Data in generative grammar: the stick and the carrot

Sam Featherston, Tübingen University [*]

March 27, 2007

Abstract

This paper is intended to lay out for broader discussion some arguments for the importance of data in work in generative syntax. These are accepted by many linguists, but a significant number of others still seem reluctant to accept them. The basic claim is that it is no longer tenable for syntactic theories to be constructed on the evidence of a single person's judgements, and that real progress can only be made when syntacticians begin to think more carefully about the empirical basis of their work and apply the minimum standards we propose. We advance two groups of reasons for syntacticians to do this, negative and positive. The negative 'stick' group concerns the inadequacy of current practice. We argue that linguists are producing unsatisfactory work with these methods. Data quality is a limiting factor: a theory can only ever be as good as its data base. The positive 'carrot' group concerns the descriptive and theoretical advantages which become available with more empirically adequate data. We hope to tempt linguists to adopt new methods by showing them the insights which better data makes available.

1 Introduction

In this article we attempt to explain why we think that research in the field of generative syntax needs to pay much greater attention to its data base. We shall put forward a range of reasons for this, beginning with some illustrations of how current practice is inadequate and failing the field. In short, much work in the generative tradition of grammar is fatally undermined by its oversimplified assumptions about the patterns that the data which it is attempting to model actually exhibits. This situation has partly come about because syntacticians are very reluctant to examine the data in any detail, but have come to regard it as unrewarding.

These criticisms are the 'stick': the unpleasant realities about the current situation. But we shall also wave the 'carrot' in front of readers, attempting to show with positive examples how increased use of data in argumention, and increased attention to the quality and evidential value of data can in fact provide

[*] This work took place within the project Suboptimal Syntactic Structures of the SFB441 *Linguistic Data Structures*, supported by the Deutsche Forschungsgemeinschaft. Project leader is Wolfgang Sternefeld. Thanks are due to Gisbert Fanselow for comment. The responsibility for the content rests with me alone.

rich descriptive and theoretical rewards. Our aim is not to point the finger at anyone in particular or criticize past or indeed current practice any more than is necessary, but rather to lay out very clear and achievable suggestions for improvement, and show the benefits that this increased attention to data offers: as far as possible we wish to encourage, not find fault. With every paradigm shift there are dearly-held positions which have to be abandoned, which is naturally unwelcome for those who hold them. This makes progress difficult, and it sometimes appears that the less evidence there is for a tenet, the more passionately it is held. We have a simple suggestion to smooth syntax's passage to empirical adequacy: minimum standards that we think any published work using judgements should adhere to.

These improved data standards do more than just eliminate the current inadequate use of data. They also yield much higher quality data, in that they provide more detail and information. To underline the advantages, we also provide some example results which we have obtained. These 'carrot' example studies reveal the rich pattern of differences between structures which provides evidence about the workings of the grammar and the nature of well-formedness. They are the result of tightly controlled experiments using judgements, but we would emphasize that we are not demanding that every syntactician carry out elaborate experiments all the time. Less time-consuming ways of gathering data are also valid and useful, but the detail in the results is proportional to the effort put into obtaining them. Linguists who use very informally obtained data must expect to be contradicted by others using more reliable and more detailed information. This ability to provide firm answers to questions is an important aspect of the 'carrot'.

Linguists have read critiques of their use of judgements before (above all Schütze 1996), to such an extent that it becomes difficult to say something new, but nevertheless this unrewarding job must be done, because the field of syntax is still continuing its inadequate practice. Although almost nobody publicly disagrees with the need for paradigm shift, the non-compliance of a significant number of colleagues makes additional persuasive articles such as this one necessary. It gives us no pleasure to find fault with colleagues, but it is our view that this immobilism is undermining the reputation of syntax and the respect that the field of syntax should enjoy in the wider academic sphere. The work of many syntacticians is entirely contained within the generative world, which makes this weakness less visible, but for those researchers whose activities overlap with other related fields, the weak empirical basis of much work in the generative paradigm is very apparent. Perhaps it is for this reason that generative linguistics has become self-referential and inward-looking. Newmeyer (1983) has a nice quote from Tom Roeper (1982, 468) "when psychological evidence has failed to conform to linguistic theory, psychologists have concluded that linguistic theory was wrong, while linguists have concluded that psychological theory was irrelevant." This is disappointing, for it was precisely the potential to provide a theory of language structure of wider application that made the generative programme so exciting in its early stages. Linguistic theory has all but abandoned its ambition to be a part of psychology, and there is no acceptance of generative grammars among psychologists, (or indeed other branches of linguistics: sociolinguistics, psycholinguistics, . . .). Generative linguists should ask themselves why.

In this paper we shall discuss only judgement data, since this is the traditional syntactician's data type of choice. The intended audience of this paper chiefly uses introspective judgements as their criterion of what structures are part of the language and what structures are not. This focus is in no way intended to belittle the value of corpus data or make out that this data type is any less relevant. There are of course still questions about the evidential status of frequency information: neither occurrence nor frequency are identical with well-formedness, but the working assumption that only those structures which are well-formed will normally occur has considerable validity. There are still questions, but it seems to us that linguists using corpus frequencies are more aware of these data issues and are solving them on their own.

Other data types such as the evidence of on-line processing are also increasingly being gathered and analysed for their implications for the representation and processing of language. This data naturally requires additional analytical steps if we wish to distinguish the encoding of the language system and its application in real time, a differentiation which is admittedly sometimes called into question. Nevertheless, the time sensitivity and event-related responses of the data type can offer valuable insights into speakers' expectations and incremental understanding (for a recent positive example see Bader & Bayer 2006).

But our aim here is to show generative linguists that paying more attention to data does not require them to throw all their acquired knowledge and expertise out the window and start again. The traditional data type for generative grammar is judgements and the mainstream generative grammarian still uses them. Generative grammar was not designed to model occurrence frequencies and they do require an inferential conversion process before one can make strong conclusions from them about well-formedness as understood in the generative framework. This additional step requires further work (see Featherston 2005 for an attempt). We shall therefore confine our remarks to judgements.

Fortunately we do not have to survey all previous related publications, since Carson Schütze has done this in his excellent work (1996). It is also well worthwhile rereading Labov (1975, 1996) and Greenbaum (1977), the critics of judgements as used in generative grammar, and Newmeyer (1983, 2003), the conservative supporter of generative practice. We agree, perhaps surprisingly, with most of what all four of these authors have said, and disagree, naturally, with parts of all of them.

2 Failings and inadequacies: the stick

The most fundamental part of the problem is that a significant number of linguists are still, in spite of all the warnings to the contrary, using as the basis of their work what we might call *linguist's judgements*. In the worse case these are introspective judgements by given linguists themselves as the data base of their own theoretical work, on the basis of a single example sentence, not checked against the intuitions of other independent informants, and often idealized to a dichotomy of good or bad. Note that this term is a mere shorthand for all the bad practice in using judgements as an evidence type – we would not claim that judgements by linguists are in themselves invalid, as far as they go.[1]

[1] We have tested this in an experiment in which we gathered well-formedness judgements. Half the participants were at least graduate linguistics students and the other half had no such background. There was no difference between the results of the two groups.

We have said that we do not wish to single out any particular linguists, for the practice that we are criticizing is not specific to individuals but rather to the field. Nevertheless, readers of draft versions of this text have demanded examples to make it clear precisely what the problem is. Here therefore are two examples of what can go wrong. We hope that we have made it abundantly clear that no personal criticism is intended, indeed we partly choose this particular linguist because he neither needs our praise nor need fear our censure.

2.1 The case of object coreference

Let us quote from Grewendorf (1988, 58) about example (1).

> Die Generalisierung [...] läßt sich für jemanden, der ein Gespür für subtile, aber nichtsdestoweniger eindeutige Grammatikalitätsunterschiede hat, an der folgenden Gegenüberstellung noch einmal illustrieren. (‘The generalization can be illustrated once again in the following contrast, for someone who has a feeling for subtle, but nevertheless unequivocal grammaticality differences.’)

(1) a. Der Arzt zeigte den Patienten$_j$ sich$_j$/*ihn$_j$ im Spiegel
The doctor showed the patient.*acc* himself/him in.the mirror
b. Der Arzt zeigte dem Patienten$_j$ ihn$_j$/*sich$_j$ im Spiegel
The doctor showed the patient.*dat* him/himself in.the mirror

Three years earlier he commented on the same data set (1985, 160):

> Selbst wenn die entsprechenden Beispiele semantisch bisweilen etwas schwer nachzuvollziehen sind, so scheinen mir die syntaktischen Intuitionen bzgl. der festgestellten Regularitäten doch relativ klar zu sein. (‘Even if the relevant examples are sometimes somewhat difficult to follow semantically, the syntactic intuitions of the regularities established seem to me to be relatively clear.’)[2]

We can therefore be sure that Grewendorf is entirely sure about his intuitions of these cases. From this data he concludes that this data supports an implicational hierarchy of anaphoric binding, in which, in this example, direct objects can bind indirect objects, but not the reverse (see also Pollard & Sag 1994, Primus 1987). We have tested precisely these structures in a fully controlled experiment gathering the introspective judgements of twenty-six informants of sixteen syntactic and and eight lexical variants, using the magnitude estimation methodology (Bard et al 1996, see also Cowart 1997, Keller et al 1998, Keller 2000).[3] The results are illustrated in Figure 1. In this graph the vertical scale shows normalized judgements, with higher scores indicating ‘better’

[2] Note that there is what we must take as a typo in the 1985 version of the examples. It should plainly read *Der Arzt zeigte den Patienten$_j$ sich$_j$*/*ihm$_j$ *im Spiegel*, not *ihn*.

[3] Using this procedure for gathering judgements under controlled conditions, speakers are able to distinguish and express multiple levels of well-formedness. We ask informants for their judgements of example structures in numerical form, relative to a reference item and to their own previous judgements. The task thus has roughly the following form: ‘You gave this structure a 10 and this one, which is twice as good, a 20. On the same scale, what score would you give this one?’. Each informant thus creates their own scale as they give additional judgements; this permits them to express all the well-formedness differences they perceive.

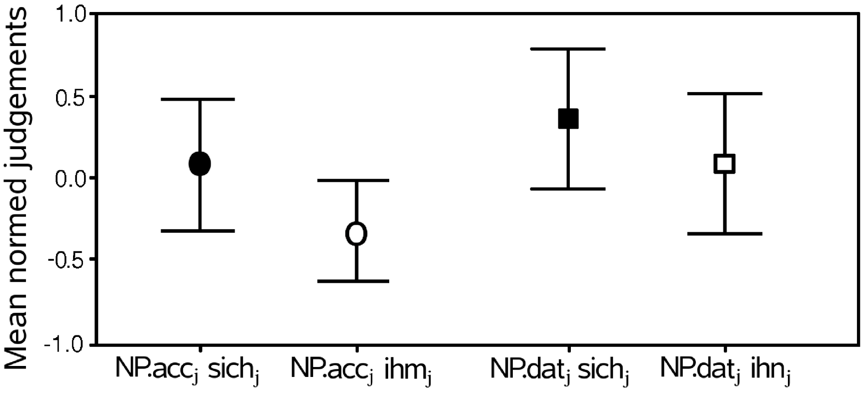

Figure 1: Experimental findings demonstrate that object coreference structures with reflexives as anaphoric elements are perceived more well-formed than with pronouns, whether the antecedent is accusative or dative.

judgements. The syntactic conditions tested are arranged along the horizontal scale. The judgements assigned to each condition are shown with an error bar. The symbol at the centre of the error bar shows the mean judgement for the condition, and the length of the bar shows the 95% confidence interval for the mean. Notice that, for clarity, we include here only the conditions which precisely apply to Grewendorf's examples, but the other conditions confirmed this finding robustly (Featherston 2002 for details).

The results do not confirm Grewendorf's intuitions, on the contrary, they show that the reflexive is judged better than the pronoun with both dative antecedents and accusative antecedents. The results also show a mild preference for antecedents to be dative, not accusative, but this is no doubt due to the independent preference for datives to linearly precede accusatives in the mittelfeld (eg Uszkoreit 1987). There is no visible effect which would support a hierarchy of binding relations.[4]

This then is a clear example of a linguist doing what Newmeyer (1983) says that linguists do, namely trusting his own intuitions more than any other source of data. However, as Schütze points out, linguists can and do err, influenced by a range of factors, including but not limited to the demands of the paper they are working on. The judgements of an individual are revealed to be inadequate as a basis for theory development. Finer data is required in part because of the number of effects which influence judgements in these examples (see Featherston 2002) and because some of these effects are relatively modest, perhaps smaller than the random 'noise' factor in the individual's judgements.

Here is another example, the well-known *that*-trace effect. In English, while both subjects and objects can be equally well extracted from a complementizerless complement clause, objects but not subjects can be extracted from complement clauses with a complementizer. The equivalent structures in German are illustrated in (2). The *that*-trace effect would predict that examples

[4] To ensure that informants get the coreference reading in every case, it is necessary to add emphatic *selbst* to the anaphoric elements. Corpus studies show that this is almost always present in naturalistic production.

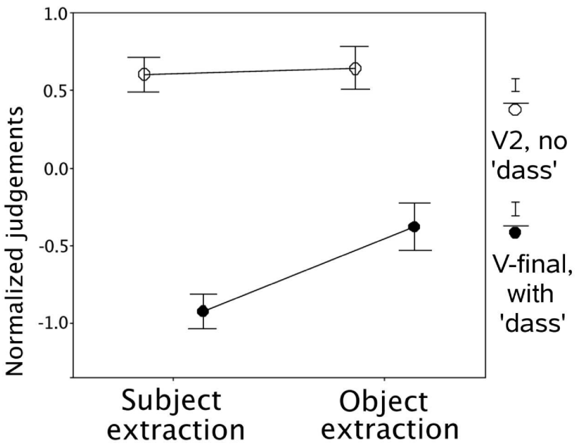

Figure 2: Our study revealed that German exhibits behaviour very similar to the *that*-trace effect well-known from English.

(2a), (2b) and (2c) should all be better than (2d). However most authors have taken the view that standard German has no *that*-trace effect (Haider 1983, Grewendorf 1988, Stechow & Sternefeld 1989, Bayer 1990, Haider 1993, Lutz 1996). Differentiating more finely, Grewendorf (1995) suggests that North German, but not standard German is sensitive to the contrast, and Fanselow (1987) claims that south German, but not standard German is sensitive to it.

(2) a. Wen$_i$ meint Lydia, liebt Jakob t$_i$?
whom thinks Lydia loves Jakob t$_i$
'Who does Lydia think Jakob loves?'
b. Wer$_i$ meint Lydia, liebt t$_i$ Jakob?
who thinks Lydia loves t$_i$ Jakob
'Who does Lydia think loves Jakob?'
c. Wen$_i$ meint Lydia, dass Jakob t$_i$ liebt?
whom thinks Lydia that Jakob t$_i$ loves
'Who does Lydia think that Jakob loves?'
d. Wer$_i$ meint Lydia, dass t$_i$ Jakob liebt?
who thinks Lydia that t$_i$ Jakob loves
'Who does Lydia think that loves Jakob?'

We tested this question, gathering judgements from twenty-five informants (see Featherston 2003 for details) who saw a total of thirty-two different conditions in eight lexical variants. The results are illustrated in Figure 2, where as before higher scores indicate better judgements. We observe that all extractions from complement clauses with a complementizer are assigned worse scores than extractions from complement clauses without a complementizer, but that the subject extractions in just this case are worse still. This finding clearly indicates that there is a *that*-trace effect in German, since the pattern here is very similar to the effects known from similar studies of English (Cowart 1997).

This syntactic phenomenon is thus another case where the judgements of the individual linguists seem not to have been sufficiently sharp to determine the full picture. That linguists disagree (with Grewendorf this time on the more accurate side) on what should be a simple empirical question must alone be regarded as a cause for concern in the quality of the data. That the majority of them should turn out to be giving the wrong answer shows that the individual linguist's judgements can mislead them.

Here therefore we have two examples of the way that insufficient attention to data can lead syntactic theory astray. It is this unfortunate and in the long term unhelpful situation for theory that we seek to see remedied. Notice however that we do not wish to condemn introspective judgements as a data type. The position that we are adopting here is that judgements are indeed a fully valid way of making generalizations about syntax, but they must be used with more care and paid more attention to. In the next section we lay out what the problems are and what can be done about them.

2.2 The inappropriate use of judgement data

An individual judgement or set of judgements is a subjective phenomenon. The loss of independence involved when a linguist uses only their own judgements as the data base of theory development is surely a clear enough reason in itself to avoid doing this. It is of course entirely reasonable for a researcher in whatever field to both gather data and interpret it as long as the data refers to some external phenomenon, but it is not good scientific practice to apply this to subjective data such as ones own introspection. This requirement is familar in the academic world: theses are read by external examiners, abstracts are read by two or three independent reviewers - the fundamental principle of peer review is well established. It is simply inadequate research practice for linguists to rely on their own unconfirmed introspective judgements as linguistic evidence.

It is also insufficient to take judgements uncritically from the literature and build upon them, since the literature is full of examples of dubious practice. I have on my desk a paper by an author with a PhD from MIT and a job at a prestigious university. The main point of the article rests upon the interpretation of two structures (both in English and German). But the author is not a native speaker of either of these languages and these critical readings are simply unavailable to me or anyone I have asked. Some related structures do show the intended readings, as do the original examples given in the first paper on the issue 25 years ago, but the author has simply not checked the facts. This is not a new problem. Greenbaum (1977) commments 'All too often the data in linguistics books and articles are dubious, which in turn casts doubt on the analyses. Since analyses usually build on the results of previous analyses, the effect of one set of dubious data left unquestioned can have far-reaching repercussions that are not easily perceived.' Practice in syntax has not changed in thirty years, it would seem.

An equally unsatisfactory data source are isolated examples from obscure, little-studied languages. Sometimes one reads an article suggesting a syntactic analysis, working it through in a familiar language and finishing off by advancing 'confirmatory' data from a language which the reader has never even heard of. Such data is uncheckable, the original data source was probably superficial, and the author themself probably knows no more of the language than exactly this point which they have taken from a descriptive grammar. This sort of data is no support at all.

It is strange that, in a field whose aim is to model the behaviour of a set of syntactic structures, investigation of this set is not valued, and the details of gathering and analyzing data are regarded as uninteresting, almost taboo. Linguists exhibit symptoms of denial of painful reality: if a syntactician reads a paper containing an empirically dubious claim, for example that a marginal structure S is not grammatical, they do not publicly question the judgement. Instead they write a reply, assuming that the structure S *is* grammatical.

If this contradiction were pointed out to them, both linguists would claim to be modelling their just own idiolect. This effectively deflects criticism, but it has the disadvantage that it removes the object to be modelled from the set of objective phenomena, since a single person's judgements are necessarily subjective. The subjectivity of individual judgements is unproblematic as long as we allow that each person's judgements are merely a sample of the judgements of the whole speech community, implying that they are generalizable (Labov 1975's Consensus Principle). But the 'my idiolect' gambit denies generalizability, and asserts that the data can never be other than subjective. This entails a sad reduction in the breadth of the field of syntax, from language, an important human phenomenon, to a single person's unreplicable feelings about language.

The 'my idiolect' definition of the object of study thus buys invulnerability to questions about its data base at an enormous price. It excludes the possibility that there is a universal grammar. It limits the applicability of any analysis to an individual. It is also demonstrably wrong, since studies gathering judgements from groups show unambiguously that judgements of well-formedness are shared across speakers with only limited systematic variation. It has, in fact, very little in favour of it, other than that it allows its proponent to duck out of a discussion of data issues (see also discussion in Labov 1975).

The origin of this disdain for evidence is probably Chomsky's (1965) suggestion that what was needed was not so much a sharpening of the data base as bold analytical steps. That was a reasonable position to take at the time, given that the generative program was making possible a wide-ranging reinterpretation of existing data. The sixties saw exciting new perspectives appear at an astonishing rate, and if good analytical progress is being made on the basis of the existing data set, then new data is indeed a low priority. Syntactic theory finds itself in a very different situation now: there is little feeling of advance or questions being definitively answered. The paper with erroneous data I mentioned above addresses an issue which was first raised in 1979, and which has not, apparently, been settled in over a quarter of a century. This is not really surprising, when one considers the mediocre quality of data which is used as the basis for analysis. It is evident that the data base underdetermines the phenomenon, and that the way forward is to improve the data base. The excitement and sense of progress in grammar nowadays is to be found in work making use of the new qualities and quantities of data which have become available.

Another significant problem has been the use of inconsistent and uncontrolled idealization. There are structures which seem clearly well-formed, and structures which seem clearly ill-formed. It would appear reasonable to idealize this opposition to a simple dichotomy and build a first draft of a theory on these clear cases. In fact, as an approach to dealing with the fuzziness in judgements this has much to recommend it. But when we idealize, we must bear in mind that we are setting aside a part of the evidence that the data contains. Idealization takes a liberty with the data, and this step requires a commitment from the idealizer to return to the full data set later on and account for why the information removed was irrelevant. Idealization is a temporary simplifying assumption, not a structure-preserving transformation. In particular, it does not follow from idealization that any new data sources are in some way less valid if they do not resemble the idealized data. The idea that experimentally obtained judgement data is somehow undesirable because it does not reveal a clear dichotomy is an example of theoretical work becoming so remote from the primary data that it no longer remembers what the data actually shows.

Note that we do not condemn idealization out of hand, in fact some idealizations have served syntax fairly well, especially the 'ideal speaker-listener' idealization, which divides narrow syntax off from sociolinguistics and processing. Even the idealization of well-formedness to a dichotomy, which divides the 'clear cases' of grammatical structures from the 'clear cases' of ungrammatical structures, is a reasonable abstraction from the basic data pattern, allowing certain factors to be held constant and thus permit analyses of wider generality to be aimed at, as an interim step.

But there is a serious problem of consistency in the use of the idealization to binarity in well-formedness. We must decide whether we are applying this idealization or not, and stick to it. In his early work, Chomsky showed that he was well aware that grammaticality is 'a matter of degree' (1965, 11; see also 1957:36; 1964 passim) but he idealized this to a binary division and developed, entirely reasonably, a model whose aim is 'to separate the grammatical sequences [...] from the ungrammatical sequences' (1957:13).

When we use a model of grammar which either generates or does not generate structures, we implicitly adopt this idealization of the data to a binary opposition, because the model has no place for other data. But if we take this theoretical option, we should, indeed we must, to be consistent, assign any and every example sentence either an asterisk (='ungrammatical') or nothing (='grammatical'). The binary model we have adopted allows no other options. But linguists do not do this; they frequently use multiple intermediate judgements at the same time as assuming a model of grammar that has categoricity as one of its enabling assumptions. It is methodologically unsound to apply such a grammar to a data set which is assigned more than two values. One might add that with such a model, syntacticians should only use evidence from the example structures which are either *clearly* acceptable or *clearly* unacceptable; for this model was explicitly designed to deal only with 'clear cases'.

It is also worth noting here that the idealization of the status of example structures to a binary opposition also entailed the idealization of the nature of grammatical constraints, reducing the set of possible 'grammatical' constraints to just those whose violation causes full ungrammaticality. Few linguists who do not work with data seem aware that this too is an idealization, dependent on the first. Those who doubt that stronger and weaker grammatical constraints exist should reread *Barriers* (Chomsky 1986).

3 Better data

In the sections above we have outlined the disadvantages of (the worst of) current practice. That was the 'stick'. This section outlines what linguists need to do in order to avoid the stick of criticism. One of our aims is to reassure doubtful linguists that quite few, common-sense steps will make a major difference. We therefore first lay out the utter minimum that linguists should do when using judgements as the basis of a linguistic argument; we then point out what additional steps would be desirable. In the next section we supply the 'carrot' and reveal the additional benefits of gathering improved data and paying attention to it. The increased effort that is necessary to gather more valid data is rewarded with more detail in the data. The additional information in turn provides more insight into the grammar.

We might summarize the essential requirement of better data in one word: control. This factor is what makes evidence gathered experimentally more valuable than any single sample. Experimental control excludes many irrelevant factors from affecting the results, so that one can be more certain that the effects one finds are due to the factors that one is interested in. An ex-colleague used to say to me that every single event of introspection was an experiment: in a way it is, but it is an extremely *poorly controlled* experiment.

There are various ways of obtaining controlled judgements, but what they have in common is much larger than what distinguishes them. The essential requirements which make for controlled elicitation are in *The Essentials*, the desirable features are in *The Desirables.* The requirements in the Essentials group are necessary for judgement data to be regarded as valid evidence at all; the requirements in the Desirables group improve the evidence further.

(3) *The Essentials*

a. multiple informants
b. multiple lexical variants of the structures

(4) *The Desirables*

a. task: responding to input
b. scale: multiple degrees of well-formedness

We shall discuss each of these requirements in turn, but let us be clear about what we are proposing. Before making any claim in published work about the pattern of well-formedness in any set of structures, linguists should gather judgements using a method which fulfills at least these minimal requirements, and publish brief details in the paper. That is all. If in a particular case a author should feel that this step is not necessary, perhaps because someone else has reported this work or because the relevant differences are so clear as to be out of any doubt, then they should briefly explain why. It is standard practice in the academic world for sources to be identified; there is no reason for introspective judgements to be made an exception to this. Reviewers should make a point of insisting upon details of judgement data quality. No author will thereby be forced to gather controlled data. But they will be forced to admit that they do not do so. Once data quality is made explicit, analyses based on weak data will, correctly, be valued less highly. There will be an incentive to base work on firmer empirical footings.

We should be clear that the Essentials are *minimum* standards. Anyone familiar with methodological practice in the field of psychology, where judgement data is also used, will be aware that these standards are very loose indeed; even within linguistics, these requirements are modest. In one of the founding works on experimental judgements in syntax, Cowart (1997) includes a sample questionnaire in an appendix. Before they take part in the experiment, informants are asked a total of thirteen personal questions, including what the highest level of educational achievement in their immediate family is, whether any members of their immediate family are left-handed, where they started primary education, whether they moved house during their primary education, and, if so, in which year of school. In comparison with this degree of control, even the requirements in the Desirables group (4) can be seen to be very limited.

There are of course additional factors which would make the data more controlled and thus more valid still. An example is random order of presentation. To avoid any effects of priming by immediately preceding context, sets of structures which are to be compared should, ideally, be presented for judgement in random order. It is not more difficult and it raises the evidential weight of the data by excluding possible irrelevant effects. The plausibility of example sentences is another very important factor; all examples should be as neutral as possible. But syntacticians should be reassured that just observing our minimum standards will bring about a decisive improvement in the validity and detail of their judgement data.

3.1 Multiple informants

This is the most important of all the steps to take. Using multiple informants assures the independence of the data. Ideally we should gather the judgements of twenty or thirty speakers, but as few as ten or twelve will suffice for some purposes. If the responses of a group of a dozen independent informants produce a given pattern, then this cannot simply be the wishful thinking of the author. It is also the case that the judgement patterns of a group of informants are replicable: we have an objective phenomenon to investigate. It is important that syntacticians should show scientific detachment and be seen to be producing theory to fit the facts, rather than the opposite. This cannot be achieved by using our own unconfirmed intuitions. A study which uses data from multiple informants is thus more intellectually rigorous than one which does not.

So much for the stick. But there are positive benefits for the linguist too. Judgements are fundamentally noisy, and show some variability both between informants and across judging events by the same person. It is likely that the second variability explains the first to some extent, that is, that a difference in judgements between two informants is at least partly the result of each individual's judgements being subject to random error. But whether or not this is the case, both these sorts of variability can be evened out, if we obtain multiple independent judgements. The errors cancel each other out and the judgements cluster around a mean, which we can take to be the 'underlying' value, free of the noise factor. Multiple informants thus deliver more accurate data.

An objection to this is that we will be amalgamating 'different grammars' (eg Newmeyer 1983). This complaint is frequently advanced but we will lay out here the reasons why we discount it (see also Labov 1975 and citations there). First and foremost, the data does not support it. If there existed among the speakers of a language grammatical sub-varieties, then we should see evidence of them in our carefully constructed experiments. But we never see this sort of result: the judgements of a given structure always cluster more or less normally distributed around a mean point. The prediction of different grammars would be that we should see separate clusters for the informants for whom the structure is good and for those for whom it is bad. But this does not occur: all we ever see is variation around a common pattern, plus occasional outliers.

In fact experimentally obtained judgements clearly show that systematic differences in judgements across groups of speakers are small and rarely of any significance (even known dialect tendencies are quite modest). The erroneous belief that there are large differences, unmotivated by geographic or social variables, has in part been caused by the assumption that we can estimate the amount of variation between speakers by looking at differences in judgements between individuals. But the real appropriate measure is between any single informant's judgements and the *mean values of a group of informants.* Each individual judgement is noisy, but since errors cancel each other out, these mean judgements effectively remove this error variance. Comparing just any two individuals' judgements on the other hand magnifies error variance because each individual brings their own noise to the comparison, and their variance in each judgement may be in opposite directions. This makes the differences and inconsistencies seem considerable; but the fallacy lies in the assumption that any person's judgements on any occasion are free of noise, and that therefore any instance of a difference in judgements between two individuals is thus evidence of a difference in their grammars. But most of the difference between individuals is just error variance. We can control this by testing groups, and then we see that groups of informants agree quite closely.

A further reason for the assumption of different grammars in the first place was the assumption of a binary model of well-formedness. When linguists were assuming well-formedness to be a dichotomy, they found that, given only two options, some informants put a marginal structure into the 'good' group while others put it into the 'bad' group. This needed an explanation, and so the idea of different grammars was born (see Labov 1975 for background). The assumption that a difference between two speakers means that they have different grammars is thus dependent on the idealization of the data to categoricity. But as we shall see in more detail below, categoricity is an abstraction from the data, not represented in the data itself. If we show informants the colour grey, and ask them to categorize it as black or white, they will show much variation, but this has no implications for their perception of colour or for the nature of the pigmentation. If we permit them to categorize grey as 'grey', they do. We then no longer need the 'different grammars' explanation, and the apparent problem of amalgamating different grammars disappears.

Our final reason for using multiple informants is that only this data is reliable. When we gather the judgements of informants, and examine the overall pattern produced by the group and the patterns produced by each individual, this becomes clear. The group produces a clear, statistically significant pattern revealing a syntactic generalization, but the judgements of each individual informant are noisy and much less visibly systematic. As an example, let us look at Figure 3 which contains the results of a judgement study of multiple wh-questions in German. As before, better judgements are shown in the charts as higher scores, worse judgements as lower scores. The ratings for each syntactic condition are shown in error bars, whose centre point is the mean value and whose length represents the 95% confidence interval for this mean value, and is thus a measure of the amplitude of variation around the mean.

Figure 3 shows mean normalized judgements of thirty-seven informants on multiple wh-questions in German, with different grammatical functions as in-situ wh-elements and raised wh-elements. All of these are possible structures of

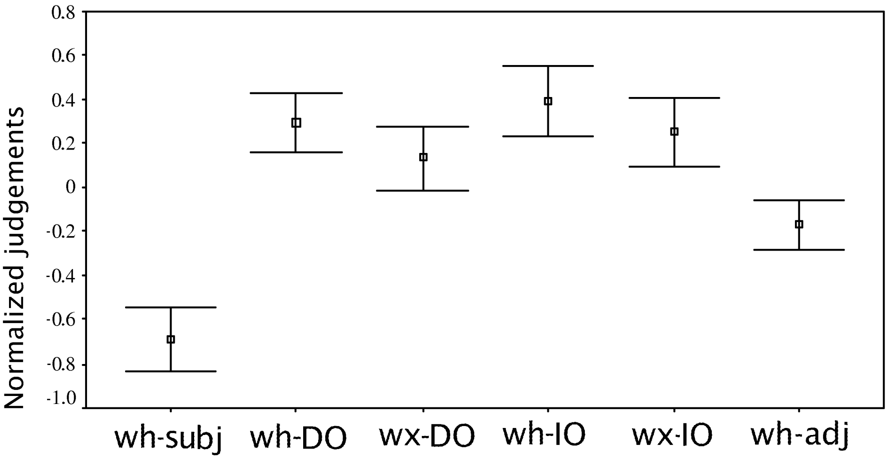

Figure 3: Judgements of multiple wh-questions in German. The six different syntactic conditions relate to the syntactic function of the in-situ wh-item: wh-subject, simple wh-item direct object, *which X* direct object, simple wh-item indirect object, *which X* indirect object, adjunct wh-item.

German, but a little marked unless used as echo questions. The results reveal two clear linguistic facts: first, that structures with in-situ wh-subjects (the *wh-subj*) are worse than all other conditions; second, that structures with in-situ wh-adjuncts (the *wh-adj*) are next worse (for detail Featherston 2004).[5]

Could the treatment as a group being obscuring different grammars? Even the group results must make this unlikely: no clear pattern can appear in the results of a group unless there is substantial agreement among them. The statistical tests demonstrate that the results are not just chance. We also tested for differences by age, sex, and handedness and found nothing, but the idea of the grammatical idiolect is not related to these variables. Did any individuals produce a different pattern? To allow the reader to judge this we have reproduced the results of the first sixteen informants in Figure 4. This graphic shows charts of the individuals' mean normalized judgements by in-situ wh-items, as in Figure 3, but here we use solid bars to show the scores, not error bars, because of the small scale. The bars originate at the zero point, which is the informant's own mean score, so bars hanging from the zero point indicate scores lower than the mean, those standing on the zero point show scores above the mean.

These sixteen informants' individual results are typical in that they show a general trend to reflect the group result in scoring the in-situ wh-subjects clearly worse (see Figure 3), but they also reveal the degree of noise contained in a single person's judgements. Most informants rate the in-situ wh-subjects (the left-most condition) worst or equal worst, in line with the group result, but two do not (numbers 2 and 16). The results of these two informants are still related to the group result however; Number 2 rates the wh-subjects second worst, while number 16 rates them third worst, but the differences between the

[5] In pairwise Tukey HSD tests, the in-situ wh-subjects (all $p<0.001$) and the in-situ wh-adjuncts (all $p<0.03$) were shown to differ significantly from all other conditions, while these others did not differ from each other (all $p>0.1$).

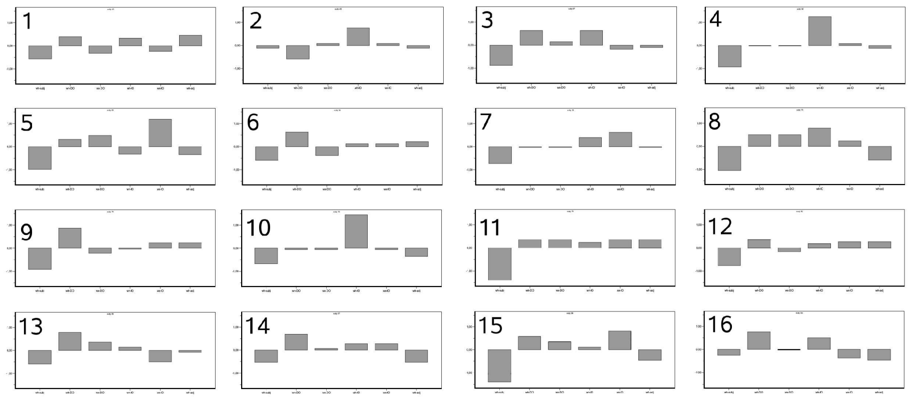

Figure 4: This graph shows the individual results of the first sixteen participants in our experiment on multiple wh-questions in German. The six conditions are as before the mean scores by syntactic function of the in-situ wh-item: wh-subject, simple wh-item direct object, *which X* direct object, simple wh-item indirect object, *which X* indirect object, adjunct wh-item.

three worst scores in this chart are quite small. Everybody rates the in-situ wh-subjects as being worse than average.

But, quite normally, these results also contain a lot of error variance. Informants 2, 4, 8, and 10 score the in-situ wh-item indirect objects (third condition from right) particularly high, while informants 5, 7, and 15 prefer the in-situ *which X* phrasal indirect objects (second from right), and informants 6, 9, 12, 13, 14, and 16 rate the wh-objects (second from left) highest. Informants 11 and 12 distinguish almost only the wh-subjects and leave the rest flat.

Could this variation reflect the idiolects of the informants? The evidence speaks against this. First, there is no alternative pattern to the group pattern. A competing rule-based sub-variety should show a competing pattern of judgements, but what we see just fairly random fluctuation around the group pattern. Second, there is noise even within a single person's judgements. If we ask the same person to judge the same structure more than once, their responses vary, that is, they do not consistently give us the same answer. This demonstrates most clearly of all that the differences between informants are not systematic, reflecting different grammars, but merely noise inherent in the process of judging.

In Figure 5 we see all 26 judgements of just the first four participants in our experiment on German multiple wh-questions in yet more detail. The six groups of conditions on the x axis correspond to the in-situ wh-items as before, but the bars distinguish these scores by raised (that is: clause-initial) wh-item type (this is the same set of wh-items in the same order, but of course no syntactic function occurs in both places at once.) Since the full experiment found very little effect of raised wh-item, the bars in each group represent judgements of almost identical conditions. The reason we focus in on the individual judgements is that these reveal clearly that the variability which we find *between* informants is also robustly visible *within* the judgements of a single informant.

Let us look at the way that these four treat just the in-situ wh-subjects

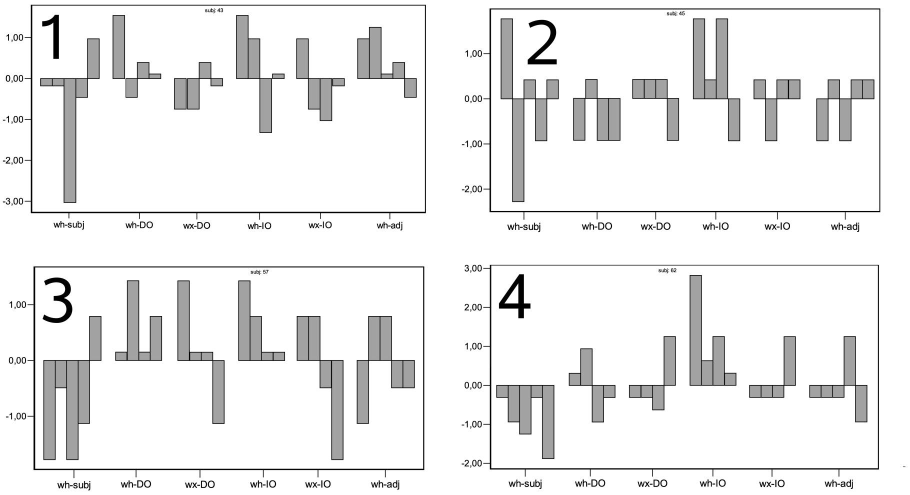

Figure 5: This graph shows the full set of 26 judgements of the first four informants in our experiment on multiple wh-questions in German. The six groups of conditions are as before the syntactic function of the in-situ wh-item: subject, wh-item direct object, *which X* direct object, wh-item indirect object, *which X* indirect object, adjunct wh-item. The bars within the groups distinguish the raised wh-item (the same set in the same order as the in-situ wh-items).

(the most left-hand group of five bars). All four of them assign their (equal) worst judgement to a condition in this group and generally they treat all the wh-subjects worse than the other groups (see Figure 4 for means). But each informant gives both more and less negative judgements for the five conditions in the group and there is no sign that there is any consistent pattern to this variation. There is thus no possible linguistic basis for such variation.

If informants produce noisy data, should we distrust their judgements? I argue that we should trust their data all the more. First, because we can only see the noise in the judgements when we have a norm, generated by a group, in comparison to which the individual's error can be identified. Second, because all the evidence suggests that linguists' own judgements are no less prone to error variance than anyone else's. The two examples we offered in the 'stick' section above and the cases reviewed in Schütze (1996) make it entirely clear that syntacticians are no different from other informants in their ability to provide well-formedness judgements. Nothing else is to be expected, since introspective judgements are orthogonal to conscious linguistic analysis (but see section 3.4 below of amplification of this). It follows that the only reliable source of perceived well-formedness evidence is the judgements of a group, in which the slips and noise cancel each other out.

When confronted with facts like these, researchers on syntax tend to resist, apparently fearing that this is a threat to theory. This fear is unjustified, however. Far from being a threat to theory, this is good news. Essentially this evidence is a firm basis on which to build a grammar, perhaps a universal grammar. The idea of individual variation in grammars was only ever an attempt, and a stipulative one, to account for the fact that individuals do not always judge structures identically, especially on a binary scale. It was always a weakening of the theory, a great reduction in its generalizability. Closer examination of the data reveals that individuals generally *do* judge structures more or less identically to the rest of the language community, always allowing for some noise. This is exactly what a theory of UG (Universal Grammar) would suggest: the formants of the core grammar are largely the same for everyone.

We shall add one more point which supports the use of multiple informants: it is the nearest thing to Chomsky's 'ideal speaker-listener' that we have. A closer look at Chomsky's (1965; 3) description of the what linguistic theory is concerned with reveals just how close the use of multiple informants allows us to get to his ideal data type.

(5) Linguistic theory is concerned primarily with an ideal speaker-listener in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by memory limitations, distractions, shifts of attention and interest, and errors (random and characteristic) in applying his knowledge in actual performance.

There are two basic aims in this specification of the object of interest of linguistic theory, the first of which can be approximated to by using multiple informants (we discuss the second in 3.2 below). This first aim is to exclude specific effects and render the data maximally general. An 'ideal' speaker-listener is an adult with no speech defect, who grew up as a monolingual, and who is of sound mind. All these should be excluded since they might distort the basic pattern; most linguists are not writing grammars to account for special cases, and those that are need a norm as a point of comparison. This generality is the aim too of the conditions about the homogeneity of the speech community, the perfect linguistic knowledge, and the avoidance of characteristic errors. The variation of the individual's knowledge from the norm, personal quirks in usage, and dialect variation are irrelevant variables which are to be controlled for, Chomsky is saying in imposing these conditions.

This requirement is not easy to meet, but it can be most nearly attained by using multiple informants. There is no such thing as an ideal speaker-listener in one person, but the mean of a reasonably sized sample has exactly the characteristics one would expect an ideal speaker-listener to have. Each individual in the sample will have their particularities, their characteristic quirks, but since these can be assumed to be normally distributed about the 'ideal' value, the mean of the group's judgements will be as free of irrelevant effects as can be achieved. Since Chomsky's definition of the concerns of linguistic theory has become something like a standard, we are therefore adhering to the traditional requirements as nearly as possible in gathering judgements from groups.

3.2 Multiple lexical variants

The primary advantage in using multiple lexical variants is to control for lexis-specific and content-specific effects. Probably every linguist has had the experience of reading an article in which a claim is made supported by an example sentence, but becoming aware that the same structure but with different lexical content reveals a different picture. Using multiple lexical variants reduces this problem because the author must come up with (say) ten different lexical forms which show the same effect. This has two positive consequences; firstly, if ten lexical forms can be found, then the effect is much less likely to be dependent on a particular lexicalization, and there is at least a class of lexical items which it applies to; second, the search for lexical alternatives will make the researcher aware of all those examples which do not work. If these are accounted for, the analysis achieves a higher degree of generalization. Developing lexical variants can thus sharpen the understanding of structures.

The second reason to obtain judgements of a structure in multiple lexical variants relates to the second of Chomsky's aims in (5), when he specifies which behavioural variables do not constitute evidence for the grammar. We may summarise the requirement in that paragraph by saying that *processing* factors are irrelevant. Memory limitations are a processing factor, as are all factors related to attention and concentration. The ultimate criterion is whether a factor reflects underlying competence or whether it is related to the real-time application of this competence in performance. Performance-related factors are not the object of syntactic theory, is Chomsky's message.

Repetition of the judging process, both by multiple informants of multiple lexical forms, directly reduces these performance effects by averaging over judging events. Any temporary absence of concentration will disappear in the averaging process. Since asking an informant to judge a single example repeatedly will cause irritation and incomprehension, precisely the sort of irrelevant processing factor that Chomsky is calling for the exclusion of, lexical variants are necessary for this too. In fact repetition has a further advantage, namely that it brings about standardization of context, creating a de facto experimental context. If you ask someone to give you one judgement, they can continue to think about what to have for dinner, admire the colour of your tie, or worry about the weather at the same time. If you ask them to give you thirty judgements, they concentrate on the task at hand: 'distractions, shifts of attention and interest' are thus reduced. The collection of data is more controlled, in that the conditions under which the data is gathered are more neutral and disconnected from language use; they are thus also held more constant.

3.3 The task

A factor which can make a big difference to judgements is the nature of the task given to the informants. The important point to bear in mind: this should avoid reference to informants' own production. If you ask people if they would say something, they are far more likely to make reference to normative factors in their response: they will often report that they would not use structures which they consider sub-standard. This can be avoided by asking informants to report their responses to linguistic input, not their production; this prevents issues of personal prestige or prescriptively defined 'correctness' playing a role.

There is however another reason, related to the criterion for judgement. Any question about production (such as *Would you use this?*) relates implicitly to occurrence, which can easily confuse the issue. Many structures which are well-formed would rarely in practice occur because there is a simpler way of expressing the same idea, which in practice blocks the quite acceptable but more complex form. This potential source of distortion can be easily avoided by phrasing the instructions for the informants in terms of receptive processing. Our own favourite is *How natural does this sound?*, which has the added advantage of focusing informants on the spoken rather than the written form.

3.4 Multiple degrees of well-formedness

When researchers are gathering judgements they should allow their informants a multi-point scale to answer on. The reason for this is simple: the data has this form, as linguists have long known. 'Like acceptability, grammaticalness is, no doubt, a matter of degree (cf. Chomsky 1955, 1957, 1961)' Chomsky writes in *Aspects* (1965, 11), citing himself three times on the subject. There can be little doubt of Chomsky's view of the issue in the light of these texts.

It is quite legitimate to idealize the data to a binary opposition in order to gain a wider picture, but the linguist doing this should take responsibility for this idealization by carrying it out on the primary language data themself. The collection of pre-idealized data by forcing informants to make a binary choice has evidently led to the current situation of collective amnesia that a process of idealization has taken place. If we asked informants to categorize people into either 'tall' or 'short' then we would obtain a binary distinction; but this does not change the original height distribution in the sample. The task contains an idealization which it transmits to the resultant data.

Instead of delegating the task of idealization to the informants, the linguist should therefore gather data in its raw gradient form and decide for themself where to draw the line between good and bad structures, This is not a trivial task, since there is no clear line of division: any given choice has an element of arbitrariness in it. Not for nothing did Quirk and Svartvik call the binary model of well-formedness 'absurdly gross' (Quirk & Svartik 1966, 49). We are confident that when linguists have done this a few times, they will rapidly realize that they are throwing useful information away, and begin to look with more interest at the wealth of detail that raw, unidealized judgements contain.

The reality of the gradient well-formedness is a key feature of the upgrading of the data base of syntax which we are arguing for, and yet we are content to leave this specification in the group of Desirables, rather than putting it into the Essentials. This needs to be explained. Succinctly, 'murder cannot be hid long; ... at the length truth will out' (Shakespeare, *Merchant of Venice*). Any data set which includes multiple judging events, whether these be by multiple informants or of multiple lexical variants, will exhibit gradience, even if this data is gathered with reference to a binary criterion. If we ask for judgements whether a given structure is 'good' or 'bad' for example, the underlying gradience would be expressed in the relative frequency of the two judgements. Marginal structures will be judged roughly equally often 'good' and 'bad', slightly better examples will have a distribution of judgements skewed to 'good'. Relative well-formedness is thus represented in the data set even if we attempt to exclude it by allowing only two values. In fact the only way to reliably obtain binary data is to ask just one person about one structure once. As soon as two people judge an example (or one person judges it twice) we have a potential of three values ({good, good}, {good, bad}, {bad, bad}), even if each individual judgement is on a binary scale.

An excursus on grammaticality judgements

We should perhaps note here some distinctions which need to be made in order for our claim of the reality of gradient well-formedness to be properly understood. More work is still required in this area, but it is already fairly clear that we need to distinguish three types of grammaticality judgements which use different criteria. The first is the type that we adopt in this paper, which we usually refer to as 'perceived well-formedness'. This is the relative judgements that we gather in our experiments, allowing informants to use as many degrees of well-formedness as they wish. This must be measuring some continuous property, since we can see cumulative effects in the data: we therefore attribute this intuition to computational effort. In the same way as we are able to say whether a sum was difficult to do or a question difficult to answer, we have some perceptible feeling of how difficult a structure is to analyse and process. This is the quality that syntacticians are referring to when they say that one example is 'better than another'. It is worth noting that the precise relationship between linguistic processing and these well-formedness judgements is still obscure.

The second criterion is the traditional binary grammaticality judgement, in which the informant is given a forced choice, grammatical or ungrammatical. We attribute this to occurrence; that is, when we ask informants to give a binary grammaticality judgement, we suspect that the most important factor that they adopt as a criterion is whether they think it would in practice occur or not. This accounts for the intuition of 'absolute (un)grammaticality'. When we feel that a given example is 'fully grammatical', we are saying that we are confident that it would occur; the judgement of 'fully ungrammatical' means that we think it would never deliberately be used.

This distinction of a mental effort criterion of well-formedness and an occurrence criterion is an important step in clarifying terms. These first two criteria for introspective judgements have in common that they are reports of our internal linguistic system. The first, we argue, measures the amount of mental work we have to do in order find an analysis - the most accessible meaningful analysis - of the form of the example structure and to associate it with a meaning. The internal procedure leading to the second type, the occurrence judgement, is less clear. It may consist of a search of our mental corpus of language experience to determine whether we have heard or used the structure type. It may however be that we measure the mental effort required, the first criterion therefore, but then apply a decision criterion, namely whether the example is *good enough* to occur. It need hardly be said that many judging events probably blend the two criteria to a greater or lesser extent, depending on circumstances. It is quite possible that both these criteria contribute to the binary grammaticality judgement, together with other factors.

The third criterion for judgements of grammaticality is very different, since it is not introspective. When syntacticians look at a structure and judge that it is grammatical, since it contains no apparent violation which is standardly held to cause ungrammaticality, then they are using this third type of grammaticality judgement. This type is quite different from the first two because it is dependent on linguistic knowledge and relates to particular assumptions about what a structure should be like. It is not the raw product of naive intuitions, it is the product of a conscious string search. To recap, in this article we make reference to the first of these three types of judgement criteria, perceived well-formedness.

3.5 Summary

Let us sum up what we have said about what steps need to be taken. Generative linguists need to take more care with their data. Specifically they should realize that a single person's judgements are very weak evidence indeed, and that theory building requires a better basis than this. Ones own judgements are fine for clear cases, and they are of course irreplaceable for hypothesis generation, but they contain considerable error variance. To obtain better data, syntacticians should therefore gather their data from multiple informants and have them judge multiple lexical variants of the structures in question. This is the very least quality of data for a study worth publishing. They will obtain more reliable data if they make sure that informants understand that their task is not to comment on their own production. Lastly, to improve the quality of the data further, but also to prevent linguists ever again forgetting that the primary language data is gradient and that categoricity is an idealization, they should allow informants a multi-point scale when gathering judgements. This will greatly raise the quality of the data and its evidential weight, allowing it to become empirical phenomenon in its own right which requires an explanation.

4 The carrot: new perspectives on the grammar

In the first part of this paper we laid out for generative linguists what we think that they are doing wrong and why. In the second part, we explained what we consider to be the most urgent steps to be taken to improve matters. In this section we shall briefly present the results of three studies to illustrate what sort of insights are available with increased use of data. In each case we gathered the introspective judgements of groups of naive informants, using our own variant of magnitude estimation (Bard et al 1996). Our methodology, *thermometer judgements*, differs chiefly in allowing informants to use a linear scale, not a magnitude scale. Experience shows that speakers are not able to produce judgements on magnitude scale even if instructed to, but default to a linear pattern (Poulton 1989). Our variant takes account of this. The name *thermometer judgements* derives from the similarity of our judgement scale to the temperature scale, which is an open-ended scale with two reference points, freezing point and boiling point. Our method also has two reference items to fix the location and amplitude of the scale, setting the lower one at twenty and the upper one at thirty to keep informants away from zero, where distortion can occur. This method thus rather resembles the simple seven-point rating scale, but has no minimum division and no maximum or minimum scores. It thus allows informants the freedom to express their intuitions with minimum constraint.

4.1 Contrastive ellipsis with *nicht*

In this study we tested the perceived well-formedness of contrastive ellipsis with the negative *nicht* ('not') in German. This is part of a larger project in cooperation with our colleagues Susanne Winkler and Andreas Konietzko (eg Winkler 2005). In English, subject and object contrastive tags have clearly different forms: subjects have a finite verb form, as in *Steve likes jogging, but Rachael* *doesn't*, while objects have none *Steve likes jogging, but not cycling.* Functionally similar contrastive ellipses in German work very differently, consisting of just the subject or object either before or after the negative *nicht.* While all four of these variants seem possible in an appropriate context, some preferences are apparent, but an individual's judgements are insufficient to clarify what factors play an important role.

We therefore tested contrastive subjects and objects before and after *nicht*, each in three contexts which differed in the expectations they triggered: subject contrast, object contrast and no contrast expectation. We illustrate these materials in (6), which first contains the three biassing question contexts and then the four answers with contrastive tags.

The subjects and objects in each lexicalization were chosen to resemble each other visually (eg *Margareta* ≈ *Magazine*), so that the grammatical function in the tag would be recognized as late as possible. The thirty participants each saw the twelve experimental conditions and twelve lexicalizations twice, as well another twenty-four unrelated items and ten standard comparison items, all in randomized order. We present the results in Figure 6. As before the error bars in the chart show the 95% confidence interval of the mean normalized judgement scores for each condition.

(6) i *no contrast expectation*
Interessieren sich deine Freunde für Politik?
'Do your friends take an interest in politics?'

ii *subject contrast expectation*
Interessieren sich Peter und Margareta für Politik?
'Do Peter and Margaret take an interest in politics?'

iii *object contrast expectation*
Lesen deine Freunde Zeitungen und Magazine?
'Do your friends read newspapers and news weeklies?'

a. Naja, Peter liest oft Zeitungen, aber Margareta nicht.
Well Peter reads often newspapers but Margaret not
'Well, Peter often reads newspapers, but Margaret doesn't'

a' ... aber nicht Margareta
... but not Margaret.

b. Naja, Peter liest oft Zeitungen, aber Magazine nicht.
Well Peter reads often newspapers but news.weeklies not.
'Well, Peter often reads newspapers, but not news weeklies.'

b' ... aber nicht Magazine.
... but not news.weeklies

The results of this experiment show that the informants were sensitive to the context, since the contrastive tags were judged better when they fulfilled the expectations raised in the preceding questions.[6] But the contrast of subjects and objects and their position before or after the negative are even stronger effects.[7] Subjects are always better before the *nicht*, objects always better after, but the

[6] Repeated measures anova of the interaction of biassing context and the grammatical function in the tag is significant ($F_1(2,58)$=7.06, p_1=0.002, $F_2(2,22)$=5.04, p_2=0.038).

[7] Repeated measures anova of the interaction of the grammatical function in the tag and its position is highly significant ($F_1(1,29)$=77.5, $p_1 < 0.000$, $F_2(1,11)$=31.3, $p_2 < 0.000$).

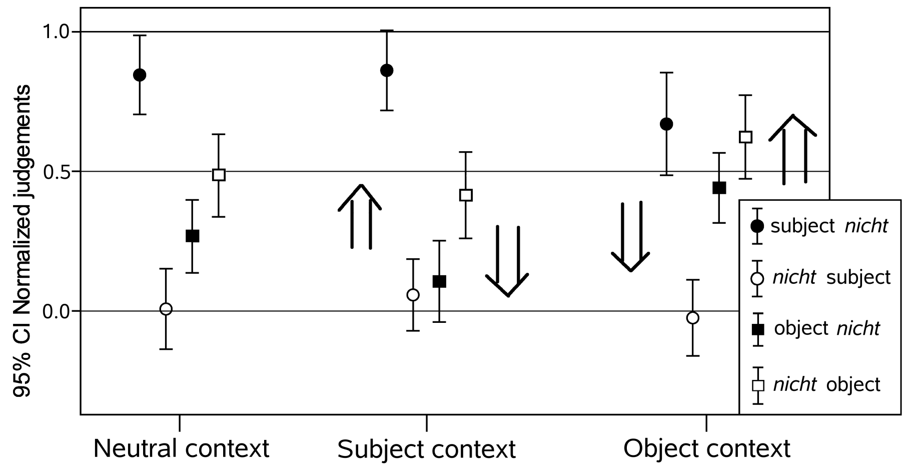

Figure 6: The results of our study on subject and object contrastive ellipses, in three contexts: neutral, favouring subject contrast, and favouring object contrast. The effect of context is visible (see arrows), but the preference for subjects to preceed, and objects to follow the negative is robust.

preference was stronger for the position of the subject than for the position of the object.

This was just a pilot study for a research project, but we shall make two points. First, it illustrates the degree of detail which this data type allows for. The results show a very fine differentiation, far sharper than any individual can produce. But these patterns are highly consistent: across all three context conditions, the perceived well-formedness of the syntactic conditions remains extremely stable, affected only by the predictable effect of the context. This demonstrates that the differences between the syntactic conditions are real effects, and not just random fluctuation.

Second, this fine grain of difference can be useful, as it sometimes gives us clues about the correct structural analysis. We may distinguish two main analyses of this sort of tag structure: those which take the tag to be a clausal structure with ellipsis of some sort and those which suggest that there are no more nodes than the overt form requires (eg Winkler 2005 vs Culicover & Jackendoff 2005). This result does not support the second analysis, in which the tag is structurally just a negated NP constituent, for this account does not predict any subject-object asymmetry in the position of the NP relative to the negative. The clausal analysis is supported, however, since the tags exhibit certain effects known from German clauses. In our study the arguments in the tags prefer their canonical clausal positions relative to the negative: subject preceding, object following. In addition, the rigidity of this restriction is stronger for the subject than for the object. Exactly these findings are the features of German clauses (cf Uszkoreit 1987, Pechmann et al 1994, Hemforth & Konieczny 2000): subjects are strongly preferred to precede anything else, while the position of objects is much more flexible, though the location after a negative is probably canonical, if we accept that a negative appears at a VP boundary. Since the patterns obtained from the tags matches the patterns found in full clauses, the analysis of these contrastive tags as elliptical clauses therefore receives support. This sort of finding thus makes the collection of more detailed data worthwhile, since it feeds directly into the syntactic analysis.

4.2 The pattern matching technique: coherent structures

This second example of the fruits of paying more attention to data relates to *coherence* in German, a sort of clausal union phenomenon in which clauses with a subset of matrix verbs seem to fuse with subordinate non-finite clauses (Bech 1955 60ff). Certain matrix predicates seem able to take both coherent and non-coherent complement clauses to a greater or lesser extent, others permit only the one or the other. In (7) both structures can have the same interpretation, but the first (7a) is monoclausal (=coherent) and the second (7b) is biclausal with apparent extraposition of the embedded complement (=non-coherent).

(7) a. [Sie wagt ihn nicht zu stören.]
she dares him not to disturb
'She does not dare to disturb him.'

b. [Sie wagt nicht, [ihn zu stören.]]
she dares not him to disturb
'She does not dare to disturb him.'

Certain other structures are more complex and coherence seems to have many further-reaching structural effects. Two of these less clear cases have given rise to controversy, because they seem to deliver contrasting messages about which matrix predicates permit coherence effects; they are the *third construction* - (8a) and the *long passive* - (8b).

(8) a. ...dass ihn der Gipser versucht zu beschwindeln.
...that him the plasterer tries to cheat
'...that the plasterer is trying to cheat him.'

b. ...dass er zu beschwindeln versucht wird.
...that he to cheat tried is.being
'...that the attempt is being made to cheat him.'

In the *third construction*, an argument of the embedded verb appears in the verbal field of the matrix verb (here: before the subject), although the embedded verb itself has been extraposed. This seems strong evidence of clausal union at some stage in the derivation. A similar conclusion is supported by the *long passive*, in which the object of the embedded verb appears as a nominative, as if it were the logical object of the matrix verb which is passivized. On the face of it, the third construction and the long passive both seem to be dependent upon the clausal union of coherence. However, only a more restricted set of matrix predicates seem to allow the long passive, which has thrown the common structural origin of the two phenomena into doubt and caused the surmise that they cannot both be symptoms of the same coherence factor (Stechow & Sternefeld 1989, cf Schmidt et al 2005,).

Our study aimed to test this using the *pattern matching technique.* In this we test structures over a range of conditions, with the hypothesis that, if two structures are related, they will respond in a parallel manner to the conditions. In this particular case, we tested the two structures *third construction*

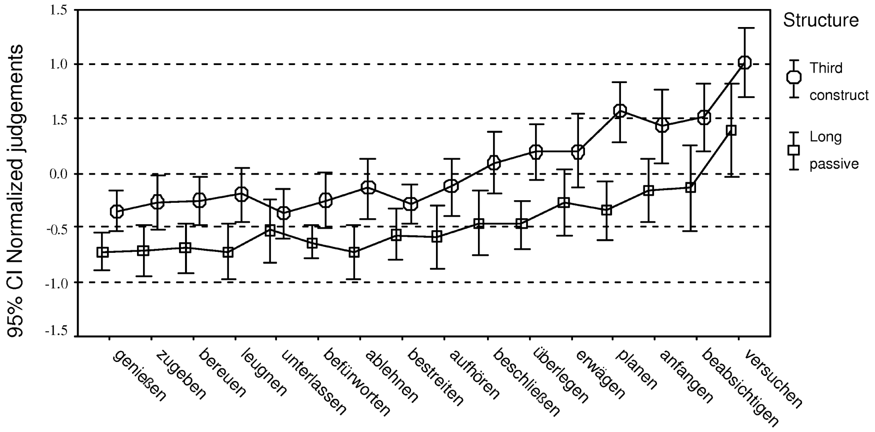

Figure 7: The *third construction* and *long passive* are not judged the same, but the effects of the predicate hierararchy are parallel. The patterns match.

and *long passive* over sixteen matrix predicates. These predicates formed a continuum of coherence-friendliness, with both strong coherence verbs such as *versprechen* ('promise') and much weaker ones such as *zugeben* ('admit'). If our two structures are both licensed by the coherence phenomenon, they should respond similarly to the degree of coherence-friendliness in the predicates (9b). It is worth underlining how this hypothesis differs from the standard hypothesis, which assumes a binary model of well-formedness (9a).

(9) a. *Binary well-formedness hypothesis*
If structures X and Y are related, both will be either possible or impossible in a given condition (here: with a given predicate).

b. *Pattern matching hypothesis*
If structures X and Y are related, they will have similar patterns of judgements over sets of conditions (here: over predicates).

We carried out this study with the thermometer judgements method. In all, twenty-six informants judged forty-five structures, of which thirty-two were part of this experiment and thirteen were fillers. The results are shown in Figure 7, where the predicates are ordered by their judged well-formedness.

The figure clearly shows three things: first, that the two structures are not judged to be equally acceptable. The long passive is consistently worse than the third construction. Second, it is plain that there is an effect of the matrix predicate: precisely those verbs which have traditionally been thought of as favouring coherence make these structures much better than the other verbs. But third, the effect of the hierarchy of predicates on the two structures is the same, as statistical tests confirm.[8] So although the long passive is a less well-formed structure than the third construction, the two structures respond to the

[8] Repeated measures anovas by subjects and items show significant effects of the structure ($F_1(1,25)$=27.0, $p_1 < 0.001$, $F_2(1,7)$=96.5, $p_2 < 0.001$), and of the verb ($F_1(15,375)$=13.7, $p_1 < 0.001$, $F_2(15,105)$=13.0, $p_2 < 0.001$). There is no interaction of the factors verb and structure ($F_1(15,375)$=1.15, p_1=0.312, $F_2(15,105)$=0.851, p_2=0.618).

features of the verb in the same way. This is striking evidence of the reality of the coherence syndrome and that these two stuctures are both coherence effects.

More generally, we can make two points. First, the binary model of well-formedness is seen to be obscuring the full picture here. The binary hypothesis in (9a) will yield the wrong answer: it is not the case that the well-formedness of the one structure implies the same of the other, even though they are both dependent on coherence. The pattern matching technique on the other hand, which makes use of a continuum of well-formedness, is clearly providing a more adequate account of the facts.

Second, we should note that most of these structures are very ill-formed and would never occur. The experimental capture of distinctions between structures which are all quite ungrammatical has important theoretical implications. Such systematic differences demonstrate that the grammar cannot possibly be merely a product of exposure to the language or of the learning of constructions via frequency patterns, as is sometimes argued (eg Bybee & Hopper 2001). The structures we tested here are vanishingly rare even with the most coherence-friendly matrix predicates, but we find the same distinction between the two structures very consistently right down among the quite ungrammatical structures, with matrix predicates that never take coherent clausal complements. An exposure-based account would predict that these structures should be judged equally bad, since they are equally frequent, that is, they never occur at all. Since the distinction is present and consistent, right down among the impossible examples, the distinctions cannot be learnt, but must rest on some system-internal factor. This would support the assumption that the grammar has an independent existence at a fairly abstract level and furthermore that parts of it are in all probability universal. Again we see that data collected under controlled conditions can have important implications for theory.

4.3 Relative clauses without relative markers

Our third example study is more speculative than the first two and we shall offer only the most tentative account of it, but we think that it provides a useful example of the extension in perspectives for explanation which better data allows. In previous work we have found that syntactic constraints familiar from English but thought to be absent from German (eg superiority, *that*-trace effect) are indeed active in this language, but are merely less obvious, put differently, they have weaker violation costs (Featherston 2003, 2005). This provides evidence that more effects apply cross-linguistically than is generally assumed, itself an interesting finding since it is a key prediction of UG.

In this study we aimed to test whether another phenomenon known in English, the ability to omit object relative markers in restrictive relative clauses (=RRCs), would be replicated if we tested the word-for-word equivalents in German.[9] Descriptive grammars of English confirm that object relative markers in

[9] We tend to agree with Quirk et al (1985 480ff) that the English relative *that* is a relative pronoun and not a complementizer. Radford (1988 482ff) and Huddleston & Pullum (2002 1056f) present the opposing view, advancing arguments based on idiosyncrasies of distribution and behaviour. We regard this as merely morphological and functional impoverishment due to lexical forms competing for a single function. But space does not permit us to discuss this issue and the question is not critical here, so we shall refer to the relative *that* as a relative marker, thus excluding neither analysis.

RRCs can be omitted, while subject relative markers cannot (Quirk et al 1985, 365ff, Huddlestone & Pullum 2002, 1054f, for factors Wasow et al 2005) - (10).

(10) a. You know the boy (that) I like
b. You know the boy *(that) likes me.

German relatives have rather different qualities to English relative clauses, since they are generally verb-final, though there are also under rather restricted conditions verb-second apparent relative clauses (Gärtner 2001) - (11). We illustrate these only to show that these are clearly different from the word string equivalents of the English structures which we intend to test.

(11) Ich kenne sogar Leute, die lesen Chomskys Bücher.
I know even people *RelMark.nom* read Chomsky's books
'I even know people who read Chomsky's books.'

Our aim is to examine whether we find a subject-object asymmetry in the perceived well-formedness of the German *word string equivalents* of the English data, which might be caused by the same factor as the asymmetry in English. If we find this, we have evidence for a universal effect, since these strings are not part of the German language. We gathered judgements of eight syntactic conditions in (12) using the thermometer judgements method. The matrix clauses in (12-i) and (12-ii) control for the effect of the case of the antecedent. Conditions (12a) and (12b) are the standard German verb-final subject and object relative clauses. Conditions (12c) and (12d) are the German string equivalents of English subject and object relatives clauses without relative markers.

(12) i. *Nominative antecedent*
Das ist der Junge ...
that is the boy.*nom* ...

ii. *Accusative antecedent*
Du kennst den Jungen ...
you know the boy.*acc* ...

a. ...der mich sehr gern mochte.
...who me very much liked

b. ...den ich sehr gern mochte.
...whom I very much liked

c. ...mochte mich sehr gern.
...liked me very much

d. ...ich mochte sehr gern.
...I liked very much

Using the thermometer judgements technique, we gathered the judgements of twenty-eight informants of the eight structures in eight different lexical forms, counterbalanced so that each person saw each lexical content and syntactic condition exactly once. These structures were presented in random order together with the twenty-eight items of another study and five standard comparison set items. We present the results in Figure 8.

The four standard German relative structures are across the top. They are all judged to be about equally good, so there is no background asymmetry

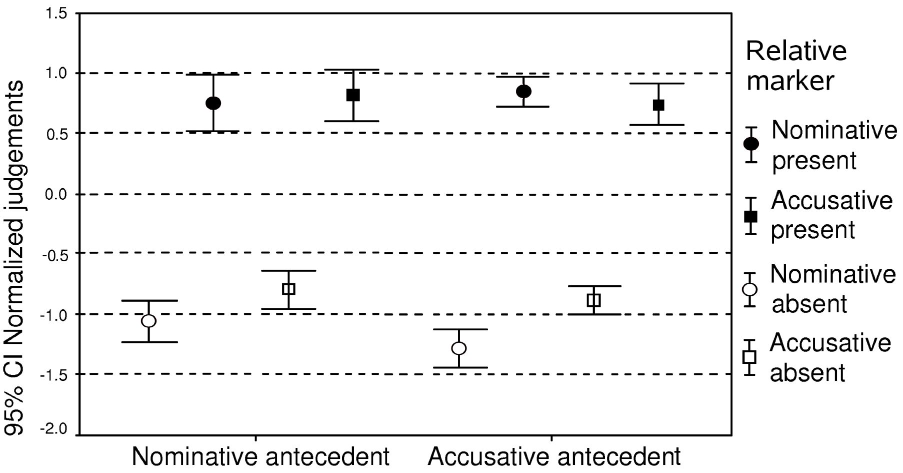

Figure 8: The results of our study looking for a subject-object relative marker dropping asymmetry in the word string equivalents of English relatives. The data shows that the impossible German structures with the same word order as English relatives show an effect similar to that found in English.

in German. The string equivalents of the English structures without relative markers are across the bottom. The left-hand pair are the relative clauses with a nominative antecedent, the right-hand pair those with an accusative antecedent. There is little difference between them, so we can conclude that the case of the antecedent plays no important role. But within each pair there is a clear difference: the structures with absent *accusative* relative markers are judged better than those with absent *nominative* relative markers, just as in the English structures.[10]

The evidence is clear: an effect we find in English relative clauses is found by German speakers in the German word-for-word equivalents too. But before we can interpret this data we must first ask whether there is any known pattern which would account for the finding. The effect is not a more general feature of German relative clauses, since we tested these too in the experiment. Nor is the effect dependent upon the case matching of the antecedent and clause, for the case of the antecedent was controlled for, and the match had no effect. The preference is not caused by the existence of a locally grammatical sub-string *... der Junge mochte mich sehr gern* with the nominative antecedent and the dropped nominative relative marker - (12c), since this condition is judged worse, not better. We can also exclude as an account of the finding as an effect of the German phenomenon of topic drop, in which a preposed (usually non-subject) topic can be phonetically null. If informants treated the two clauses as separate main clauses, but treated the lower clause as an example of topic drop, then this would be expected to improve judgements. But this cannot account for our finding, since this wrongly predicts the (12c) to be better than (12d). Last of all, there is no reason to believe that knowledge of English influenced our

[10] Repeated measures anovas by subjects and by items shows a significant interaction of the case of the relative pronoun and its presence or absence ($F_1(1,27)$=9.57, p_1 = 0.005, $F_2(1,7) = 11.94$, $p_2 = 0.011$).

informants, since the data was gathered in Germany in an entirely German context. In sum, none of these irrelevant effects offer an account for our finding.

We therefore tentatively suggest that the effect in German is caused by the same factor, whatever that may be, as the effect in English. Although these structures are quite ungrammatical in German, probably due to the very different clausal positions of German verbs, it would appear that the effect known from English is nevertheless measurable under experimental conditions. Effects such as we found here do not just occur at random (and they certainly are not learnt from exposure): they must have a cause.

Argumentation of this type is unfamiliar in syntax, and no doubt many readers will react incredulously, but this sort of finding is predicted by our conception of UG. Since the primary data shows no sign of binary well-formedness, we abandon this idealization. It follows that UG cannot have a binary quality: we therefore conceive of UG as a set of probably cognitively-driven non-categorical constraints on well-formed structural representations (cf Bresnan et al 2001). Not all of them will be immediately apparent in all languages, but they are present all the same, and if the occasion arises they will be active even in the minds of speakers of languages which do not usually reveal evidence of them.

We thus predict that German speakers will find object relative marker dropping better than subject relative marker dropping, for the same reasons that English speakers do. We also predict that speakers of languages without wh-movement will be sensitive to island constraints and speakers of non-pro-drop languages will find phonetically null subjects better than phonetically null objects, as in Italian. Exactly what is causing this finding must await further work but the point, we hope, is clear. There is a lot of evidence about the grammar available which we have so far hardly started to gather, let alone analyze. Carrying out careful and systematic comparisons and gathering controlled data reveals new facts about the language and offers new explanatory possibilities.

4.4 The value of evidence

We used experimental techniques to obtain these results but we must stress that this sort of finding does not always necessitate the degree of control that we put in here. Empirically grounded grammars do not require tightly controlled experiments, but they do require basic respect for the quality of evidence. The intuitions of a small group of people gathered with only fairly modest controls will produce a less detailed but generally similar picture to our experiments. All that the syntactician working with introspective data need do to produce empirically grounded work is move some way along the continuum of data quality, like adopting the Essentials (3) and the Desirables (4).

If all syntacticians took this step we would find a marked improvement in the linguistic quality of work in the field, since theoretical work would be accounting for empirical facts. This would cost grammarians little, since judgements gathered from groups by the methods we recommend here are the identical data type to an individual's judgements, but with the irrelevant effects and performance factors reduced. Indeed one of the attractive features of controlled judgements is that the findings can be checked for plausibility by any individual, that is, the judgements of the group are accessible and replicable for a single speaker. The approach thus fulfils Chomsky's requirement that 'the speaker-hearer's linguistic intuition is the ultimate standard that determines the accuracy of any proposed grammar, linguistic theory, or operational test' (1965, 21).

Studying grammar by gathering 'better' judgement data and examining it closely thus has many advantages. First and foremost it allows the linguist to avoid the stick of criticism for producing theory without an appropriate basis in data. There can be little satisfaction in producing or reading work which so clearly fails to satisfy scientific and academic standards. But further, collecting more and better data permits new insights into phenomena in the grammar. Quite small parts of very few languages have been studied using controlled elicitation of judgements, in part because of the widespread view that this data was invalid. But findings such as those that we have presented in this section show clearly that there are aspects of the nature of the grammar which are being obscured by insufficient attention to the primary language data. There are many previously unknown facts to be discovered, analysed, and interpreted, with only moderate change in the paradigm of research of generative syntacticians. For linguists who dare to take this step, there is a wealth of new descriptive, theoretical, and explanatory progress to be made. This is the carrot.

5 For and against

In this section we will summarize some of the positive implications of the increased attention to data that we are calling for, but also try to answer some objections and counter-arguments that we have met in the past. But let us first restate what we wish to see happen. Each syntactician should take it upon themself to apply at least the Essentials (3) and normally also the Desirables (4) in their work. The first requires that we gather judgements from multiple informants and construct multiple lexical variations of structures. The second proposes that we elicit judgements only as responses to others' output, and that we allow informants a range of values for their answers. Syntacticians should adhere to these standards in their own work and apply them to others when reviewing, for example. If these simple requirements were met, all published work using judgements would meet basic standards of empirical adequacy.

5.1 Generality

All other things being equal, any insight of greater generality is more interesting than one of less generality. Theories constructed on the basis of data gathered as we have specified are of greater generality. If we test an effect on twelve different lexicalizations of a structure, we exclude the possibility that applies just to a single example, and we have a stronger claim that it might apply in all relevant cases. If we obtain the judgements of twenty-five informants, we can reasonably assume the findings to be valid for the whole population of speakers and not just for a single person. Furthermore, such findings become hard facts which need accounting for, not the subjective impressions of one person, which can be put down to whim. If we then in an additional step go on and find the same syntactic effect in another language, we immediately have a strong candidate for a linguistic universal. It is not obvious why a single person's judgements, with their idiosyncrasies and noise component, should be even remotely as interesting as an object of study as these generalizations. The big questions in generative linguistics are those which refer to all speakers, not just one speaker, and to the whole language, if possible to all languages, not just to a single lexical string.

5.2 Progress in grammatical theory

The use of data as we have suggested will allow theory to proceed in new, descriptively adequate directions. First, because it allows theory to exit from its current immobilism. It cannot be a healthy position for the theory of grammar to be in, when for many phenomena, multiple widely varying analyses are possible and the field offers no procedures for deciding between them. Generating new hypotheses may be more intellectually rewarding than testing old ones, but if the methods used in the field do not even allow us to detect which of the widely differing syntax architectures (eg optimality theory, minimalism, HPSG, LFG ...) is the most correct framework, then our theoretical alternatives are simply too weakly constrained by their empirical base. If we cannot test and discard hypotheses, we can make no progress.

Second, the use of data will assist progress in grammatical theory because empirically unfounded assumptions are closing researchers' eyes to possibilities. In the recent work on evidence for innateness (Penke & Rosenbach 2004) some authors effectively dismissed UG for the simple reason that they could not find exceptionless constraints on the form of human language. If they looked at the primary data in a little more detail, they might conclude that they were asking the wrong question. Experimentally gathered judgements reveal constraints on language to be survivable effects, not categorical restrictions, but often to apply cross-linguistically. Universal grammar is richly structured and readily accessible, we would argue, but it is not categorical. Linguists need to look at the data first and develop their models afterwards, not the other way round.

5.3 Improving debate

One more advantage of treating data with greater respect is that it might lead to an improvement in the quality of linguistic debate. When we concern ourselves with data, then we appreciate what conclusions can be drawn from a particular set of findings but also how further work might support or disprove an analysis. In the light of this, if we read that some phenomenon can be accounted for if we assume an additional functional projection with an α-morpheme then we shall be ready to look for contradictory evidence. If we are forced to conclude that there can be no such evidence, for the hypothesis makes no empirically testable predictions, then this would be an important step towards the development of an evaluation measure for linguistic analyses in terms of *empirical adequacy*, for falsifiable hypotheses are stronger than unfalsifiable ones. We would therefore hope that the use of data would more tightly constrain the range of accounts advanced for syntactic phenomena.

5.4 Some frequent objections answered

Does this not require the abandonment of the generative model? No. Chomsky's great merit was in formulating an essentially psycholinguistic question which has inspired and still inspires linguists to look for a model of the grammar which is psychologically real and learnable by a two-year-old. The question remains. *Syntactic Structures* and *Aspects* are as relevant as ever. Chomsky was also taking an entirely reasonable position at the time when he suggested: 'The critical problem for grammatical theory today is not paucity of evidence but rather the inadequacy of present theories of language to account for masses of evidence that are hardly open to serious question' (1965). The problem occurs when later linguists attempt to use the same data-light approach on evidence which *is* open to serious question. This is clearly not in line with Chomsky's intentions. Our own approach, on the other hand, is fully consistent with his views.

> '...one whose concern is for insight and understanding [...] must ask whether or to what extent a wider range and more exact description of phenomena is relevant to solving the problems he faces. In linguistics, it seems to me that sharpening of the data by more objective tests is a matter of small importance for the problems at hand. One who disagrees with this estimate [...] can justify his belief in the current importance of more objective tests by showing how they can lead to new and deeper understanding of linguistic structure. Perhaps the day will come when the kinds of data we now can obtain in abundance will be insufficient to resolve deeper questions concerning the structure of language.' (1965)

The case for data in syntax is precisely that the time has come, as Chomsky foresaw, for more detailed data to be consulted (Schütze 1996, 27), since methods in elicitation have advanced. The results would suggest that deeper understanding is indeed obtained by higher quality data.

You are just gathering acceptability judgements. I am only interested in grammaticality. It is true that performance factors are still amply represented in judgements gathered under controlled conditions. But they play a smaller role than in informal linguist's judgements, not a larger role. Using multiple informants reduces individual differences and using multiple lexicalizations and a standard experimental context reduces the effect of known performance factors. This objection is thus misguided.

But more generally, we need to treat the differentiation of acceptability and grammaticality with care. Linguists usually dismiss known performance factors as non-grammatical, and tend to limit the domain of 'grammatical' effects to the sentence. But another important criterion adopted is whether a constraint causes categorical ungrammaticality. As we have seen however, this assumption is problematic, since the idea of categorical syntactic constraints is dependent on the idealization to binary grammaticality.

In our cross-linguistic studies on the superiority effect, for example, we have found that the same effect which is thought to be narrowly grammatical in English is regarded as merely stylistic, and thus non-grammatical, in German. One of these assumptions must be wrong, but even with this information we have no obvious way of determining which one it is. As Chomsky says, '...there is no reason to expect that reliable operational criteria for the deeper and more important theoretical notions of linguistics (such as 'grammaticalness' [...]) will ever be forthcoming.' (1965, 19). It is thus not clear that the distinction is well-defined enough for it to be usable as a analytical tool.

Our own approach is to discount from being grammatical all effects which can be accounted for by known performance or processing factors, and all effects whose domain exceeds the sentence level. All other effects we assume to be narrowly grammatical. This results in us attributing to the grammar certain restrictions whose violation costs are not sufficiently great to cause the violating structure to be excluded absolutely from being part of the language. We see no principled reason to exclude these.

Isn't it the case that the core grammar is categorical while markedness and other performance factors are gradient? This position cannot be disproved but we find no evidence to support it. If it was correct, we would expect to find at least occasional glimpses of categoricity under controlled conditions. In fact we rarely if ever find effects in judgement studies which look like what we would expect of categoricity. One symptom might be that a structure breaking a categorical grammatical rule should be judged so bad that it can get no worse. In fact it is more or less always possible to make a structure worse. Another sign of categoricity would be if a constraint violation caused a structure to sink to a low point of perceived well-formedness, no matter how good or bad the structure was otherwise. We do not find this either; a violation generally causes a structure to be judged worse by a certain amount.

The reason that this is not proof of the non-existence of categoricity is that other factors play a role too, and we cannot distinguish them. First, we find some distortion towards the ends of the scale and the picture becomes less clear. The comprehensibility of a structure and the correctability of violations play a role in judgements too. It is apparent that exactly those violation types which are good candidates as categorical (eg subject-verb agreement) are most correctable and have least effect upon comprehension, so that an example such as *What Mandy want do next?*, which should be categorically bad, is judged *better* than a mere subjacency violation *What did Mary ask who had taken?*. The reason is probably that the agreement violation is easier to correct, and leaves the example fairly readily comprehensible. These irrelevant factors are plainly not narrowly grammatical, but we cannot prove that they are not obscuring the features we would expect of categoricity. We can thus not prove categoricity not to exist, but there is no evidence which supports it either.

What about the intuition of absolute (un)grammaticality? The intuition is real, but our research into judgements leads us to conclude that intuitions of absolute (un)grammaticality are different in nature to our intuitions of perceived well-formedness, in that they reflect potential or actual *occurrence*. One piece of evidence which leads to this conclusion is the fact that even naive informants can readily distinguish the two types. One often hears the judgement: 'It's better than the other one but I would never say it.' Here the speaker is distinguishing perceived well-formedness from occurrence information. So an example which is perceived to be absolutely grammatical is *well-formed enough* to be produced, while one which is felt to be absolutely ungrammatical is *ill-formed enough* for it not ever to occur. This suggests that intuitions of absolute (un)grammaticality, real though they are, are not responding to the same criterion as standard perceived well-formedness (see the excursus in 3.4). We should also note that the intuition of absolute grammaticality applies only to individual example structures, not the scale of well-formedness: there are plenty of examples which are intuitively neither fully grammatical nor fully ungrammatical. The intuition of absolute grammaticality cannot replace the requirement for a continuum model of well-formedness.

This is just data, not theory. Data doesn't replace for theory. This reproach (made in a review of a paper) illustrates the intellectually corrosive effect of paying insufficient attention to data. A theory is always a theory about a set of data. If the data can be accounted for by known surface factors (weight, focus, ..) without additional theory, then no additional theory is necessary. Furthermore, this account is *more explanatory* than an alternative which requires us to posit an additional mechanism. Data is a pre-condition for theory, and the quality of a theory can never exceed the quality of the data set which it is based on. Descriptive adequacy is a precondition for explanatory adequacy.

You can't ask us to do an experiment each time we need a judgement! No, that isn't necessary. In fact an individual's judgements give a fairly good idea of what the judgements of a group will show, and ones own judgements can be informally checked by asking a couple of other independent speakers. But when a paper is to be published, it does not seem unreasonable to expect the author to check any judgements it contains carefully by (at least) preparing multiple lexicalizations and formally asking ten or more other people. The author should then account for the data that they find, noting any idealizations that they apply to it. This process need not be repeated endlessly: when one linguist has tested a data set properly, others can simply refer to their findings. This is one more reason to include details of data gathering in published work.

There are also initiatives to systematize the gathering of an individual's judgements so as to improve their quality without the additional effort of running an experiment. Our own contribution to this is to develop a set of standard reference examples, relative to which judgements can be made more accurately. This method can be insightfully related to that of estimating temperature. Without a temperature scale, we would find it difficult to judge and communicate our judgements of temperature. The Celsius scale makes this much easier, by giving fixed points relative to which other temperatures can be identified. We use five (sets of) example structures which span the range of perceived well-formedness and thus provide known points of comparison. This is thus a well-formedness scale parallel to a Celsius scale and should aid the giving of more differentiated judgements and the communication of judgements. We use these in our experiments, but mention them here only as a example of how judgements as a data type can be hardened up without demanding disproportionate effort from working linguists.

What evidence have you for arguing for a gradient model of grammaticality? First, if we elicit judgements giving informants a free choice of scale, they always use a continuum of well-formedness. There is no clear step or dissociation between well-formed and ill-formed structures, and our tightly controlled tests leave little room for the suggestion that irrelevant performance factors are hiding a binary opposition. Second, frequency data backs this up: there are frequent structures and less frequent structures which nevertheless occur. Frequency and perceived well-formedness tend to co-vary. We can account for frequency variation in terms of gradient well-formedness if we accept that the well-formedness of a structure is a (just 'a' not 'the') causal factor of frequency. Third, syntacticians standardly assume a binary model of grammaticality, but they tacitly admit that this is an idealization by using intermediate values (cf Chomsky 1986, for some extreme cases see Lakoff 1973, Müller 1995, Wurmbrand 2001). Fourth, this feature allows us to account for historical language change and dialect variation. A gradient system allows language change to creep, instead of having to jump from one state to another.

But allowing gradience in the grammar is a weakening of the theory. Theories are far more constrained by being required to account for the data than they are constrained by the requirement to be elegant. If linguists really want a constrained theory, they should reduce the amount of the idealization and restrict themselves to making only claims that the data actually supports. Constraining theory by such standards as 'elegance' and 'economy' is hardly any constraint at all since we have so little idea how to identify elegance and economy in a theory, and even if we did, this would constrain theory to be elegant, rather than constraining it to be accurate. Data gathered under controlled conditions, on the other hand, is very restrictive and unyielding, but permits the strongest theory of all, that which accounts for the facts.

6 References

Bader M. & Bayer J. (2006) *Case and Linking in Language Comprehension.* Dordrecht: Springer.

Bard E., Robertson D. & Sorace A. (1996) Magnitude estimation of linguistic acceptability. *Language 72* (1), 32-68.

Bayer J. (1990) Notes on the ECP in English and German. *Groninger Arbeiten zur Germanischen Linguistik 30*, 1-55.

Bech G. (1955) *Studien über das deutsche verbum infinitum, volume 1.* Copenhagen: Munksgaard.

Boersma P. & Hayes B. (2001) Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry 32*, 45-86.

Bresnan J., Dingare S. & Manning C. (2001) Soft constraints mirror hard constraints: Voice and person in English and Lummit. In: Butt M. & King T. (eds) *Proceedings of LFG01 Conference*, 13-32. Stanford: CSLI.

Bybee J. & Hopper P. (eds) (2001) *Frequency and the Emergence of Linguistic Structure.* Amsterdam: Benjamins.

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

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

Chomsky N. (1964) Degrees of grammaticalness. In: Fodor J.A. & Katz J. (eds) *The Structure of Language: Readings in the Philosophy of Language*, 384-389. Eaglewood Cliffs, NJ: Prentice-Hall.

Chomsky N. (1965) *Aspects of the Theory of Syntax.* Cambridge, Massachusetts: MIT Press.

Chomsky N. (1986). *Barriers.* Cambridge, Massachusetts: MIT Press.

Culicover P. & Jackendoff R. (2005) *Simpler Syntax.* Oxford: OUP.

Fanselow G. (1987) *Konfiguationalität: Untersuchungen zur Universalgrammatik am Beispiel des Deutschen.* Tübingen: Narr.

Featherston S. (2002) Coreferential objects in German: Experimental evidence on reflexivity. *Linguistische Berichte 192*, 457-484.

Featherston S. (2003) That-trace in German. *Lingua 1091*, 1-26.

Featherston S. (2004) The Decathlon Model: Design features for an empirical syntax. In: Kepser S. & Reis M. (eds) *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives* Berlin: de Gruyter.

Featherston S. (2005) Universals and grammaticality: Wh-constraints in German and English. *Linguistics* 43 (4), 667-711.

Gärtner H.-M. (2001) Are There V2 Relative Clauses in German? *Journal of Comparative Germanic Linguistics 3* (2), 97-141.

Greenbaum S. (1977) The linguist as experimenter. In: Eckman F. (ed) *Current Themes in Linguistics: Bilingualism, Experimental Linguistics, and Language Typologies*, 125-144. New York: Wiley.

Grewendorf G. (1985) Anaphern bei Objekt-Koreferenz im Deutschen. Ein Problem für die Rektions-Bindungs-Theorie. In: Abraham W. (ed) Erklaerende Syntax des Deutschen, 137-171. Tübingen: Narr.

Grewendorf G. (1988) *Aspekte der Deutschen Syntax. Eine Rektions- Bindungs-Analyse.* Tübingen: Narr.

Grewendorf G. (1995) German. A grammatical sketch. In: Jacobs J., von Stechow A., Sternefeld W. & Vennemann T. (eds) *Syntax: An International Handbook of Contemporary Research*, 1288-1319. Berlin: de Gruyter.

Haider H. (1983) Connectedness effects in German. *Groninger Arbeiten zur Germanischen Linguistik 23*, 82-119.

Haider H. (1993) *Deutsche Syntax - Generativ.* Tübingen: Narr.

Hemforth B. & Konieczny L. (eds) (2000) *German Sentence Processing.* Dordrecht: Kluwer.

Huddleston R. & Pullum G. (2002) *The Cambridge Grammar of the English Language.* Cambridge: Cambridge University Press.

Keller F. (2000) Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. PhD Thesis, University of Edinburgh.

Keller F., Corley M., Corley S., Konieczny L. & Todirascu A. (1998) WebExp: A Java Toolbox for Web-Based Psychological Experiments. Technical Report HCRC/TR-99, University of Edinburgh.

Kiziak T. (in press) Long Extraction or Parenthetical Insertion? Evidence from Judgement Studies. In: Dehé N. & Kavalova Y. (eds). *Parentheticals*, 121-144. Amsterdam: Benjamins.

Labov W. (1975) Empirical foundations of linguistic theory In: Austerlitz R. (ed) *The Scope of American Linguistics*, 77-133. Lisse: Peter de Ridder.

Labov W. (1996) When intuitions fail. In: McNair L., Singer K., Dolbrin L. & Aucon M. (eds) *Papers from the Parasession on Theory and Data in Linguistics. Chicago Linguistics Society 32*, 77-106.

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

Lutz U. (1996) Some notes on extraction theory. In: Lutz U. & Pafel J (eds) *On Extraction and Extraposition in German*, 1-44. Amsterdam: Benjamins.

Müller G. (1995) *A-bar Syntax. A Study in Movement Types.* Berlin: de Gruyter.

Newmeyer F. (1983) *Grammatical Theory: Its limits and possibilities.* Chicago: University of Chicago Press.

Newmeyer F. (2003) Grammar is grammar and usage is usage. *Language 79*, 682-707.

Pechmann T., Uszkoreit H.; Engelkamp J. & Zerbst D. (1996) Wortstellung im deutschen Mittelfeld: Linguistische Theorie und psycholinguistische Evidenz. In: Habel C., Kanngießer S. & Rickheit G. (eds) *Perspektiven der Kognitiven Linguistik*, 257 - 299. Wiesbaden: Westdeutscher Verlag.

Penke M. & Rosenbach A. (2004) *What Counts as Evidence in Linguistics? The case of innateness.* Studies in Language 28:3. Amsterdam: Benjamins.

Pollard C. & Sag I. (1994) *Head-driven Phrase Structure Grammar.* Chicago: University of Chicago Press.

Poulton E.C. 1989 *Bias in Quantifying Judgments.* Hove & London: Erlbaum.

Primus B. (1987) *Grammatische Hierarchien. Eine Beschreibung und Erklärung von Regularitäten des Deutschen ohne grammatische Relationen.* München: Fink.

Quirk R. & Svartvik J. (1966) *Investigating linguistic acceptability.* London: Mouton.

Quirk R., Greenbaum S., Leech G. & Svartvik J. (1985) *A Comprehensive Grammar of English.* London: Longman.

Radford A. (1988) *Transformational Grammar.* Cambridge: CUP.

Roeper T. (1982) Review of Halle, Bresnan and Miller (eds) Linguistic Theory and Psychological Reality. *Language 52*, 467-470.

Schmid T., Bader M. & Bayer J. (2005) Coherence effects: An experimental approach. In: Kepser S. & Reis M. (eds) *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, 435-56. Berlin: de Gruyter.

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

Stechow A. von & Sternefeld W. (1989) *Bausteine syntaktischen Wissens.* Opladen: Westdeutscher Verlag.

Uszkoreit H. (1987) *Word Order and Constituent Structure in German.* CLSI Lecture notes no.8, Stanford: CSLI.

Wasow T., Jaeger F. & Orr D. (2005) Lexical variation in relativizer frequency. Paper at DGfS annual conference 2005, Cologne.

Winkler S. (2005) *Ellipsis and Focus in Generative Grammar.* Berlin: de Gruyter

Wurmbrand S. (2001) *Infinitives: Restructuring and Clause Structure.* Berlin: de Gruyter.