta-governed and so they can only be extracted under circumstances that permit their trace positions to be antecedent-governed. In a multiple wh-question, Lasnik and Saito (1984) argue, the overtly raised wh-item prevents the other wh-item, covertly raised and adjoined outside it, from antecedent-governing its trace. It therefore follows that, while an object can remain in situ in the overt syntax and be raised only at LF, this is not possible of a subject or adjunct. The contrast between (1a) and (1b) follows from this distinction in proper government of subject and object positions.

The second group of accounts is related to economy conditions and might be termed "competitive," since the feasibility of any given wh-item in situ in a multiple wh-structure is dependent upon the wh-item type which is superior to it. This approach therefore defines the grammaticality of pairs of wh-items, which compete with each other for the clause-initial position, the closer winning. The motor of this type of account is an economy condition such as the minimal link condition (MLC, Chomsky 1993). This requires the assumption of an economy measure: essentially a distance metric within which movement from a subject position is shorter and thus also more economical than movement from, for example, an object position. It also requires that less economical derivations be blocked. Essentially the original "superiority condition" (Chomsky 1973) was of this type, though the effect was not specified as being driven by economy. Note that both of these types of accounts exist in various forms. Intermediate positions exist too: for example, the *Barriers* (Chomsky 1986a) version is based upon the ECP, but also contains a competitive element in the minimality condition, which prevents antecedent government by the further of two raised wh-items.

One aim of our study was to find evidence which might allow us to adjudicate between these two groups of accounts of the superiority effect, by examining the behavior of indirect objects and adjuncts as wh-items. The predictions of the categorical ECP-related approach for these are rather dependent on a range of other assumptions, such as the relationship between an indirect object and its subcategorizing verb, and the status of adjuncts. For example, in Lasnik and Saito (1984) superiority should not exclude adjunct extraction, because adjuncts are gamma-marked only at LF. However, we can at least derive the prediction that adjunct and indirect object extractions should be each be either uniformly legal or illegal, since the ECP is a condition on base positions taken individually. This contrasts with the rather stronger prediction of the competitive MLC account that the superior wh-item of any pair may be (overtly) raised while the inferior item may not. We should thus expect the data to form a hierarchy of wh-items on which, for any pair, the higher wh-item would be overtly extracted and the lower would be left in situ. Now the precise nature of this hierarchy is an empirical question, but we might expect to see a contrast of grammatical functions and an argument/adjunct distinction, perhaps as in (2).

(2) Subject > Direct Object > Indirect Object > Adjunct

On fairly standard assumptions about the structure of the clause and the concept of distance, this might be identical with an obliqueness hierarchy (Pollard and Sag 1994, but see also the similar Grewendorf 1988). Some relevant examples from Chomsky (1973) are in (3). He makes the comment about (3c) and (3d) that movement over a wh-phrase is permissable if it is located to the right of the verb, which would imply a superiority hierarchy consisting only of a contrast of subjects and all other items. It is likely that he was not making any distinction between bare wh-items and *which X* wh-phrases.

(3) a. John knows what books to give to whom.
b. John knows to whom to give what books.
c. John remembers where Bill bought which book.
d. John remembers to whom Bill gave which book.
e. *John knows what who saw.

We may summarize the difference in predictions thus: the categorical account type requires that a particular wh-item in situ in a multiple wh-question will either be possible or impossible; for example, indirect objects will or will not be acceptable in situ. The economy-related account type, on the other hand, suggests that grammaticality will be a function of pairs of wh-items. For a given pair, only one behavior is possible, but for the individual item, the feasibility of remaining in situ may vary: for example, an in-situ direct object might be acceptable with a raised subject, but unacceptable with a raised adjunct.

3. Discourse linking

Pesetsky (1987) notes that there are cases in which the superiority effect does not appear, even though the syntactic conditions would lead us to predict that it should. When an in-situ subject is "discourse linked" (or "d-linked"), that is, 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 *what* form. Thus unlike (1a) and (1b) above, (4a) and (4b) show no very apparent grammaticality contrast.

(4) 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 (4a) and (4b) 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 the d-linking of the raised wh-item plays a role. Another sub-aim of the experiments reported here was to clarify this point and test the effects of d-linking more generally. A further reason for the relevance of the factor of d-linking is its ability to function as a characteristic feature of the superiority effect. If a subject–object asymmetry shows this specific unpredicted exception to its application, we can be sure that the effect we are seeing is indeed related to superiority. This allows us to make cross-linguistic comparisons, secure in the knowledge that we are observing the same phenomenon in the two languages.

4. Movement constraints in German

The applicability of these constraints to German is controversial. There are certain differences between clause structures in English and German that make it credible that different constraints on movement apply. First, German complement clauses come in two types, which we shall refer to as "V-final" and "V2." There are, for our purposes, two important differences between them: the first type has a complementizer in initial position, while the second never has one, and the verb in the V-final type is clause-final, while the verb in the V2 type is at the beginning of the *Mittelfeld*, generally as second constituent. The contrast between a complement clause with and without a complementizer is thus in German part of a larger syntactically significant contrast; a complementizer cannot be optionally omitted in German. The second relevant difference to English is that the order of nontopicalized arguments (and adjuncts) is far freer in German, so that, while there are ordering preferences of subject and object (e.g. Lenerz 1997; Uszkoreit 1987), it is much less clear than in English that they occupy structurally distinct positions.

Examples (5a) and (5b) show the structures in which superiority should apply in German, if it exists. If (5a) is a possible expression of German and (5b) is not, then we may say superiority is part of German grammar.

(5) a. Wer hat was gelesen?
who has what read
b. Was hat wer gelesen?
what has who read

We may perhaps generalize that the standard view is that superiority does not apply (e.g. Grewendorf 1988; Müller 1991; Haider 1993; Lutz 1996; Fanselow 2001) or does so only in more restricted circumstances than in English. For example, Wiltschko (1997) claims that superiority fails to apply to d-linked wh-items in German (as in English), but that bare wh-items (like *wer* 'who,' *was* 'what') are more easily interpreted as d-linked in German than in English. Grewendorf (2001) accepts that there is "long" multiclausal superiority, but denies "short" monoclausal superiority.

It will be clear that anyone denying superiority in a language must also necessarily deny Pesetsky's d-linking effect, since this is an exception to the former. We illustrate the constructions in which d-linking might apply in (6). If superiority applies, then (6a) should be bad, but if d-linking also applies, then (6b) should be better.

(6) a. Was hat wer gelesen?
what has who read
b. Welches Buch hat welcher Mann gelesen?
which book has which man read

We test this contrast and the effects of d-linking in the raised and in-situ positions in our second experiment below.

As we have seen, researchers tend to deny that these movement constraints apply in German. Since the major underlying issue may be reduced to whether or not there are subject–object asymmetries in German, this denial has had far-reaching consequences for analyses of the structure of the German clause. In particular, the strong conclusion has sometimes been drawn that German shows no evidence of an IP, and that, therefore, the German clause has a less hierarchical structure (e.g. Uszkoreit 1987; Haider 1993; Pollard and Sag 1994; Pollard 1996), which itself has significant further implications. Such conclusions justify examination of their empirical base. The reason that researchers have denied these movement constraints is that counterexamples can be fairly readily found, that is, examples in which the constraints should apply, but which are nevertheless tolerably acceptable. Haider (1993) states explicitly what others tacitly assume:

> 'Wenn im Deutschen Subjektsätze die Spec-I Position einnehmen, verbietet CED Extraktion, und zwar ausnahmlos. Um dies zu widerlegen genügt aber schon ein einziges Beispiel. (If clausal subjects occupy the spec-IP position in German, then the condition on extraction domains forbids extraction, and that without exception. But only one single example is sufficient to refute this). (Haider 1993: 159 [our translation])

This position we may term the "counterexample model," and it is of impeccable intellectual heritage. However, it rests crucially on the assumption that well-formedness in the grammar is categorical, that is, that a structure which violates a constraint is necessarily completely excluded as a part of the language. But this assumption is, and has long been known to be, an abstraction from the primary data of syntactic well-formedness.

5. Grammaticality

For the discussion of this issue, we need some clarification of terms. We shall use the term "grammaticality" to refer to atheoretical syntactic well-formedness, the construct we attempt to measure with "grammaticality judgements." In this, we follow Schütze (1996), who notes that using the obvious alternative "acceptability" is equally theoretically laden. This is also our experience: the use of "grammaticality" for judgement outcomes invites the stock criticism that the judgements can never be demonstrated to be free of performance factors, while the use of "acceptability" triggers a dismissive response that one is measuring ‘‘mere’’ acceptability. The first of these two at least confronts the issue head on.

But the choice we have made is supported by other factors as well: first, this usage is largely in line with the most frequently quoted definition of what linguistic theory is concerned with. Let us look at the specification in Chomsky (1965: 3) once more. We first need an ‘‘ideal speaker–listener in a completely homogeneous speech community, who knows its language perfectly:’’ in the studies reported here, we use the mean values of samples of at least 25 informants. This is as near to access to an ideal informant as is achievable, given that it counterbalances individual variation and variation across the speech community.[1] Next, informants should be ‘‘unaffected by memory limitations, distractions, shifts of attention and interest, and errors (random and characteristic).’’ The effects of random and characteristic errors are excluded by averaging process over multiple informants, as above. The judgement task we employ is specifically designed to exclude as far as is possible memory and on-line processing effects: the sentences to be judged are displayed for as long as the informant chooses, and the multiple lexicalizations help us measure the effect of the form of the structure, rather than its content. Precisely the irrelevant factors identified by Chomsky are therefore controlled for as far as the constraints of the real world permit, as are any other effects, such as euphony, lexical frequency, and plausibility, which are grammatically irrelevant but affect judgements (for further discussion of the methodology, see below and references there). In sum, therefore, while the judgements we shall discuss are not fully in line with the abstraction of ‘‘grammaticalness,’’ Chomsky himself admits that this is unattainable (Chomsky 1965: 11, 19); our approach is as close as is currently practical (Schütze 1996; Cowart 1997).

Additionally, this use of the term ‘‘grammaticality’’ is in line with standard informal practice in the literature. For example, in the recent *Handbook of Contemporary Syntactic Theory* (Baltin and Collins 2001) 23 of the 30 authors use this term to denote the quality judged in example structures without further discussion or definition, and without any indication of how acceptability factors were controlled for (only one author uses ‘‘acceptability’’ in the same sense). We thus use the term ‘‘grammaticality’’ (similarly ‘‘grammatical, ungrammatical’’) here to refer to the quality we measure in judgements, while accepting that this is only an approximation to Chomsky’s ideal of ‘‘Grammaticality,’’ which we capitalize (likewise: ‘‘Grammatical, Ungrammatical’’). We suggest that this practice be more generally applied, noting that the distinction of a theory-specific and a more general term by means of capitalization has functioned very well in the case of ‘‘Case/case.’’

Our aim in these studies was to use experimental methods to learn about the grammar. To this end, we exclude as far as possible any confounding factor from our linguistic materials, take great care in their construction (for details, see below), apply the best methodology currently available for elicitation (magnitude estimation, Bard et al. 1996), and test multiple informants. In the analysis of our judgement results, we take a conservative position and attribute to the syntax only those effects which cannot be accounted for in any other way. We therefore discount anything for which an account in pragmatic, processing, and experimental terms can plausibly be advanced. The remainder is therefore as free of performance factors as can practically be attained. But this rigorous exclusion has another side to it: any consistent effect which cannot be accounted for in performance terms we treat as "syntactically relevant," whether or not it has been traditionally considered to be so. Note that we shall discuss this issue in terms of "syntactic (ir)relevance" rather than as the Grammaticality/Acceptability distinction, in order to free ourselves of traditional assumptions about where this boundary lies.

The judgements we obtain cause us to draw some nonstandard conclusions about the nature of the grammar. First, that common assumptions about the boundary between syntactically relevant and irrelevant data need to be reassessed, and second, that grammaticality is not categorical, but a continuum. Third, in consequence of the second, constraints have a violation cost but are in principle always "survivable." We take these in order.

The first point results from our finding that phenomena regarded as syntactically relevant in one language are discounted as nonsyntactic in another. Note that we are not attempting to contest the existence of such a division, merely reassessing on the basis of far richer data where the border should be drawn. Our own preference is to exclude effects from the group of syntactically relevant factors in case of doubt: since the existence of this group is at the heart of the motivation for the generative approach to grammar, it seems to me to be only proper to be careful about which effects we admit.

The second point should be a surprise to no one; Chomsky (1957) notes that some data "requires that we generalize the grammatical-ungrammatical dichotomy, generating a notion of degree of grammaticality" (Chomsky 1957: 35, fn 2), in Chomsky (1964) he addresses "degrees of grammaticalness," and in Chomsky (1965), he states "grammaticalness is, no doubt, a matter of degree." However, at the same time as making these remarks in favor of scalar grammaticality, Chomsky's frameworks simultaneously abstracted from it, not unreasonably, since Chomsky explicitly states the need for idealization (e.g. Chomsky 1975 [1955]: 145, fn 15). The need for the noncategoricity to be restated and empirically demonstrated is due to a tendency among linguists to forget that categoricity is still an abstraction from the primary data. The counterexample model of Haider in the previous section is an example of this: syntacticians have sometimes felt it to be a prerequisite for syntactic relevance for a constraint to be categorical. We hope to demonstrate in this work that the standard assumptions are obscuring certain parts of the data set and that real syntactic progress can be made when they are not applied. But we must make one more thing clear: our relative grammaticality judgements show not a continuum between two fixed endpoints of "grammatical" and "ungrammatical," but a true continuum with no fixed points at all. It follows that there is, however, no point at which the accumulated violation costs make a structure "ungrammatical": there is no lower bound or sudden drop-off in the pattern of judgements, they just continue a linear pattern of getting worse. It follows, therefore, that a structure is never absolutely "grammatical" and "ungrammatical" in this model of grammaticality, only ever more or less grammatical.

The third point is perhaps the most crucial to keep in mind while reading this article. Our experimentally obtained data shows that constraint violations have a violation cost, but are "survivable." This means that any given constraint violation will cause a structure to be judged worse than an otherwise equivalent structure which does not violate that constraint, but it will not necessarily prevent it from being part of the language. Keller (2000) has shown convincingly that these constraint violation costs are cumulative — we also observe this in our studies. We are therefore assuming a weighting model of constraint interaction: all constraints are applied to all candidate structures and the appropriate violation costs applied. Independently and subsequently, the form to be output is decided by a probabilistic competition function. The form which has the least cumulative violation costs is therefore the most common output, but forms which are only a little worse will also be produced on occasion. When we refer to a "constraint" in this article, we mean a factor which systematically reduces the judged grammaticality of a structure, but no more than this: it does not mean that the violating structure cannot be a part of the language, though it may appear less frequently. Those data sources which apparently reveal a dichotomous grammaticality (frequency data, binary grammaticality judgements) are, we argue, drawing on intuitions of frequency as well as structural well-formedness. So the standard question "Is this structure grammatical?" is understood as "Is this structure sufficiently grammatical for it to be produced?" as a method of imposing binarity onto a continuum. In this article we use the term "acceptable" as an informal shorthand for "sufficiently grammatical to be produced."

6. Methodology and procedure

The magnitude estimation methodology allows us to obtain maximally differentiated grammaticality judgements from a group of informants and compare them meaningfully (Bard et al. 1996; Cowart 1997; Keller 2000). It derives from a methodology used to grade physical sensations such as heat and brightness and was developed from there for use in attitude and opinion measurement (Stevens 1975; Lodge 1981). It is a variant of the standard grammaticality judgement task: informants are asked to give judgements of sentences, in numerical form, based upon their intuitions about the well-formedness of the linguistic structures presented to them. It varies from standard elicitation of grammaticality judgements in several ways. The first innovation is that subjects are asked to provide purely relative judgements: at no point is an absolute criterion of grammaticality applied. This helps avoid the distortion of prescriptive norms. Judgements are relative both to a reference item and the individual subjects' own previous judgements. Also, all judgements are proportional; that is, 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. Furthermore, the scale on which judgements are made is open-ended: subjects may always add a new highest or lowest score if they feel the need to express a more extreme value. Lastly, the scale has no minimum division: subjects can always produce an additional intermediate rating. The net result is that subjects are able to produce judgements that distinguish all the differences they perceive, with minimal interference from external factors.

Subjects logged themselves on to the experimental website and participated in the experiment remotely, as the experiment was made available on the web using the WebExp experimental software package (Keller et al. 1998, http://www.language-experiments.org/). The experiment proceeded as follows: first subjects encountered a page of instructions outlining what they were being asked to do. The nature of the task was explained and the phases of the experiment described. It was stated that the object of interest was the spoken language, rather than the written form, and that the criterion they were to judge by was whether the sentences "sound natural." They were additionally instructed to complete the task briskly since it was their first impressions that were of interest. Subjects next filled in a personal details form, and subsequently carried out two practice phases. The first practice phase was designed to familiarize them with magnitude estimation; they were instructed to assign numeric values to line lengths relative to a reference line. This was followed by a second practice phase that extended the use of magnitude estimation to judging sentence naturalness. The materials in this phase were carefully chosen to reduce the initial shock of meeting some of the less grammatical sentence types from the experiment, but still aimed to represent the full range of acceptability. Only after this stage did the elicitation of the judgements of structures reported here begin.

Thirty-eight subjects recruited by flier took part in this experiment. Each saw a version of the materials such that each syntactic condition and each item appeared once, randomly mixed among another eighteen filler items. Participants were asked to supply their names, ages (mean age 29.1, range 19–52), sex (16 females, 22 males), occupations (all but 3 students or graduates), and dialect backgrounds (16 from the southern areas of Bavaria or Baden-Württemberg, 5 from the central areas, 5 from the north, 12 claimed no dialect background).

7. Superiority

In order to test for superiority, we tested twenty-six different multiple wh-question structures, hoping thus to establish whether German has such an effect, and if so, which combinations of grammatical functions as wh-items would trigger it. Twenty-six sentences of the form of (7) were constructed, such that each could be transformed into the multiple wh-questions we tested.

(7) Der Zahnarzt hat dem Patienten die Zahnpasta
the dentist has the.DAT patient the.ACC toothpaste
empfohlen
recommended
"the dentist recommended the toothpaste to the patient"

The structures and lexis 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 (7–11 letters, mean lengths: subjects 8.6, indirect objects 8.6, direct objects 8.6) and lemma frequency from the CELEX Lexical Database[2] (mean per million figures, subjects 30.4, indirect objects 41.6, direct objects 35.5).[3] The twenty-six target structures are shown in Table 1. We tested multiple wh-questions containing every pair of wh-subject *wer* ('who'), wh-direct object *was* ('what'), d-linked wh-direct object *welches X* ('which X'), wh-indirect object *wem* ('to whom'), d-linked wh-indirect object *welchem X* ('to which X'), and wh-adjunct *wann* ('when').

Table 1. *Superiority experiment design*

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

The first five experimental structures, corresponding to the first row of Table 1, were thus as in (8). We do not list all twenty-six forms for reasons of space.

(8) a. Wer hat dem Patienten was empfohlen?
who has to.the patient what recommended
"Who has recommended what to the patient?"
b. Wer hat dem Patienten welche Zahnpasta empfohlen?
who has to.the patient which toothpaste recommended
"Who has recommended which toothpaste to the patient?"
c. Wer hat wem die Zahnpasta empfohlen?
who has to.whom the toothpaste recommended
"Who has recommended the toothpaste to whom?"
d. Wer hat welchem Patienten die Zahnpasta empfohlen?
who has to.which patient the toothpaste recommended
"Who has recommended the toothpaste to which patient?"
e. Wer hat dem Patienten die Zahnpasta wann empfohlen?
who has to.the patient the toothpaste when recommended
"Who recommended the toothpaste to the patient when?"

Note that the direct and indirect objects appear twice: once as d-linked and once as bare wh-elements. This was for two reasons. First, we intended to test for d-linking effects in this experiment, and the inclusion the two different forms allows us to do this. Second, the German wh-item *was* ('what') is frequently used in the spoken language as an indefinite, a short form of *(irgend)etwas* ('something'); to a more restricted extent *wem* 'to whom' can be used as a short form of the indefinite *irgendwem* ('to somebody'). We controlled for this potential misreading by including the *which NP* form of *was* and *wem*, since these (*welches X*, *welchem X*) do not permit the alternative reading as indefinites. Let us note two unavoidable irregularities in the materials. Each NP may appear in three different forms: as a standard definite NP, as a bare wh-item, and as a *which NP* wh-item, but it is not possible to match these three forms for length — bare wh-items are always shorter. Some length effects must therefore be allowed for in the results. Additionally, the temporal adjunct was only included in the experimental materials when it was a wh-form. The reason for this is that the location within the Mittelfeld of such adjuncts is subject to a range of interacting factors, and so we should either have had to position it differently in different structures, or else put it between the direct object and the verb, both of which would, in our estimation, have caused greater variation than its omission except as a wh-form. When the wh-form *wann* was in situ, it was always placed in Mittelfeld-final position before the verb. This corresponds to its probable generation site, but is a somewhat marked position when, as here, it is preceded by two full NPs. Nevertheless, we felt that the highest priority was that it should appear in its presumed generation position.

8. Results

For the charts below, the data was first normalized by subtracting the subject's mean from each score and then dividing it by the subject's standard deviation (z-scores). This effectively unifies the different scales that the individual subjects adopted for themselves and allows us to inspect the results visually. The most significant result for our present purpose is presented in Figure 1, which shows, on the scale axis, the mean normalized grammaticality judgement score and 95% confidence interval by sentence type. 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 abbreviate bare wh-direct object as *wh-Do*, *which-X* direct object as *wx-Do*, and wh-adjunct as *wh-adj*.[4]

The most obvious result is the poor grammaticality of structures with in-situ subjects compared to all the others. Multiple wh-questions with in-situ subjects are clearly less acceptable than other types of multiple wh-questions. In a repeated measures ANOVA, the effect of the in-situ wh-item type is highly significant both by subjects ($F_1 = 30.73$, $p_1 < 0.001$)

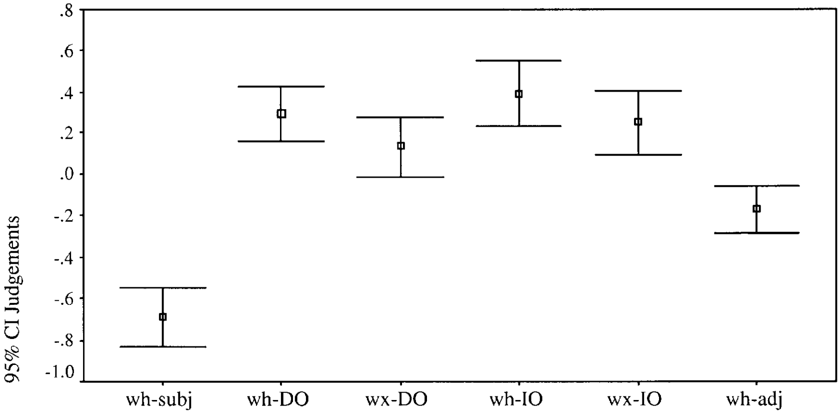

Figure 1. *Results of experiment on superiority in German. Judgements are distinguished by in-situ wh-item. Error bars show mean judgement and 95% confidence interval of mean*

and by items ($F_2 = 45.86$, $p_2 < 0.001$). 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 $p > 0.1$).[5]

Since a dispreference for in-situ subjects is precisely the core empirical content of the superiority constraint, it would appear that we have detected a superiority effect in German. There is also some variation among the other conditions, however. First, the direct object and indirect object bare wh-items are better than their *which-NP* equivalents. We do not wish to assign any syntactic significance to this, however, because we can account for this variation in other ways. Recall that the bare wh-items *was* and *wem* are used in the spoken language as indefinite pronouns: it seems likely that this could be improving the scores of the wh-DO (*was*) and wh-IO (*wem*) conditions over the d-linked wx-DO (*welches NP*) and wx-IO (*welchem NP*) conditions. Also, the bare items are significantly shorter than their *which-NP* equivalents, and constituent heaviness is known to affect scores. Lastly, this difference does not correspond to any known syntactic effect: certainly it is not d-linking, since this would predict the opposite effect, that is, the in-situ *which-NP* types should be better than bare wh-items, not worse. In the light of these considerations, we need not attribute this variation to any syntactic cause.

The rather larger difference between the score of the wh-adjuncts and the bare wh-object conditions can be accounted for by similar factors. Firstly, the sentence types with wh-adjuncts are longer and thus heavier, for just in these cases we have three, rather than two constituents in the Mittelfeld — see (8) above. It is also the case that the in-situ position of the *wann* ('when') immediately before the verb in our experimental materials is somewhat marked (cf. [9a]). As a "light" element and a pro-form it would appear more naturally in an earlier surface position before or among the arguments in the Mittelfeld (9b) and (9c).

(9) a. Wer hat dem Patienten die Zahnpasta WANN empfohlen?
who has the patient the toothpaste when recommended
"Who recommended the toothpaste to the patient when?"
b. Wer hat dem Patienten WANN die Zahnpasta empfohlen?
c. Wer hat WANN dem Patienten die Zahnpasta empfohlen?

This word order dispreference is very probably depressing these scores, but since this cannot be quantified, we are unable to exclude the possibility that other factors are also playing a role in the relative weakness of the in-situ adjuncts.

One more aspect of these results which may be syntactically interesting: the data reveals only the weakest of contrasts between the direct and indirect objects. This offers a potential contrast with English, where sentences such as (10a) seem better than (10b).

(10) a. Who did you send what?
b. ?What did you send who?
c. What did you send to who?

The existence of superiority effects between direct and indirect objects is a key question in the investigation of the trigger of the superiority effect (see the discussion of categorical and relative accounts of superiority above). However, the contrast in (10a) and (10b) may have a different cause, namely, the restricted circumstances in English in which NPs can be used as indirect objects. PPs are often the more natural form: in our specific case, (10c) seems rather better than (10b). In order to compare the German and English findings more precisely, we carried out a parallel study on superiority in English which we report below. There we perform additional crosslinguistic comparisons.

Further light can be thrown on the German results by looking at the data in greater detail. In Figure 2 below we see the in-situ wh-items on the horizontal axis (as in Figure 1), but the error bars show the scores differentiated by raised wh-item. As above, these scores are given as 95% confidence interval bars, with the marker type at the mean point indicating the raised wh-item (see legend beside graph).[6]

Let us state our major conclusions on the basis of this data straight away: there is no syntactically relevant interaction of raised and in-situ wh-item types; that is, the ability of a wh-item to appear in situ is largely

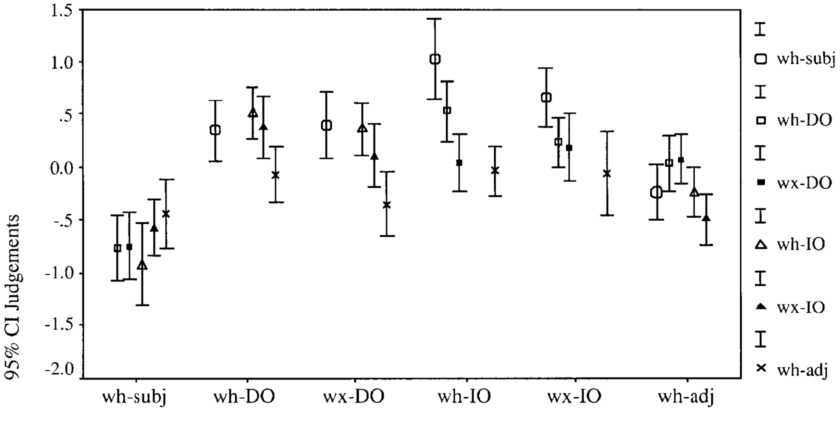

Figure 2. *Superiority in German: results by in-situ wh-item on the horizontal axis as above, but with scores differentiated by raised wh-item within the groups of error bars*

unaffected by the wh-item in the raised position. In the following section of the text we shall seek to justify this interpretation of the data in Figure 2. This needs to be argued for, because in fact there is some other systematic variation in the data.

Figure 2 shows the same data as Figure 1, but additionally distinguishes the effect of the raised wh-item. On the left, we see the clearly degraded status of the in-situ subjects, and furthermore we see that there is little difference by raised wh-item: the error bars are closely bunched. Across the middle of the graph are the direct and indirect objects, all four conditions showing rather more spread than the in-situ subjects, but with fairly even mean values by in-situ wh-item. The in-situ wh-adjuncts on the right pattern more closely together, but are a little weaker than the objects. Across the four in-situ object conditions, the weakest condition is always the one with the raised wh-adjunct. We mentioned above that the conditions with wh-adjuncts, raised or in-situ, were always one constituent heavier than those without wh-adjuncts. Figure 2 shows us that this dispreference for structures with wh-adjuncts is independent of their position: the scores of the conditions with in-situ wh-adjuncts are very consistent with the score of the conditions with raised adjuncts. We may therefore attribute these lower scores to this heaviness factor with some confidence. There is, interestingly, one exception to this tendency: the condition with an in-situ wh-subject is actually best with the raised wh-adjunct. This effect is probably caused by the same word order factor which causes some conditions with initial subjects to be judged better. We therefore turn to these to outline why this might occur.

The raised wh-subjects pattern with the other conditions with the in-situ wh-adjuncts and direct objects, but seem to be scored rather better when combined with the in-situ indirect objects. We do not think this is a factor relevant to superiority, however, since there is a ready explanation in terms of a surface word order effect. It is a descriptive generalization that light constituents and pro-forms tend to precede full NPs in the Mittelfeld (Behagel 1909; Lenerz 1977; Uszkoreit 1987; Featherston 2002). The *wh-subj wh-IO* sentence type fulfils this constraint (11a), and is thus judged better than the other in-situ object scores, which do not, for example, (11b). It is no surprise that the bare wh-IO *wem*, which is a very light constituent, shows this effect more strongly than the heavier wx-IO *which-NP*. It seems likely that this effect is also responsible for the relatively good performance of the *wh-adj wh-subj* condition (11), since the in-situ wh-item fulfills this constraint here as well. It is interesting to note that superiority violating conditions benefit from this effect, while most superiority fulfilling conditions violate it. The apparent superiority effect is thus reduced by the non-negligible violation cost of this word order effect.

(11) a. Wer hat wem die Zahnpasta empfohlen?
who has to.whom the toothpaste recommended
b. Wer hat dem Patienten was empfohlen?
who has to.the patient what recommended
c. Wann hat wer die Zahnpasta empfohlen?
when has who the toothpaste recommended

Let us note that it is this difference which is almost solely responsible for the slightly better scores of in-situ wh-IOs than wh-DOs, which we noted in Figure 1. There is thus clearly nothing in the German data which corresponds to the DO/IO extraction asymmetry which would be predicted by a competitive superiority constraint.

We shall mention one more effect which is visible; there is a tendency for bare wh-items to be better than d-linked wh-items both in-situ and raised, which is probably merely a length effect. This explanation is supported by the across-the-board appearance of the effect. Other than these "heaviness" and "pronoun first" effects (and it seems likely that the latter is a sub-form of the former) effects, the in-situ direct and indirect objects show a noteworthy consistency both across object type and d-linking status, and across raised wh-items. We can thus conclude that there is no evidence of syntactically relevant differences between the nonsubject wh-items either in raised or in-situ positions in this data.

To summarize, we may say that this data shows a clear degradation in acceptability for in-situ wh-subjects, but no other variation which requires a specific syntactic explanation relevant to superiority. It is clear on this basis that German does have a superiority effect, with a plain dispreference for any other wh-item being raised over a subject wh-item. The standard assumption in the literature is therefore defeated. Note also that the claims that superiority applies in a narrower range of circumstances in German than in English are also falsified, since we tested for short superiority, which Grewendorf (2001) excludes, and with non-d-linked forms, which Wiltschko (1997) denies.

We raised the question above whether these results might allow us to adjudicate between the relative (e.g. minimal link condition) and the categorical (e.g. ECP) accounts of superiority. Interestingly, they go some way towards doing so. Recall that the ECP-related account predicts that a particular wh-item type will be either acceptable or unacceptable in situ, while the MLC approach requires that grammaticality will be a function of the behavior of pairs of wh-items. On this German data, the ECP account wins convincingly, for our finding of a subject-versus-the-rest superiority is readily compatible with the ECP account. We see no sign of pair-related extractability, though admittedly we cannot entirely exclude it on this basis of this data. We cannot absolutely rule out the MLC account type for there are possible explanations. For example, if the distance metric feeding superiority makes use of hierarchical position, but German has something like a flat VP, then even an economy-based superiority would produce exactly the data we find. While such possibilities exist, see Chomsky's (1973) suggestion that our movement from the right of the verb is equidistant, and the examples in (3), our data must nevertheless strongly favour an ECP-type account.

One question on which this data could give us no information was that of d-linking in German. Since d-linking essentially takes the form of the cancellation of otherwise predicted superiority effects, and we only tested for this in the direct and indirect objects, where however no superiority effect was found, our data fails to throw any light on this question. For this reason, we conducted an additional experiment to investigate this question further.

9. Discourse linking

This experiment was a follow-up experiment to the previous one and used eight of its items as its materials. It aimed to check for d-linking effects between subjects and objects, where we had demonstrated a superiority effect. We therefore tested structures with just subject and direct object wh-items, but with bare wh-items and d-linked wh-items at each position (cf. [12]).

(12) a. Wer hat dem Patienten WAS empfohlen?
who has to.the patient what recommended
b. Wer hat dem Patienten welche Zahnpasta empfohlen?
who has to.the patient which toothpaste recommended
c. Welcher Zahnarzt hat dem Patienten WAS empfohlen?
which dentist has to.the patient what recommended
d. Welcher Zahnarzt hat dem Patienten welche Zahnpasta
which dentist has to.the patient which toothpaste
empfohlen?
recommended
e. Was hat WER dem Patienten empfohlen?
what has who to.the patient recommended
f. Was hat welcher Zahnarzt dem Patienten empfohlen?
what has which dentist to.the patient recommended
g. Welche Zahnpasta hat WER dem Patienten empfohlen?
which toothpaste has who to.the patient recommended

h. Welche Zahnpasta hat welcher Zahnarzt dem Patienten
which toothpaste has which dentist to.the patient
empfohlen?
recommended

The methodology was as in the previous experiment. Thirty subjects took part and were asked for their age (mean age 26.1, range 20–42), sex (20 males, 10 females), occupation (24 students, 6 university employees), and dialect background (15 from the south, 2 from the center, 5 from the north, 8 claim no dialect influence). Subjects also saw 25 other sentences as fillers. Unlike in the previous experiment, bare in-situ wh-items (*wer*, *was*) were capitalized. 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 indefinites (*was* = *etwas* 'something,' see above).

Figure 3 illustrates the results of this experiment using the same conventions as the previous 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: in each pair of alternative orders, the version with a raised subject and in-situ object (12a)–(12d) is always judged better than that with a raised object and in-situ subject (12e)–(12h); the precondition for testing for a discourse linking effect is thus fulfilled. And this effect is indeed visible: the worst two scores are precisely those which have an in-situ bare wh-subject, while the two with in-situ d-linked subjects are almost as

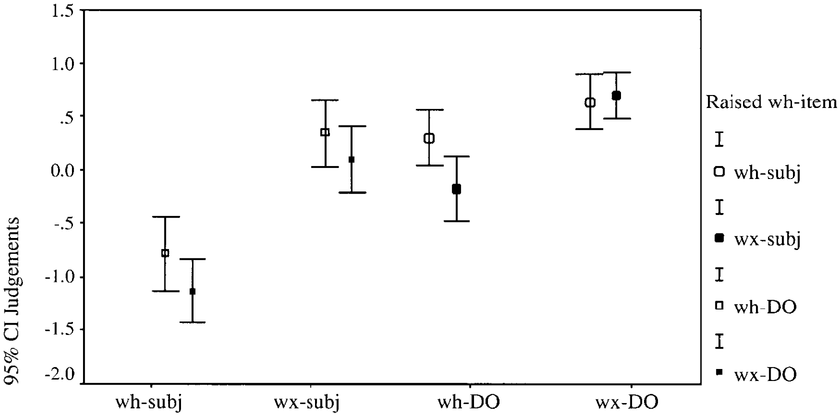

Figure 3. *Results of experiment on the discourse linking effect in German. In-situ wh-items on the horizontal axis, error bars distinguish raised wh-items. There is a clear improvement when an in-situ wh-subject is in d-linked form*

good as their equivalents which do not violate superiority. This is striking empirical confirmation of the reality of the discourse linking effect.[7]

A repeated measures analysis of variance reveals the statistical robustness of these effects: there is a significant effect for "argument order" (Subject < Object, Object < Subject) both by subjects ($F_1 = 46.2$, $p_1 < 0.001$) and by items ($F_2 = 68.9$, $p_2 < 0.001$), as well as an interaction of "argument order" and "in-situ wh-item type" (bare wh-item, d-linked wh-item) both by subjects ($F_1 = 5.71$, $p_1 = 0.025$) and by items ($F_2 = 51.65$, $p_2 < 0.001$). We may conclude that Pesetsky's d-linking phenomenon exists in German as in English. So not only does German respect superiority, but it displays exactly the same exception to it as English. This must effectively quash any suspicion that the dispreference for in-situ wh-subjects we identified in German is anything other than the superiority effect generally acknowledged in English. An effect with such a specific exception cannot be mistaken.

There are two other effects in this data we might mention. First, clause-initial d-linked wh-items are generally worse than non-d-linked ones, an effect which was only marginally visible in our previous experiment. There is also a consistent trend for in-situ d-linked items to be better than bare ones, independent of the Pesetsky's d-linking phenomenon. This is in stark contrast to the data of the first experiment, where the reverse was true. We attribute this difference to our capitalization of the in-situ bare wh-items in this experiment: our attempt to prevent these from being interpreted as indefinite pronouns seems to have worked. We discuss possible causes for these effects in Featherston (2005).

10. Superiority effects and clause type

A reviewer comments that it might have been better to test superiority in embedded clauses, arguing that monosentential multiple wh-questions pragmatically require a sorting procedure, with the first wh-element as the sorting key. In a pragmatically unmarked setting, the first wh-item is taken as the topic and as the unmarked sorting key (Kuno 1982). The appearance of a wh-subject as the second wh-item could thus be pragmatically dispreferred. Now there may well be some truth in this; in particular we give some credence to the argumentation of Garret (1996) about the reasons for the restrictions on adjunct wh-items in initial position, but we doubt that it would radically change our findings. Our own intuitions of superiority effects in embedded and main clauses in English reveal no such differential effect (13).

(13) a. Who sent the customer what?
b. *What did who send the customer?
c. I want to know who sent the customer what.
d. *I want to know what who sent the customer.

We might also note that other syntacticians have discussed the superiority effect on the basis of examples in main clauses (e.g. Chomsky 1981: 255).

Nevertheless, in the light of the suggestion that there might be a differential superiority effect by clause type, we carried out a corpus study of some parts of the data set in question. Here we had two main aims: first, to learn whether occurrence frequency would support our judgement data in showing a strong dispreference for embedded wh-subjects; and second, to test whether main and embedded clause types show signs of behaving differently in these structures. The table below shows the numbers of multiple wh-questions of different sorts in the COSMAS corpus of German.[8] We divided the results into four different clause types: subject clauses, complement clauses, free V-final clauses, and main clauses. We illustrate these in (14).

(14) a. Wer was bauen soll, ist noch offen.
who what build should is still undecided
b. Er weiß nicht, wer was will.
he knows not who what wants
c. Wer was am Sonntag erwarten darf.
who what on Sunday expect may
d. Wer hat damals was gewußt?
who has then what known

Only very few examples found did not fit into any of these categories. Note that both subject and complement clause types occurred both sentence-initially and sentence-finally. We did not distinguish these.

Table 2. *Frequency of multiple wh-question types with wh-subjects in COSMAS corpus of German*

Clause type	*wer*	*was*	proportion	*wer*	*wen*	*wer*	*wem*
	was	*wer*	(%)	*wen*	*wer*	*wem*	*wer*
Subject	129	5	3.9	144	0	37	0
Complement	513	31	6.0	281	0	117	0
Free V-final	12	1	8.3	23	0	5	0
Main clause	309	7	2.3	534	0	106	0
Total	963	44	4.6	982	0	265	0

Table 3. *Frequency of multiple nonsubject wh-question types in COSMAS corpus of German*

Clause type	*wem was*	*was wem*	*wem wen*	*wen wem*
Subject	0	1	0	0
Complement	19	20	0	0
Free V-final	1	1	0	0
Main clause	6	10	0	0
Total	26	32	0	0

Let us first note that the superiority effect we found in our judgement data is replicated in this frequency data. The frequencies of in-situ subjects (*was wer*, *wen wer*, *wem wer*) are far lower than their equivalents with initial subjects (*wer was*, *wer wen*, *wer wem*), which may be taken as a reflection of the superiority effect. There is no similar effect among the nonsubject pairs *wem was* and *wem wen*, again in line with our finding in the judgement data of no strong effects between nonsubjects. Notice that the attested occurrence of superiority violations does not, as is often thought, disconfirm our claim of a superiority effect in German. The reason is that we are continuing to assume the continuum of grammaticality that our relative judgements reveal. The effect of a constraint violation in such a model of grammaticality is only ever to reduce the violating structures judged grammaticality, not to render it ungrammatical. This "reduced grammaticality" is exactly what the frequency data shows. Every violation imposes a violation cost which means that the structure is less likely to be produced, but does not exclude it as part of the language. In sum, this frequency data would tend to corroborate our previous findings from judgement data.

Since we have established that the frequency data corresponds to expectations on the basis of our experimental findings, we now consider whether there is reason to suspect that the judgement results we found in main clauses would have been materially different had we tested complement clauses. This data shows little support for this suggestion. It is true that the proportion of *was wer* is rather higher in the complement clause type than in the main clause type (6.0% vs. 2.3%), but it is very clear that the background dispreference for in-situ subjects is consistent and overwhelming with all raised wh-items and across all clause types. Nor is this difference between complement and main clause types repeated with any other pair of wh-items. We thus see no reason to suspect that testing for superiority in embedded clauses would have produced a materially different result.[9] It is possible, probable even, that a sorting effect exists, but it is clear from this data that our finding of a superiority effect is independent of it.

11. Superiority in English

Since the significance of some of our findings in German have depended on their correspondence with English, we decided to obtain judgement data on superiority in English as well, using the same methodology. Thirteen basic sentence items were constructed of the form of (15) from which twenty-six multiple wh-questions were derived.

(15) The dentist showed the patient the toothpaste.

Three changes were made to the syntactic conditions used in German. First, in English no adjunct wh-item was tested. There seem to be extra factors which regulate the possibility of adjunct wh-items in multiple wh-questions, and these cause different wh-adjuncts to behave differently. In the light of this, we decided to omit wh-adjuncts from this study. A second change to the German experiment was that we included d-linked wx-subjects straight away, in the light of our findings in the first experiment. The last change from the German conditions was an addition: we tested the indirect objects in initial position both as "dative-shifted" NPs and as PPs (cf. [16]).[10] The reason for this was that we feared that the relative restrictedness of the first alternative might give rise to effects which would cloud the data, if we failed to control for them.

(16) a. Who did the dentist show what?
b. Who did the dentist show what to?

We did not test for in-situ prepositional indirect objects, however, since we doubt that these will reveal any additional relevant effect. Our own judgements detect no difference between these two structures (cf. [17]).

(17) a. What did the dentist show who/the patient?
b. What did the dentist show to who/the patient?

The complete set of conditions that we tested in English is presented in Table 4.

The first two rows of structures are exemplified in (18).

(18) a. Who showed the patient what?
b. Who showed the patient which toothpaste?
c. Who showed who the toothpaste?
d. Who showed which patient the toothpaste?
e. Which dentist showed the patient what?
f. Which dentist showed the patient which toothpaste?

The lexis was strictly controlled to minimize background variation. The sentences were as closely matched to their German equivalents as possible

Table 4. *Syntactic conditions for experiment on superiority in English*

Initial wh-item	In-situ wh-item					
	wh-subj	wx-subj	wh-DO	wx-DO	wh-IO	wx-IO
wh-subj			who sent IO what	who sent IO which thing	who sent who DO	who sent which IO DO
wx-subj			which S sent IO what	which S sent IO which DO		
wh-DO	what did who send IO	what did which S send IO			what did S send who	what did S send which DO
wx-DO	which DO did who send IO	which DO did which S send IO			which DO did S send who	which DO did S send which IO
wh-IO	who did who send DO		who did S send what	who did S send which DO		
wx-IO	which IO did who send DO		which IO did S send what	which IO did S send which DO		
wh-IO … *to*	who did who send DO to		who did S send what to	who did S send which DO to		
wx-IO … *to*	which IO did who send DO to		which IO did S send what to	which IO did S send which DO to		

with respect such factors as animacy and definiteness. NPs were matched for length (5–10 letters, mean lengths: subjects 7.1, indirect objects 7.0, direct objects 7.4) and lemma frequency from the CELEX Lexical Database (mean per million figures, subjects 33.8, indirect objects 61.5, direct objects 42.3). Thirty subjects were recruited by advertisement in British university newsletters. Each saw a version of the materials such that each syntactic condition appeared once and each item twice, randomly mixed among another thirteen filler items. Participants were asked to supply their names, ages (mean age 34.5, range 21–56), sex (18 females, 12 males), occupations (all but 3 students or employees of British universities) and dialect backgrounds (11 southern England, 12 northern England, 7 other [Scots, South Africa, US, etc.]). The data were processed as in the previous experiments.[11]

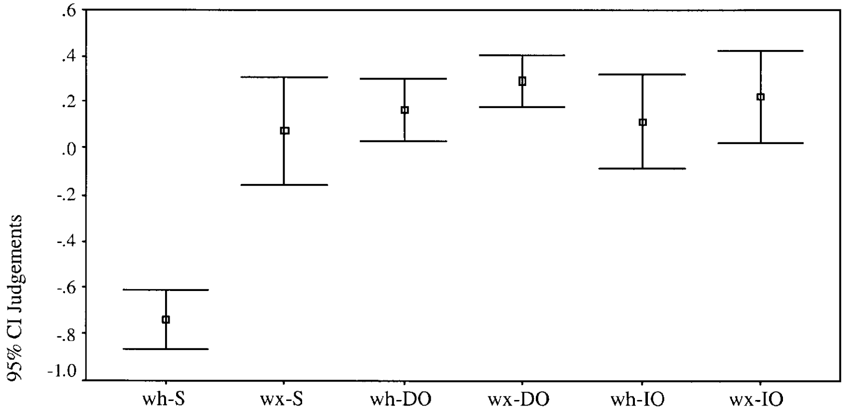

Figure 4. *Results of English superiority experiment, showing judgements by in-situ wh-items only*

The most obvious feature of these English results is that they closely resemble the German results. The in-situ subjects are clearly worse than any other in-situ wh-item type, and the differences between the others are fairly marginal. In particular, there is no apparent asymmetry between direct and indirect objects; on the contrary, they show a remarkable uniformity. This is of importance for our investigation into the nature of the superiority effect. We may next note that the d-linked in-situ subjects are very nearly as good as the other conditions: here too we find a close correspondence to the German discourse linking data. Further, we see that there is a general preference for in-situ wh-objects, direct and indirect, to be d-linked. This too is parallel to our finding in Experiment 2.

Statistical tests show the robustness of these findings: in repeated measures analysis of variance there is a significant effect for the factor "in-situ wh-item type" both by subjects and items ($F_1 = 18.95$, $p_1 < 0.001$; $F_2 = 27.10$, $p_2 < 0.001$), while a Tukey post hoc test demonstrates that the in-situ subjects are different from the other groups (all $p < 0.001$), while these others do not differ significantly (all $p > 0.05$). The preference for d-linked wh-items is significant ($F_1 = 26.32$, $p_1 < 0.001$; $F_2 = 20.07$, $p_2 = 0.001$) as is the interaction of the factors "grammatical function" and "d-linking" ($F_1 = 12.13$, $p_1 < 0.001$; $F_2 = 9.41$, $p_2 = 0.001$).

This result has strong implications for our concept of superiority. This result must exclude absolutely any residual doubt that the effect we noted in German was indeed superiority as we know it from English, since the effects in German and English are so similar. This is solid evidence that German does indeed know this constraint, although German syntacticians had assumed otherwise. We discuss why this surprising situation might have come about and what lessons we might draw for syntactic practice below, but we shall note here that we attribute this failure to syntacticians' assumption of a categorical model of grammaticality, which is obscuring relevant data.

We turn next to the full pattern of all 26 conditions in this experiment presented in Figure 5. At first glance, this data seems to show a greater degree of variation than the German results, but we hope to convince the reader that the additional variation does not in any way defeat the more general conclusions we have drawn so far; it merely reflects two additional subregularities.[12]

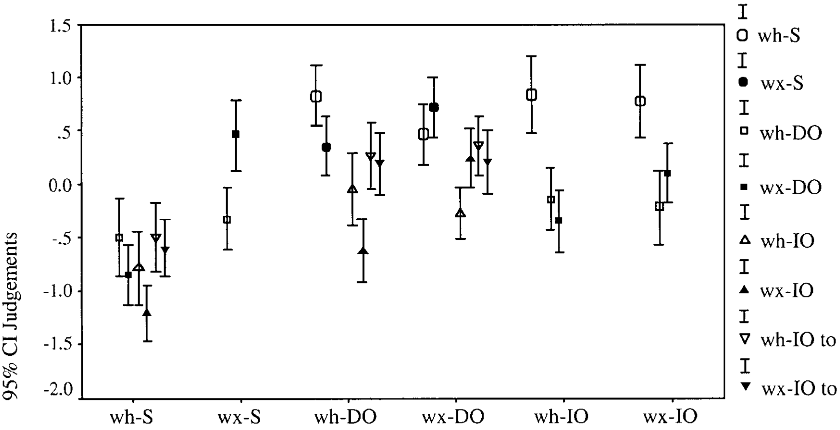

Figure 5. *English superiority experiment results, distinguishing in-situ wh-item on the horizontal axis, and raised wh-item in the groups of error bars*

The basic contrast of in-situ wh-subjects versus all other conditions is clearly visible, as is the discourse linking effect in d-linked in-situ wh-subjects. But we can also clearly see one interesting factor which is a real difference to the German results: in German there was only the most marginal effect of the grammatical function of the raised wh-item. In this English data, on the other hand, we see a clear preference for wh-subjects in clause-initial position, independent of what wh-item type is in situ. In all four conditions with nonsubjects as in-situ wh-items, the best scores are obtained when a wh-subject, bare or d-linked, is in the raised position: all others are scored lower (Tukey HSD test: all $p < 0.001$). This is a particularly interesting finding, since it gives us a possible account of why superiority violations are judged fully unacceptable in English, but only dispreferred in German. We can see that the presumed single superiority effect in English is in fact made up of two cumulative effects: first, a dispreference for in-situ bare wh-subjects, but second, another dispreference for any raised nonsubject wh-item.[13] The net effect of these two constraints is to exclude any structure with an in-situ wh-subject. In German on the other hand, while the dispreference for in-situ bare wh-subjects is quite clear, there is no equivalent dispreference for raised wh-nonsubjects (Tukey HSD test: wh-subj vs. wh-DO $p = 0.598$, wh-subj vs. wh-IO $p = 0.294$). Since German has only one of the two contributing constraints, the cumulative "superiority effect" appears weaker. It therefore rather looks as if the difference between English and German superiority effects is due merely to the preference in English for subject-initial clauses.[14]

Another factor causing variation between conditions seems to be a preference in English for consistency in d-linking status between multiple wh-items. Independent of Pesetsky's d-linking effect in in-situ subjects, there is a preference for the d-linking status of the superior and inferior wh-items to be the same, that is, both bare types or both d-linked (see Figure 6 below). This interaction is absent in German; there in-situ d-linked wh-items were consistently judged better and clause-initial d-linked wh-items were judged worse, but there was no interaction between the two (see Figure 2). This English d-linking matching effect seems to be behind the fairly poor score of the *wx-IO wh-DO* condition in Figure 6, but it is present across all other conditions except those with a raised PP-type indirect object. We have no real explanation for this effect, but we doubt that it is narrowly syntactic: it seems to be more likely to be related to the suitability of the structure for use as an echo question. An echo question, we might say, is a discourse-linked structure, and it is natural for a discourse-linked wh-item to be used for its question element. If a multiple wh-question is interpreted as an echo question (cf. Sobin 1990), only the

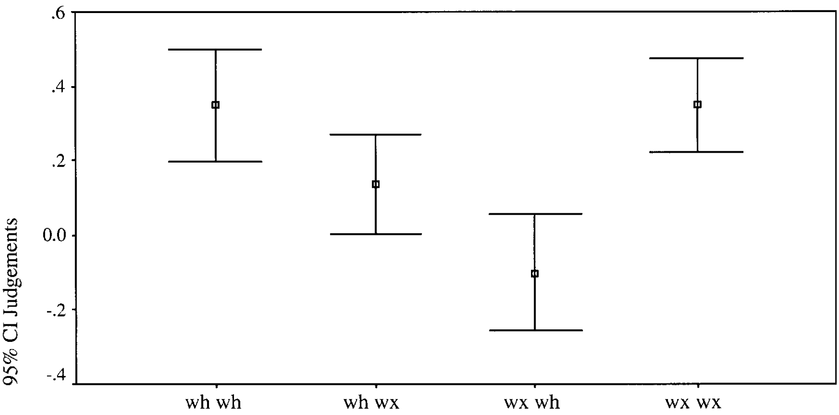

Figure 6. *The interaction of d-linking status: an in-situ, d-linked wh-item is better after a raised d-linked wh-item, but clearly worse after a raised bare wh-item. In general, there is a dispreference for a mismatch in d-linking status*

in-situ wh-item can be the element questioned in the echo question, since there is no felicitous precedent in discourse for the alternative; that is, (19a) is a possible conversational exchange, (19b) isn't, because the A contribution contains an in-situ wh-item.

(19) a. A: Who/which baker sent the duchess the birthday cake?
B: Who/which baker sent the duchess what/which cake?
b. *A: The baker sent the duchess what/which cake?
B: Who/which baker sent the duchess what/which cake?

It follows that informants, given a structure such as B's, assume that it has the antecedent as in (19a), that is, that the in-situ wh-element is the question element added in the echo question. Since, by assumption, a discourse-linked wh-item is very felicitous in a discourse-linked utterance such as an echo question, it is unsurprising that structures with the pattern *wh* ... *wx* are judged better than those with *wx* ... *wh*. In our opinion, felicity in echo questions is at the base of many, if not all, d-linking effects, but the specific pattern in this data set offers particularly strong support for this, since precisely the structure which the account predicts to be actively dispreferred, the *wx* ... *wh*, is clearly weaker, while the other combinations show less variation.

A final feature we might look for in these results is the behavior of the conditions in which neither wh-item is a subject, since these provide the test cases for the issue of the nature of the superiority effect. In fact, the English data also shows no evidence of any superiority effect between

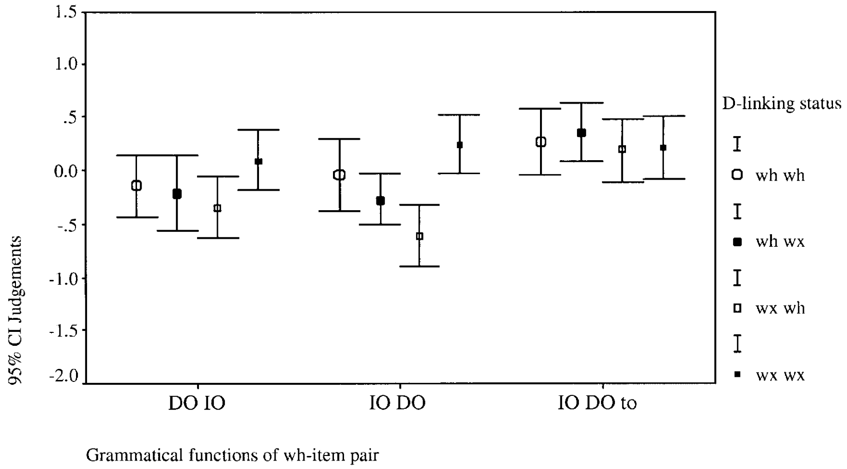

Figure 7. *Testing for effects among nonsubject wh-items: this graph shows all the conditions without a wh-subject. On the horizontal axis, we distinguish the combinations of arguments tested: direct objects, NP indirect objects, and PP indirect objects. The groups of error bars distinguish the conditions by d-linking status*

nonsubjects. Recall that the "shortest move" competitive account of the superiority effect would predict that we should find an inverse correlation: to the extent that one wh-item of a given wh-item pair is extractable, the other should not be. Figure 7 shows all the conditions which consist of nonsubject pairs, and allows us to judge whether there any asymmetry between direct and indirect objects. In fact, there is no relevant difference; there are some slight d-linking related preferences, but there is absolutely no sign of an extraction asymmetry between the *DO IO* and the *IO DO* pair (Tukey HSD test $p = 0.978$). There is some difference in the *DO IO* and the *IO DO to* pair, but then these are not true equivalents, and anyway most of this difference is related to the unexplained lack of d-linking effect in the *IO DO to* condition.

The superiority effect in English as in German, we may conclude, is restricted to wh-subjects, and is related to individual wh-item types, not wh-item pairs. This effectively excludes any account based on competition or economy as a possible trigger of the effect. Now it is true that this data does not in itself falsify a shortest move account coupled with a completely flat VP: technically, this approach is possible. But this account is an uneconomical way of motivating an effect which is specific to subjects, and we therefore interpret these results as strong support for a generation site related superiority trigger, such as the ECP.

12. Discussion

Contrary to the general assumption amongst linguists, our experimental data shows a very clear picture of movement constraints operating in German. The results show consistent and significant effects for the wh-movement constraints tested: superiority and discourse-linking. Let us note here that we have also found a similar pattern of data in the case of the *that*-trace effect, but space has not allowed us to discuss this here (see Featherston 2003b). This has a number of important implications for syntax. Firstly, those approaches to the structure of the German clause which rely on the nonexistence of these movement constraints as part of their evidential base must be weakened by this finding. In particular, the argument that is sometimes made that German, unlike English, has no specific position for the subject, or equivalently in phrase structure grammar terms, no separate subject valence feature, is undermined, for we have found very clear evidence of syntactic differences in behavior between subjects and other complement types in our experiments. This must count against the view that German subjects inhabit a flat clause structure with other arguments and adjuncts, and are differentiated from them just by linear precedence restrictions (cf. Uszkoreit 1987; and the discussion in Pollard and Sag 1994; also Haider 1993). This assumption of a flat clause or at least a flat Mittelfeld would demand that there be at most differences of degree between subject and object extraction. But this prediction is not supported. Overall, therefore, this data supports syntactic models in which assign to the subject a rather different role to that assigned to other arguments.

For the nature of the superiority effect we have found quite conclusive evidence. In both the English and German data we observed the same pattern of judgements: in-situ wh-subjects worse than any other in-situ wh-element but no differences between other conditions. We see no evidence of a hierarchy of extractability or any sign of wh-item pair-related effects in either German or English. In both languages, the dispreference for in-situ wh-subjects was neutralized by the d-linking of the in-situ subject. This data set thus offers no support at all for competition-based models of superiority that utilize a distance or economy measure: the effect is related merely to the subject position and shows no effects of pairs of wh-items. We cannot entirely exclude some additional effects with wh-adjuncts, but we have only the most feeble suggestion that these exist, and anyway it seems likely that these may be caused by other factors, the untangling of which may require some care. A question about which we can do no more than speculate is that of the motor of the superiority effect (Ginzburg and Sag 2000 is interesting on the subject). We can be confident that it is related only to the subject position and to the status of the subject as a wh-item, since subjects which are not wh-items can very felicitously remain in situ while objects are raised. We hope to address this question in further studies.

Our results will be also be of interest to syntacticians concerned with the nature of universal grammar (UG). The apparent absence of these movement constraints in languages like German suggests that these constraints cannot be universals. Our finding that these effects can indeed be found in German by experimental means must reawaken expectations that such idiosyncratic and unpredicted effects might well be found in other languages where their presence has been doubted. If further research were to show that these constraints were experimentally measurable in other languages which had previously been thought to show no sign of them, then this would be strong evidence for their being driven by universal characteristics of language. One way that this problem has been addressed is by the use of the construct of parameters. If a syntactic effect is found which would be expected to be universal, but which cross-linguistic comparison shows is not universal, then it is hypothesized that this is the outcome of parametric choice in the principles and parameters model (Chomsky 1986b). The universality of the system can then be saved by the claim that the parameter is universal. Let us note here that our data does not support this model, since it relies upon a dichotomous model of grammaticality, which, as we have seen, is merely an abstraction. Since superiority is not absent from German, we have no need to posit parametric choice.

But how could it come about that linguists have tended to deny the existence of these phenomena in German, when our data clearly demonstrates that they do indeed exist? The answer seems to be that the linguists denying these constraints and we in our own empirical work have been asking two different questions, and it is for this reason that we can come to two different answers. The "condition on extraction domains" forbids extraction "without exception," Haider argues, but "one single example is sufficient to falsify this" (Haider 1993: 159 [my translation]). But this is making an assumption about the functioning of linguistic constraints which does not on closer inspection hold. It presupposes that a violation of a grammatical constraint must rule out a structure absolutely, but it seems that precisely this condition does not apply in this case. Linguists have argued that these constraints do not apply in German because it is possible to find acceptable counterexamples, but it is clear that the counterexample model of argumentation is inappropriate. These constraints demonstrably do apply in German, but this does not necessarily lead to full ungrammaticality. It would appear that these constrains are "survivable," by which we mean that the reduction of grammaticality that their violation causes is not, or is not on its own, sufficient to exclude the structure from being part of the language.

It is worth pointing out here the difference between the notion of "survivability" and that of "violability" used in optimality theory (Prince and Smolensky 1993). In optimality theory (OT), under certain circumstances, a constraint can fail to have any effect upon the output, even though a candidate structure exhibits the structural description to which it applies. This occurs when the constraint would apply to all remaining candidate structures, or when only one candidate is left. In this case, the constraint has no effect upon the selection of the optimal candidate, put differently, it fails to apply. Our own concept of survivability has a very different nature. All constraints apply to all candidate structures and their violation costs are applied without exception. However, a given violation cost inflicted upon an otherwise perfect structure may not be sufficient to exclude the structure absolutely. Superiority in German is an example of this; its violation carries a cost in terms of diminished grammaticality, but this cost is survivable. The final status of a structure depends not on any one constraint, but upon the sum of the violation costs of the constraints that it violates. Notice, too, that our grammaticality judgement data shows that there are constraint violations with larger and smaller violation costs in perceived grammaticality.

Now the idea that there are stronger and weaker constraints, and that they are cumulative, as this approach presumes, is not new. Chomsky (1964) discusses the possible implications of varying constraint strengths, and he utilizes the differential cost of subjacency and ECP violations and violation cost cumulativity in *Barriers* (1986a), but others have noted differences in strength and made use of them as well. For example, Lakoff (1970) has "relative grammaticality"; Keller (2000) makes use of very similar data to our own and argues for two types of constraints, hard and soft, which have rather different characteristics. Our own view is that this differentiation is not necessary: we assume only a continuum of violation cost strength. Some constraints will have violation costs sufficiently large to prevent a violating structure ever appearing as a part of the language except as a slip of the tongue. Such constraints are still in principle survivable in the technical sense, although they will in practice rarely be survived. Note that the final selection of the structural candidate for output is no doubt probabilistic, so less optimal candidates for the expression of a given content will occasionally be produced (see data in Note 15).[15] This therefore provides an account of why structures thought to be ungrammatical do occur in large corpuses, and removes the need for them to be explained away as errors or slips of the tongue. The key point is that whether a constraint is in practice survived is merely incidental; it has no bearing on the nature of the constraint. Constraints only ever reduce a structure's perceived grammaticality: the intuition of absolute grammaticality or ungrammaticality is a function of frequency. When a structure is sufficiently grammatical to be produced, it is regarded as fully grammatical; when it is sufficiently ungrammatical not normally to be produced, it is regarded as fully ungrammatical.

Although the idea of variation in violation cost is not new, the default assumption in syntactic practice does still seem to be that syntactically relevant constraints must produce automatic categorical ungrammaticality, and that any syntactic constraint which does not produce full ungrammaticality is merely a stylistic preference (or markedness, or similar), not part of narrow "Grammaticality." These assumptions must be regarded as questionable however, at least in the phenomena which we have addressed in this article. Superiority in English has been thought of as producing a fully Ungrammatical output and has thus been considered related to Grammaticality. Superiority in German does not correspond to this idealization, and so the existence of the Grammatical constraint superiority in German was denied; since, by assumption, Grammaticality implies categorical ungrammaticality. The results of our judgement elicitation experiments show that both English and German superiority produce very similar effects, and that both even show the same exception for in-situ, d-linked wh-subjects, which is strong evidence that the effects have the same syntactic nature. On the basis of our data, we may claim that previous assumptions about the nature of superiority have been misguided, since it is implausible for superiority in English to be Grammatical, whereas the identical effect in German with the identical exception for d-linking is merely markedness.

Precisely what assumption should change to correct this implausibility is an interesting question. If we allow German superiority — whose violation does not absolutely exclude a structure — to be Grammatical in nature, we offend against the general assumption that Grammatical constraints are categorical. If we demote English superiority to being merely an Acceptability factor, we must cast off the tradition that superiority is Grammatical and we offend against the feeling that Acceptability should be related to preferences rather than absolute grammaticality. The third possibility is our own preferred option: the abandonment of the Grammaticality/Acceptability distinction as currently constituted. We may advance two reasons for this: first, studies such as this one demonstrate that the dividing line is being drawn in the wrong place, and second, it is not obvious on what grounds we might decide where the right place is. For a construct such the Grammaticality/Acceptability distinction to be of any use, it must be possible to judge where the dividing line is located. But this criterion is lacking: in practice linguists tend to assume traditional assignments in the literature, and in new cases apply the criterion of categoricity; a few seem to use it indiscriminately as a weapon (if data supports my theory it must be Grammatical, if it supports your theory it is just markedness) without offering any evidence to support the assignment. Even Chomsky agrees that no operational criterion might be available for Grammaticality (Chomsky 1965: 11). This criticism of the Grammaticality/Ungrammaticality distinction has of course been made before (e.g. Ross 1972; Lakoff 1973; Lightner 1976; see also the discussion of the boundary of structure factors and functional factors in Newmeyer 1999). But the point seems still to need to be made anew, as is evidenced by linguists' assumptions in the field of syntax that we have investigated here. If we have no means of detecting whether superiority is Grammar or Acceptability, the distinction cannot help us in constructing a grammar. Whereof one cannot speak, thereof one must be silent.

Which is certainly not to suggest that all differences in judgements are syntactically relevant. Garden path effects, memory overload, and all other processing effects, euphony, frequency and all other lexical effects, plausibility, truth, and all other content-related effects, individual preferences and idiolect — all of these are syntactically irrelevant and should be controlled for where possible, discounted when encountered. But all other differences in judgements, where no performance factor can be identified or even suspected, are surely best assigned to the syntax, at least until such time that an alternative explanation is forthcoming. We have applied this approach in the studies reported here, and are of the opinion that the additional insights gained into the constraints investigated justify and validate the approach.

One question we have not yet addressed concerns the nature of the difference between German and English superiority. We have established that the German version exists: why then does it have an apparently weaker violation cost? What distinguishes the two? Perhaps the first thing to point out here is that this difference may be less marked than German linguists tend to assume: the frequency data in Tables 2 and 3 above show that in the 951 million accessible word forms of the COSMAS corpus, we found no single example of a structure with *wen* ... *wer* or *wem* ... *wer*. Nevertheless, *was* ... *wer* was attested, and more often than one might expect for the equivalent structure in English. This does require some explanation.

One possibility is that constraint violation strengths can vary cross-linguistically, but this might be seen as requiring some justification. Our own view is that variation in constraint strength across languages is less problematic than it is sometimes assumed. First, this is not a stipulation we make in order for a syntactic account to go through: the primary data of language presents us with this phenomenon. These wh-constraints simply do exist in both English and German, but in German they do not wholly exclude a structure from being part of the language, which the English variant more nearly does. Second, it is not incompatible with our understanding of mental processes that constraints in language should be gradable. If we assume, as we do here, that syntactic constraints are at base a symptom of computational complexity, it seems fairly natural for this complexity to depend in part upon other features of the language.

Once we have allowed crosslinguistic variation in constraint violation cost, we might apply it to the superiority data set in a number of ways. It might, for example, be the case that the strength of the dispreference for in-situ subjects differs between the two languages: our data shows no sign of this, but as we cannot readily quantify the violation cost strength in such a way that it could be compared across languages, this possibility cannot be ruled out. Another possible reason why the superiority effect in German does not trigger full ungrammaticality might be that multiple wh-questions in English are already "lower down the scale" than they are in German. The additional violation cost of superiority, although of equal amplitude in the two languages, would in this scenario be enough to push superiority violating structures right into the group of structures generally excluded from the language, while in German, this violation cost merely would push violating structures into the zone of markedness.

We noted above that superiority in English seems to have two components: a dispreference for in-situ wh-subjects and a dispreference for clause-initial wh-nonsubjects, and suggested that the absence of this second component in German might be the cause of the perceived difference in strength of the superiority effect in the two languages. This account is tempting because of its simplicity, and it no doubt plays a role in the different perception of superiority in English and German, but we would still need to accept cross-linguistic constraint strength variation for other data sets, such as the *that*-trace data (Featherston 2003b).

To summarize, experimentally obtained judgements such as those presented in this article make it clear that syntactic constraints may be violated, but have an inevitable violation cost in terms of the grammaticality of the structure. This cost is quite consistent and is even quantifiable within a given study. Constraint violations are survivable, that is, a single violation does not necessarily remove the structure from the set of possible structures of the language. This has further implications: if constraint violation costs vary in strength and are cumulative, it follows that grammaticality itself is not binary, but a continuum. Now we shall not here discuss at length whether the idea of absolute grammaticality is an epiphenomenon of perception, as we argue in Featherston (2002), or due to frequency in production (Featherston 2003a), nor shall we outline the decathlon model (Featherston forthcoming), which attempts to situate judgements in a psycholinguistic model. Our aim here is to highlight the implications that these findings have for syntactic theory building.

The conclusions we have been forced into sound like bad news for syntax. The previous simplifying assumption on which much syntax based itself, namely that it dealt only with categoricals, looks to be too flawed to be further employed: syntacticians must deal with yet one more variable, the strength of violation cost. It is undeniable that this additional complication will make syntactic theory more complex and cumbersome. The implications are not all bad, however, since the recognition that constraints are not always absolute may allow a lot of awkward data to be satisfactorily dealt with. Wh-constraints in German need no longer be a problem: they exist, but they are survivable, and so, counterexamples can be tolerated instead of needing to be accounted for with additional categorical sub-rules. We argue in Featherston (2002) that the same applies to parts of the binding theory; again a murky descriptive area whose many twists and apparent exempt anaphors can be accommodated more naturally when survivability is embraced. The abandonment of the simplifying assumption that constraints are absolute should therefore make syntactic theory more empirically adequate, surely a most desirable result. But the acceptance of violability opens a further welcome perspective, too. The realization that constraints are violable and that counterexamples are not forcing evidence may permit a reconsideration of the universality of certain parts of the grammar. For just as the categoricity assumption is applied within languages to argue that the grammar of a language cannot contain a constraint since counterexamples can be found, it can equally be applied between languages in the search for universals. This has made it possible to argue against the universality of movement constraints on the basis of data from German: if German does not have it, it can't be universal. This way other syntactic phenomena, too, apparently missing from other languages, have been seen as ruling these out as universals. The magnitude estimation methodology and constraint survivability allow these assumptions to be tested. It seems quite plausible that constraints thought to be absent from a language will be revealed to be present and measurable in an experimental approach, but to have a small violation cost, so that they have little effect upon the language. It may also prove that constraints which have no usual domain of application in a language may appear if such a domain is artificially created. For example, in a language with no wh-movement we create structures implementing wh-movement of subjects and objects in multiple wh-questions. Speakers of the language will naturally judge these sentences all as bad, but will they judge superiority violations as worse? If they do, then the case for universality is strong. Universals may prove to be much more common than generally assumed, which can only support the claim of syntax to be an approach to the study of the human mind. This, therefore, is the sort of advance we hope and expect from the approach to syntax we have adopted here.

Received 28 August 2001
Revised version received
21 October 2003

Eberhard-Karls-Universität, Tübingen

Notes

* This work was supported by the Deutsche Forschungsgemeinschaft. Thanks are due to Wolfgang Sternefeld, Frank Keller, Roland Meyer, and many other members of the SFB441 in Tübingen for support and advice. Thanks also to Bob Borsley and two anonymous reviewers for helpful comments. All remaining weaknesses are my own. Correspondence address: SFB441 Linguistische Datenstrukturen, Universität Tübingen, Nauklerstr. 35, 72074 Tübingen, Germany. E-mail: sam.featherston@uni-tuebingen.de.

[1] Whether the ideal subject should know the language perfectly is a moot point: too much knowledge might be a dangerous thing, in that it could make our subject untypical. For our purposes, the ideal subject should know the language "perfectly normally."

[2] The CELEX Lexical Database, Release 2, 1995. Centre for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen.

[3] The full set of items can be inspected under http://www.sfb441.uni-tuebingen.de/~sam/db/soup.mazl.html.

[4] The figures are, by in-situ wh-item (mean, standard deviation): wh-subj −0.6887, 0.9798; wh-DO 0.2945, 0.8401; wx-DO 0.1307, 0.9061; wh-IO 0.3947, 0.9810; wx-IO 0.2545, 0.9593; wh-adj −0.1705, 0.7613.

[5] None of the data discussed in this article 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. While the nonsubjects were always normally distributed, the in-situ subject conditions were sometimes not fully normally distributed. Inspection of the data spread reveals that this is due to occasional outlying positive judgements of these conditions, which are generally judged to be bad. In the light of the robustness of the effects observed, this cannot undermine the basic empirical facts from which we argue. In analyses of variance, the Huynh-Feldt correction was applied when appropriate. The interactions and post hoc tests of the superiority data sets had to be carried out with the MANOVA procedure, since the

linguistic data of superiority force an incomplete experimental design (e.g. there can be no cell in the design with a wh-subject superior to a wh-subject).

6. The figures are, by in-situ wh-item and raised wh-item (mean, standard deviation):

   - in-situ wh-subj (raised: wh-DO −0.7646, 0.9327; wx-DO −0.7430, 0.9469; wh-IO −0.9192, 1.1767; wx-IO −0.5711, 0.8219; wh-adj −0.4457, 0.9704);
   - in-situ wh-DO (raised; wh-subj 0.3495, 0.8594; wh-IO 0.5195, 0.7553; wx-IO 0.3784, 0.8818; wh-adj −0.0069, 0.7727);
   - in-situ wx-DO (raised: wh-subj 0.3989, 0.9443; wh-IO 0.3624, 0.7141; wx-IO 0.1097, 0.8838; wh-adj −0.3484, 0.8966);
   - in-situ wh-IO (raised: wh-subj 1.0311, 1.1324; wh-DO 0.5346, 0.8867; wx-DO 0.0046, 0.7925; wh-adj −0.0033, 0.6952);
   - in-situ wx-IO (raised: wh-subj 0.6557, 0.8354; wh-DO 0.2327, 0.7176; wx-DO 0.1832, 0.9343; wh-adj −0.0053, 1.1846);
   - in-situ wh-adj (raised: wh-subj −0.2397, 0.8009; wh-DO 0.0037, 0.7839; wx-DO 0.0073, 0.6847; wh-IO −0.2327, 0.7024; wx-IO −0.4908, 0.7239).

7. The figures are, by in-situ wh-item and raised wh-item (mean, standard deviation):

   - in-situ wh-subj (raised: wh-DO −0.7845, 0.8558; wx-DO −1.1317, 0.7345);
   - in-situ wx-subj (raised: wh-DO 0.3511, 0.7852; wx-DO 0.1037, 0.7547);
   - in-situ wh-DO (raised: wh-subj 0.2967, 0.6334; wx-subj −0.1751, 0.7544);
   - in-situ wx-DO (raised: wh-subj 0.6356, 0.6317; wx-subj 0.7042, 0.5319).

8. Institut für Deutsche Sprache, Mannheim, COSMAS Corpus W-PUB, 951.67 million words accessible [http://corpora.ids-mannheim.de/~cosmas/]. The figures here represent the number of occurences of each type found with a search for the two specified wh-items at up to three words distance between the two. Each ordered pair was searched for twice: both with and without an initial capital W (i.e. *wer was* and *Wer was*), and the results summed. When searches threw up more than 500 hits then samples were counted (normally 25%, but minimum 200 and maximum 500) and the frequencies adjusted proportionately to the total hits. Prepositional arguments were discounted (i), as were occurences not in clauses (ii), occurences where neither of the wh-items in question were in clause-initial position (iii), and examples containing *was* which is not a direct object (iv).

   (i) Was gehört zu wem?
   (ii) Wer gegen wen?
   (iii) Wann hat *wer wem* etwas weggenommen?
   (iv) Wem gehört eigentlich *was* in Europa?

9. A point of interest in this frequency data is the contrast between *was* and *wen*. While there are approximately equal numbers of *wer was* and *wer wen* structures, the 44 *was wer* examples reach 4.6% of the 963 total of the *wer was* examples, but the ordered pair *wen wer* is not attested. This effect is not visible in our judgement data, since there all direct objects were inanimate. We can only speculate here, but let us note that the overall frequency of *was* and *wen* is very different. We investigated their frequency in COSMAS by searching for wh-items with capitals followed by a question mark up to eight word forms later. We did not examine these results in detail, but even raw data here gives a reasonable idea of relative frequency: *Was ...?* occurred 81,200 times and *Wen ...?* just 1,286 times (*Wer ...?* 33,484, *Wem ...?* 1,481). Such a clear frequency contrast between *was* and *wen* (60:1) opens up a range of processing reasons for

constrasts between *was* and *wen*, we would therefore hesitate to attribute this to a syntactic factor. See also Note 15 for equivalent data in English.

10. We use the term "dative shifted" merely as a convenient name for this argument type.
11. The figures are, by in-situ wh-items (mean, standard deviation): wh-S −0.7394, 0.8612; wx-S 0.0072, 0.9046; wh-DO 0.1637, 0.9075; wx-DO 0.2884, 0.7820; wh-IO 0.1154, 0.9719; wx-IO 0.2228, 0.9526.
12. The figures are, by in-situ wh-item and raised wh-item (mean, standard deviation):

    - in-situ wh-S (raised: wh-DO −0.5031, 0.9755; wx-DO −0.8475, 0.7779; wh-IO −0.7829, 0.9173; wx-IO −1.2080, 0.7217; wh-IO to −0.4973, 0.8609; wx-IO to −0.5976, 0.7332);
    - in-situ wx-S (raised: wh-DO −0.3189, 0.7602; wx-DO 0.4643, 0.8774);
    - in-situ wh-DO (raised: wh-S 0.8264, 0.7748; wx-S 0.3560, 0.7348; wh-IO −0.0041, 0.9114; wx-IO −0.6142, 0.7825; wh-IO to 0.2644, 0.8273; wx-IO to 0.1911, 0.8014);
    - in-situ wx-DO (raised: wh-S 0.4659, 0.7589; wx-S 0.7192, 0.7414; wh-IO −0.2679, 0.6490; wx-IO 0.2422, 0.7252; wh-IO to 0.3645, 0.7270; wx-IO to 0.2068, 0.7830);
    - in-situ wh-IO (raised: wh-S 0.8340, 0.9485; wh-DO −0.1437, 0.7684; wx-DO −0.3441, 0.7659);
    - in-situ wx-IO (raised: wh-S 0.7796, 0.9004; wh-DO −0.2115, 0.9372; wx-DO 0.1004, 0.7513).

13. A reviewer asks whether our claim of a dispreference for any raised wh-nonsubject means that there is something problematic about examples such as "What did Kim do next?" The answer is yes, and no. To appreciate this requires an understanding of the two different levels that we distinguish within our model of the grammar and grammaticality. The answer is yes, because this example violates this constraint, and the constraint's weighted violation cost is applied to it. We see this reflected in the relative judgements we are discussing here. If all other factors are held constant, a structure with an initial subject is judged better than one without an initial subject. No, on the other hand, because this constraint is "survivable," which means that its violation does not automatically exclude the violating structure from the language. Whether a candidate structure can be part of the language is decided on the basis of the cumulative effect of all the constraints which it violates. Now there are other constraints which this structure does not violate, notably the constraint which prescribes initial position for wh-items. All other factors again held constant, a structure with a wh-form in initial position is judged better than one with a wh-form in noninitial position. Now, since this latter constraint has a greater violation cost than the former, the form "What did Kim do next?" is preferred to "Kim did what next?" These two constraints are contradictory in a structure containing a object wh-item, but if there were a form which could combine satisfaction of both constraints, it would be preferred to both of these forms, all other things being equal. When carried out under strictly controlled conditions, introspective judgements always produce this pattern of of data (see Keller 2000). Our own "decathlon model" (Featherston forthcoming) attempts to account for the data by dividing the computation into two parts. In the first, the constraint application module, constraint application is blind and exceptionless, constraint violation costs being quantifiable and cumulative. The second, the output selection module, functions competitively, the candidates bearing the least total violation costs being selected, probabilistically, for output. Our relative judgements tap in just to the output of the first module: structures are only better or worse. A different picture is obtained if subjects are asked for binary judgements, since in this case frequency of occurence additionally plays a role: only structures which are "good enough" to win through in the competition for

output are judged to be part of the language, all others are rejected. Further details of the decathlon model, the grammatical model which attempts to account for the different data patterns found in relative grammaticality judgements, binary choice grammaticality judgements, and frequency data are in Featherston (forthcoming).

14. This data leads one to the question whether a syntax can have positive preferences or only negative dispreferences. Since our assumption here is that syntactic regularities are at base the reflexes of computational complexity, we would not exclude the possibility that a grammatical factor could make a structure better. However, this question requires and deserves further study and a more detailed treatment than we can afford here. We are planning to address this in future work.

15. A search in the British National Corpus (BNC, Oxford University, 100 million words) revealed only two examples of *what . . . who* structures. The second example is from a transcription of a public meeting; I think we may assume that the "who's" should be read as "who'd". This might be a slip of the tongue or else an erroneous transcription.

(i) What who meant by what?
(ii) I've been a teacher for twenty years (. . .) and I've never been asked by anybody at the playhouse what play's who's like to see in my school.

This is consistent with our other experimental and corpus findings: the BNC is only one tenth of the size of the COSMAS corpus of German, but nevertheless it produces one main clause and one complement clause example. This is only about half the rate we would predict on the basis of the German frequency data, but then we have noted that the effective strength of the superiority constraint in English seems to be greater. The data is thus consistent with our other findings, but it is plain that such limited data can never be more than suggestive. What is required is much more source data, and our sole current access to this is through web search engines. We make no claim about the representativeness of this data source, noting only that work such as Keller et al. (2002) has shown that search engines can produce results which corelate with judgements even better than data from the BNC does.

A Google search of "what did who" produced 371 hits, of which Google selects 236 as being not merely repeats. Of these, 112 were true superiority violations (we excluded: all linguistics sites, and all non-anglophone sites). A search for "who did who" produced 831 hits, of which Google determines 486 as being not merely repeats. These contained just five genuine superiority violations, three with raised wh-DO and two with raised wh-IO. It is noteworthy that this relative distribution corresponds to the equivalent German frequencies in Table 2, where *was wer* occurs, but *wen wer* and *wem wer* do not. This data too, for what it is worth, is thus consistent with our findings elsewhere.

(iii) Hello . . . Hey, buddy . . . All right . . . Who did WHO play for?
http://www.post-gazette.com/columnists/20000712gene.asp
(iv) Person A: last week who did they play
Person B: who did who play
http://www.ujournal.org/users/brat143/
(v) "So, who did she take?" Said a voice. "Who did who take?" Replied a second voice.
www.egiantess.com/stories/greatescape4.txt
(vi) Who did who send it to and what happened?
http://www.watchforum.com/cgi-bin/wbbs/mf_config.pl?read=4007

(vii) Person A: it does say that whoever lied to both (. . .)
Person B: IN WHAT VERSE and who did who lie 2?
http://forums.extremeoverclocking.com/archive/topic/32682-1.html

Interestingly, the first three of these are direct echo questions, but the last two are not. They do not therefore support the claim that superiority violations are only ever produced as echo questions. When the context sentence is exactly copied in the echo question, it is possible to argue that the structure is not being generated by the producer of the echo question, it is merely quoted as a sound string, and one element of it is questioned. In examples such as (vi) and (vii), there is no identical preceding pattern and so we may conclude that the writers have generated these questions themselves. The presence of superiority violations independently generated would tend to support our view that grammatical constraints are, in principle, survivable.

References

Baltin, Mark; and Collins, Chris (eds.) (2001). *The Handbook of Contemporary Syntactic Theory*. Malden, MA: Blackwell.

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

Behaghel, Otto (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25, 110–142.

Chomsky, Noam (1957). *Syntactic Structures*. The Hague: Mouton.

—(1964). Degrees of grammaticalness. In *The Structure of Language: Readings in the Philosophy of Language*, Jerry Fodor and Jerome Katz (eds.), 384–389. Eaglewood Cliffs, NJ: Prentice-Hall.

—(1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

—(1973). Conditions on transformations. In *A Festschrift for Morris Halle*, Stephen Anderson and Paul Kiparsky (eds.), 232–286. New York: Holt, Reinhart & Winston.

—(1975 [1955]). *Logical Structure of Linguistic Theory*. New York: Plenum Press.

—(1981). *Lectures on Government and Binding: The Pisa Lectures*. Berlin: Mouton de Gruyter.

—(1986a). *Barriers*. Cambridge, MA: MIT Press.

—(1986b). *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.

—(1993). A minimalist program for linguistic theory. In *The View from Building 20*, Ken Hale and Samuel Keyser (eds.), 41–58. Cambridge, MA: MIT Press.

—; and Lasnik, Howard (1977). Filters and control. *Linguistic Inquiry* 8, 425–508.

Cowart Wayne (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgements*. Thousand Oaks, CA: Sage.

Fanselow, Gisbert (2001). Features, θ-roles and free constituent order. *Linguistic Inquiry* 32, 405–437.

Featherston, Sam (2002). Coreferential objects in German: experimental evidence on reflexivity. *Linguistische Berichte* 192, 457–484.

—(2003a). Magnitude estimation in syntax: or Galileo and the telescope. Talk given at Potsdam University, Department of Linguistics, 28th February 2003.

—(2003b). *That*-trace in German. Ms, Tübingen University.

—(2005). Magnitude estimation and what it can do for your syntax: some wh-constraints in German. *Lingua* 115(11), 1525–1550.

—(forthcoming). The decathlon model of empirical syntax. In *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, Marga Reis and Stephan Kepser. Berlin: Mouton de Gruyter.

Garret, Edward (1996). Wh-in-situ and the syntax of distributivity. In *Syntax at Sunset*. UCLA Working Papers in Linguistics, Edward Garret and Felicia Lee (eds.), 129–145. Los Angeles: UCLA Linguistics Department.

Ginzburg, Jonathan; and Sag, Ivan (2000). *Interrogative Investigations: The Form, Meaning and Use of English Interrogatives*. Stanford: CSLI Publications.

Grewendorf, Günther (1988). *Aspekte der deutschen Syntax*. Tübingen: Narr.

—(2001). Multiple wh-fronting. *Linguistic Inquiry* 32(1), 87–122.

Haider, Hubert (1993). *Deutsche Syntax–Generativ*. Tübingen: Narr.

Keller, Frank (2000). Gradience in grammar: experimental and computational aspects of degrees of grammaticality. Unpublished doctoral dissertation, University of Edinburgh.

—; Corley, Martin; Corley, Steffan; Konieczny, Lars; and Todirascu, Amalia (1998). WebExp: a Java toolbox for web-based psychological experiments. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh.

—; Lapata, Maria; and Ourioupina, Olga (2002). Using the web to overcome data sparseness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Jan Hajic and Yuji Matsumoto (eds.), 230–237. Philadelphia: Association for Computational Linguistics.

Kuno, Susumo (1982). The focus of the question and the focus of the answer. In *Papers from the Parasession on Nondeclaratives*, Robinson Schneider (ed.), 134–157. Chicago: Chicago Linguistics Society.

Lakoff, George (1970). *Irregularity in Syntax*. New York: Holt, Reinhart and Winston.

—(1973). Fuzzy grammar and the performance/competence terminology game. *Chicago Linguistics Society* 9, 271–291.

Lasnik, Howard; and Saito, Mamoru (1984). On the nature of proper government. *Linguistic Inquiry* 15, 235–289.

Lenerz, Jürgen (1977). *Zur Abfolge nominaler Satzglieder im Deutschen*. Tübingen: Narr.

Lightner, Theodore (1976). Review of 'Goals of Linguistic Theory' by Stanley Peters. *Language* 52(1), 179–201.

Lodge, Milton (1981). *Magnitude Scaling: Quantitive Measurement of Opinions*. London: Sage.

Lutz, Uli (1996). Some notes on extraction theory. In *On Extraction and Extraposition in German*. Linguistik Aktuell 11. Uli Lutz and Jürgen Pafel (eds.), 1–44. Amsterdam: Benjamins.

Müller, Gereon (1991). Beschränkungen für Wh-in-situ. *Groninger Arbeiten zur Germanistischen Linguistik* 34, 106–154.

Newmeyer, Frederick (1999). Some remarks on the functionalist-formalist controversy in Linguistics. In *Functionalism and Formalism in Linguistics*, Michael Darnell, Edith Moravscik, Michael Noonan, Frederick Newmeyer, and Kathleen Wheatley (eds.), 467–480. Amsterdam: Benjamins.

Pesetsky, David (1987). Wh-in-situ: movement and unselective binding. In *The Representation of (In)Definiteness*, Eric Reuland and Alice ter Meulen (eds.), 98–129. Cambridge, MA: MIT Press.

Pollard, Carl (1996). On head non-movement. In *Discontinuous Constituency*, Natural Lanuage Processing 6, Harry Bunt (ed.), 279–305. Berlin and New York: Mouton de Gruyter.

—; and Sag, Ivan (1994). *Head-driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Prince, Alan; and Smolensky, Paul (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report No. 2, Center for Cognitive Science, Rutgers University.

Ross, John (1972). The category squish: Endstation Hauptwort. *Chicago Linguistics Society* 8, 316–328.

Schütze, Carson (1996). *The Empirical Base of Linguistics: Grammaticality Judgements and Linguistic Methodology*. Chicago: University of Chicago Press.

Sobin, Nick (1990). On the syntax of English echo questions. *Lingua* 81, 141–167.

Stevens, Stanley (ed.) (1975). *Psychophysics: Introduction to its Perceptual, Neural and Social Prospects*. New York: John Wiley.

Uszkoreit, Hans (1987). *Word Order and Constituent Structure in German*. CLSI Lecture Notes No. 8. Stanford, CA: CSLI.

Wiltschko, Martina (1997). D-linking, scrambling and superiority in German. *Groninger Arbeiten zur Germanischen Linguistik* 41, 107–142.