Thermometer judgements as linguistic evidence

Sam Featherston, Tübingen University*

July 6, 2007

Abstract

In this paper I describe the method that we use to gather introspective judgements in an experimental context. We start by briefly outlining the rationale of gathering judgements at all, given that this data type has been questioned so often in the literature. We then describe our own preferred methodology, contrasting it with other common approaches and pointing out its advantages. We report a small-scale study on embedded wh-questions with complementizers such as *Mary would like to know who that Jackie will take to the party* to illustrate these points.

1 Introduction

The world of linguistics is changed, changed utterly. The quantity and evidential quality of data now available to linguists exceeds the wildest dreams of previous generations. We may discern two main developments, both of which are dependent on the rise of the computer. The first is the growth of corpus linguistics: the speed of data processing and the limits of data storage have been revolutionized over the last twenty years, which has meant that linguists have access to enormous records of language use and the processing power to analyse them in detail. It has consequently become entirely mainstream to use corpus data, since no special knowledge, privileged access, or laborious collection is required. And corpus data is authentic, documented

*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. Responsibility for weaknesses and errors rests of course with me alone.

evidence about real language use, which allows us to make wider generalizations about what people really say and write.

At the same time, there has been a clear growth in interest in the psycholinguistics of language. This has no doubt also been driven by the development of the computer, since this branch of linguistics uses computers too in order to gather evidence about the time course of language use, right down into the range of milliseconds. The result is that we can see far more exactly how speakers and listeners apply their knowledge of language when they use it. This data type additionally offers us insights into the way the our minds process language, surely a valuable perspective.

With this wealth of rich data available, it may seem surprising that any linguists would continue to take an interest in introspective judgements, that step-child of language data, which has often been controversial and often criticized. Yet it is this data type that this article is concerned with, and I hope to persuade the reader that judgements can be objective, quantifiable data and that they can be valuable linguistic evidence and even have certain advantages over other data types.

In the first section I will briefly outline the experimental approach which we adopt to gather judgements and describe our own methodology, *thermometer judgements*, which offers a good combination of ease of use, validity, and result detail. In the second part of this paper I will sketch the sort of result pattern that this data type delivers, and try to show that it offers certain analytical and explanatory advantages.

2 Judgements: problems and solutions

A number of articles have been written whose main content has been an attack upon judgements as a data type (Labov 1972, 1975, 1996, Ringen 1977, Sampson 1975, 2001). These articles have in the most part made sensible and valid but critical points about judgements as a data type and their value. Schütze (1996) reviews the literature and Newmeyer (1983) argues for the value of judgement data and models based upon it. Space does not allow us to discuss this in detail here, but we shall just go over the three main criticisms made of judgements and answer them from the perspective of the linguist who studies judgements gathered experimentally.

The first reproach made about judgements is that they are insufficiently precise. Different people give different judgements and even the judgements of an individual may vary on different occasions. It has also been pointed out that if you ask people what they say, their answers do not correspond to the externally observed facts. The second complaint about judgements is

that they are fundamentally subjective and thus unsuitable as a data base for scientific study. Furthermore, because they are subjective impressions, they cannot easily be quantified or meaningfully compared across informants; we cannot know my judgement of 'good' is the same as your judgement of 'good'. The third alleged weakness of judgements is that we do not have a full understanding of what they measure. The process of judging is not clearly identifiable as directly related to speaking or listening, which makes it a rather marginal phenomenon, sometimes referred to as 'metalinguistic competence'. Additionally, a range of factors such as plausibility, context, and lexical choice can affect introspection, which makes it appear to be a motley collection of effects.

These are the disadvantages of judgements. However, I shall argue that none of these disadvantages are in fact as serious as they might seem and most of them can be fairly readily corrected. The key to this is the collection of judgements using standard experimental techniques and controls (Schütze 1996, Cowart 1997). The single most important step is to gather the judgements of groups of informants (usually 25 -30). Labov himself says in his critique (1972, 109) "We find again and again that the grammar of a speech community is more regular than the behavior of the individual". This is perfectly true. When we collect the judgements of multiple informants, we find that they agree quite closely. Many of the cases of disagreement in judgements advanced as evidence of their unreliability (eg Labov 1975, 1996) disappear when the mean of a group is taken as the fixed point from which to measure variance. Each individual act of judgement does indeed contain some variability, perhaps idiosyncratic, perhaps random. But the variation is always fairly minor and always around a mean value; this becomes clear as soon as we look at the judgements of a group.

The collection of the judgements of groups of informants also addresses the issue of subjectivity. When a group judges structure A to be better than structure B, each individual's introspection may be subjective, but the totality of the group result is at very least inter-subjective; it has real existence outside the consciousness of any individual. It therefore represents an appropriate object of study for science, since it is a replicable fact about the world. We can predict that any other group of informants will behave the same way. This also has the favourable methodological effect of removing the data from the linguist themself, thus ensuring impartiality.

Another important step is to gather relative judgements, not absolute binary judgements. We do this by giving informants a multi-point scale on which to locate their intuitions of well-formedness; we do not just give them a binary choice. There are good precedents for this: Quirk and Svartik for example call the binary model of acceptability "absurdly gross" (Quirk &

Svartik 1966; 49). Most of the claims of unreliability in judgements relate to cases where one group of linguists say that a structure is grammatical, and another group says that it is ungrammatical. The prima facie disagreement disappears when we realize that the structure was in fact marginal. The inconsistency was only apparent: a product of an imposed binary scale which experience has shown to be inadequate. The choice of a multi-point response scale is thus an important empirical issue.

We call introspective data given on a multi-point scale *relative judgements*. When using such a scale, it is very natural to have the judgements expressed in numerical form, especially since this has additional independent advantages. For instance, this enables these judgements to be processed with standard statistical tools. Group means and standard deviations can be calculated, which permits the degree of variance to be captured and statistical tests of the significance of the effects in the results to be applied. Both of these greatly add to the evidential value of the data.

We have argued that fairly simple methodological steps can make judgements into a 'hard' data type. The question what precisely introspective judgements are measuring cannot be quite so easily dealt with. Labov complains "The data [...] may be cited as evidence of some underlying construct such as a linguistic rule, when it may in fact be [...] the product of many factors and represent no single factor at all" (Labov 1972, 104). His comment is quite true, but relates more the the naive use of judgement data than to the data type itself. Nevertheless, we should consider carefully what judgements quantify, put differently, what judgements are judgements of.

We shall not be able to give a exact answer, but realistically, judgements share this disadvantage with (probably) all other linguistic data types. We find that all language data is inherently complex and reveals information about (the) language only by implication and in comparison. If we measure frequency, for example, we can state clearly that we are measuring how often a particular structure, lexical item, or combination S occurs in a given sample of naturally occurring language use C. But what does the ascertainment that S occurs n times per million word forms in C mean for the language? Nothing, until certain assumptions and analytical conventions are applied to it. And how representative a sample is C? The questions are numerous and the conclusions always indirect . The same can be said for processing data such as reading speed or eye movements. Language data requires assumptions and interpretation before it can become meaningful. Judgement data is no different here.

The results of judgement studies also closely resemble the findings from other data types. There is an overwhelming correspondence between what is judged to be well-formed and what speakers choose to produce, for example.

Similarly, structures which are judged to be acceptable are generally easier for the parser to deal with than structures which are judged to be iunacceptable We may draw a minimum conclusion: judgement data is not measuring anything so very different from what frequencies and processing data relate to. We can usually only draw inferences about the language from differences between judgements (or frequencies, or processing load) when these differences co-vary with linguistic variables. Our evidence is thus always inferential, and judgements are no different from other data.

So much for the minimal position. But in fact we can make more positive statements. When we gather judgements with experimental controls, we are excluding or reducing the influence of certain factors on the results; we can therefore list those things which we are *not* measuring. If we elicit response from a reasonable group of informants, we can say that the data which we are gathering is a generalization over speakers. If we use multiple lexical forms of our experimental materials, we can say that our data is a generalization over the structure. We are not measuring all those factors which are held constant in the experimental design. These will typically include lexical frequency, word length, plausibility, context, individual differences, communicative aim, social variables and so on. The power of the experimental method comes from its exclusion of irrelevant factors: we can be sure that the effects we find are related to the variables that we manipulate in the design.

The nature of the task can also sharpen our grasp upon what we are measuring. In the articles I cited above criticizing judgements, one strand of evidence for the claimed unreliability of judgements was the outcome of studies in which informants were asked whether they found certain structures acceptable. The critics of judgements then follow this with observational data which reveals that the informants' reported judgements of acceptability do not correspond with their observed use. This mismatch is argued to show that we should not trust judgements. However, these claims establish little because they relate to studies asking informants, at least implicitly, about their own production. It will be clear that normative and social prestige factors may then distort results. Relative judgements should be gathered explicitly as responses to the production of others. This effectively prevents the normative or socio-dialectal status of expressions affecting the results.

An example of this is the study quantifying speakers' use of parting greetings, reported by Labov (1996). When asked which parting greetings they used, speakers apparently rejected the use of *bye-bye*, saying that it was only for babies. Observation however revealed that *bye-bye* was the single most common expression used. Labov cites this as evidence that we cannot trust judgement data. We shall make two comments. The first concerns the linguistic basis; the study fails to distinguish baby *bye-bye* in which both

syllables are given their full value, and adult *b-bye*, in which the first vowel is a schwa. This is a crucial mistake since the first is indeed childish, while the second is common in adult contexts. It is apparent that the researchers obtained opinions about the use of the first but observed the use of the second; so the failing here is in the linguistic analysis, not in the data type.[1]

It is also important to be more specific what we mean by 'judgements'. The introspective judgements we gather are not just informants' opinions of anything, but very specific linguistic intuitions of well-formedness. One cannot expect any useful data to emerge from asking informants to give judgements of frequency or use, since this information is not part of the speaker's sprachgefühl. Labov (1996) is right when he notes that we cannot reliably give these 'judgements', but he is wrong is classing these opinions together with the introspective data that we collect, which is just unconscious intuitions of difficulty, not conscious or reflected knowledge. The criterial question which we put to informants is: *How natural does this example sound?* This has the double advantage of highlighting the receptive aspect and the speaking mode and, crucially, avoids confusing or leading informants with association-laden terms such as 'grammatical'.

We shall refer to the construct measured with relative judgements as *perceived well-formedness*. Although we cannot precisely say what mental processes it quantifies, we can rule out certain factors and we can make certain positive statements. Introspective judgements yield something like a measure of how difficult informants find it to process, analyze the structure of, and compute the meaning of an input sentence. We shall see below in the presentation of standard pattern of results that there is quite strong evidence for the view that judgements measure the ccomputationalcost of assigning a structural representation to a string and associating it with a matching interpretation. Relative judgements are thus most closely related to processing cost and are thus at base a psycholinguistic variable.

We shall finish this section by highlighting a particular of judgements: they are the syntactician's traditional data type. Intuitions have always been used as one of the major ways of deciding what examples are and are not grammatical. While the method of gathering these intuitions has been greatly improved, the data type itself is very familiar. We are therefore collecting the traditional data type, but with finer differentiation. To the question 'What is your experiment measuring?', we can thus always answer 'The same as judgements have always measured, but in greater detail.'

[1] In the first edition of the CELEX lexical database of German one form of the verb *heuen* ('to make hay'), the 1st and 3rd person simple past form *heute*, was recorded as astonishingly frequent. Closer inspection revealed that this was because it was a homograph with *heute* ('today'). This is not a sufficient basis to reject databases as a data type.

3 Methodology: thermometer judgements

There are several ways of collecting relative judgements, but they have more in common than they have differences. Traditionally a discrete point scale with five or seven values has often been used. More recently magnitude estimation has been quite popular (Bard et al 1996). This method consists of standard judgement elicitation on a slightly more sophisticated scale. First, judgements are given relative to a reference item, which provides an anchor, and to their own previous judgements. Informants thus create their own coherent scale while judging, but there are no absolute values, and the scores are relative just to each other. Second, since judgements are given numerically, on an open-ended scale with no minimum division, informants can express all the differences in well-formedness they perceive with no constraint from an imposed scale. These relative judgements form an interval scale and can be tested by analyses of variance; the data yields robustly significant results.

We make two changes to this methodology, which yield what we have dubbed *thermometer judgements*. The first change is in the scale. In classic magnitude estimation judgements are given relative to a single reference item. So if a subsequent example is twice as good as the reference example, informants are told to give a score twice as large. If it is only half as good, it should receive half the score. The idea for this was derived by Bard et al (1996) from the psychophysical studies of Stevens (1975). However, our experience has shown that informants do not in fact do this; instead they use a simple linear scale, in which distances above and below the reference score are given numerically equal values. It is clear why they do this: in order to produce scores in multiples, informants would have to have access to a zero value, but in a continuum of well-formedness there is no zero value. They cannot therefore estimate what 'twice as good' might mean.

Scales of perception have been much debated in ppsychology there are a range of factors which can affect judgement patterns, as Poulton (1989) and Laming (1986) discuss at length, but for linguists, the important issue is to gather judgements as simply as possible, on the scale that our informants in practice use. In our thermometer judgements method, we therefore omit the instructions telling informants to use a magnitude scale. This has no apparent effect upon the results, since informants did not in practice use the magnitude scale even when instructed to, but it seems more valid to adapt the instructions to the informants' abilities.

The other change we make relative to magnitude estimation is to supply two reference items instead of one. The reason for this is the distortion which tends to occur around the value zero. In theory, in magnitude estimation all scores should be positive, though decimals below one are permitted. In

practice, when giving scores near zero, informants tend to simplify or stick to integers. In thermometer judgements we therefore present the informants with two reference items, one quite bad and one quite good, and assign them the values 20 and 30. This sets both the location and amplitude of the judgement scale in such a way that no informant has the need to approach zero. The method thus has a scale with two fixed points anchored to known values, which define the scale in the same way that the 0^oC and 100^oC points, linked respectively to the freezing and boiling points of water, fix the Celsius scale. The name *thermometer judgements* reflects this similarity.

The thermometer judgements scale thus fairly closely resembles the simple five- or seven-point scales we mentioned above, but retains the flexibility of open-endedness and no minimum division from the magnitude estimation scale. No assumptions are made about the data pattern, and it keeps the the advantage of being an interval scale for statistical analysis. It is necessary to point out to informants that they are entitled to use values between, above, and below the two reference values, but this is very naturally done as part of the practice phase. Here is a simplified example of what participants read:

(1) This example is worth 20.
The father fetches food and drink the child.

This example is worth 30.
The father fetches for the child the food and drink.

If the first one is worth 20 and the second one 30, how much would you give this one?
The father fetches the child the food and drink.

The method has proved itself to be very effective since, on the one hand, studies can be very exactly focused on just the issue of interest, but the approach delivers very finely differentiated results.

4 Evidential value

In the first part of this article I have laid out some details of the nature and collection of the data type of relative judgements, and I hope to have convinced the reader that data collected in this way is sufficiently objective and quantifiable to count as scientific evidence. Our second aim is to persuade the reader that this data is linguistically interesting, that is, that it can provide rresearcherswith productive insights into the nature of language. In fact I shall suggest that relative judgements have certain advantages over other data types in certain contexts.

4.1 What data patterns relative judgements show

Let us first examine the typical data pattern that this data type reveals.

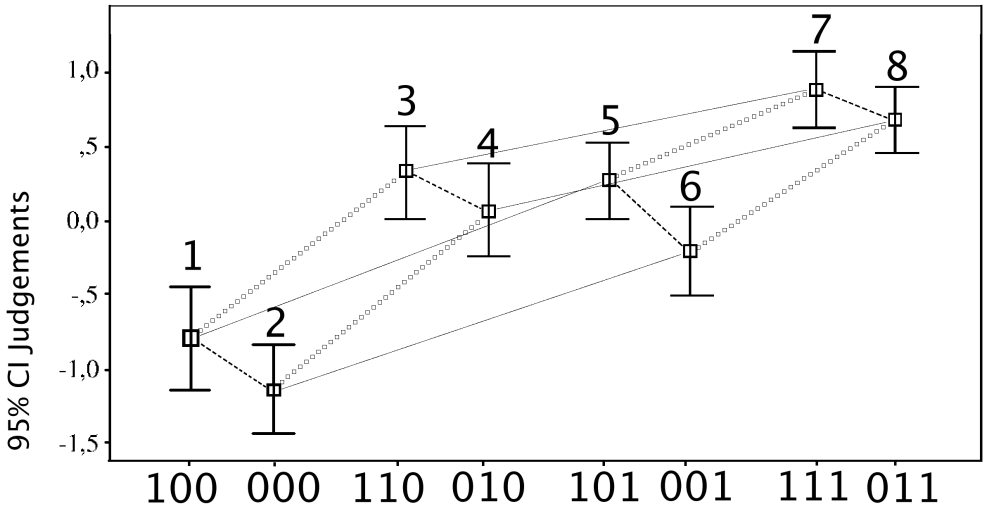

Figure 1: This figure shows typical features of the results of judgement experiments. The y-axis shows the normalized judgements, the x-axis the eight conditions, made up of three parameters, each with two values. The pattern of results shows that constraint violation costs are automatic, consistent, quantifiable and cumulative.

Figure 1 shows the sort of results that we get from experimental judgements. The values of the informants have been normalized so that they all have the same mean point and standard deviation: in the chart higher scores on the y-axis indicate judgements that the structures are more 'natural', but there is no absolute criterion. The figure 0.00 merely indicates the mean point of all scores. We tested eight syntactic conditions here, which were derived from three binary parameters. The scores of each condition are expressed with an error bar, in which the mean value is shown by a square marker and the 95% confidence interval is shown by the length of the bar. The shorter the bar, the less variation there was in the scores given to the condition.

Since we are looking at a typical result to observe the normal sort of pattern obtained, not a specific one, we have just assigned the conditions numbers 1 - 8. Each condition has a value, 0 or 1, for each of the three binary parameters. Each condition has its values on the three parameters marked on the x-axis. We can think of the parameters as constraints on well-formedness, so a zero on the x-axis implies that a structure has violated a constraint; a 1 indicates that it has fulfilled the constraint. Condition 7

thus fulfils all three constraints (shown by *111*), condition 3 on the other hand fulfils the first two constraints but violates the third (so *110*).

It will be apparent that conditions 3 and 7 are thus a minimal pair. Conditions 4 and 8 are also a minimal pair, the only difference from 3 and 7 being that both 4 and 8 violate the first constraint, whereas 3 and 7 fulfil it. Conditions 1 and 5, and 2 and 6 are similarly minimal pairs varying only on the same parameter, whereas 1 and 2, 3 and 4, 5 and 6, and 7 and 8 are minimal pairs which vary on the first parameter. In fact all the pairs of conditions joined by lines in the graphic are minimal pairs, in which always one pair member fulfils a constraint, and one pair member violates it. Having clarified what information the chart shows us, we are now ready to discuss the data pattern (building on the work of Schütze 1996, Cowart 1997, and Keller 2000).

The first thing that we find in results of studies gathering relative judgements is that the effect of a violation is quantifiable. If a structure violates a given constraint, then it is judged worse by a certain extent. Different constraints have different violation costs, some greater, some less, but the effect is consistent. Whether the structure is otherwise good (like number 7) or otherwise bad (like 1) the effect is about the same; thus 8 is as much worse than 7, as 2 is worse than 1. We thus identify constraint-specific violation costs. Furthermore, we see that, given that the effects of constraint violations are consistent, the application of constraints to structures must be automatic. It is not the case that constraints apply or do not apply to structures depending upon some external factor, such as whether the structure is otherwise good or bad. Lastly, let us note that these constraint violation costs are cumulative. The drop in perceived well-formedness between, for instance, condition 3, which satisfies the first two constraints, and condition 2, which violates the first two constraints, is exactly the separate violation costs of these two constraints, added together.

The data from experimentally obtained relative judgements thus shows us a lot about the way that constraints on structure interact, and the results of violations of these constraints. Constraints are applied automatically to all structures, blindly and without exceptions. Any violation costs incurred are applied consistently, without any reference to the status of the structure. These violation costs vary according to the constraint violated, but for any given constraint type, all other things being equal, the effect is consistent. Lastly, these constraints are cumulative. This is what we always find when we gather this data type.

It is perhaps also worth noting explicitly what we do not find. We find no sign of any binary division between 'grammatical' and 'ungrammatical', nor do we find any sign that syntactic constraints are ordered or applied

conditionally on external factors such as the well-formedness of a structure or other structures in its potential comparison set. Well-formedness thus seems to be absolute, but not binary. We also see no sign of grammaticality status being variable, depending on random factors (such as stochastic theories would suggest): the status of each condition seems to be inherent and determined by the structure itself, not dependent upon probabilistic factors which might alter outcomes in separate cases of derivation or competition.

Let us underline that these findings are common to all sets of judgement data obtained in a sufficiently controlled manner to deliver this degree of detail: they are not methodology dependent, but rather a constant of the data type relative judgements. This is worth saying, because these findings are of real importance to our understanding of the way that the 'grammar' works. In fact these findings directly contradict the assumptions of certain models of the way that constraints interact in determining which structures are well-formed and ill-formed. This makes them potentially of great empirical and theoretical importance.

4.2 Judgements vs frequency data

But if relative judgements show such an unambiguous pattern, why is it that linguists have not generally taken this on board? Why would linguists entertain theories which assume other interactions of constraints? The answer seems to lie in the technological advance of judgements and frequencies. Before the coming of the personal computer, linguists used individuals' judgements and corpus frequencies, but both of these were only possible on a small scale, and so little detail was available. The coming of computer technology made above all the collection, storage, and interpretation of corpuses much easier, so those linguists who were interested in data-based linguistics turned most commonly to frequency data as their evidential base. The interest was such that the field boomed, and descriptive and explanatory generalizations were made on the basis of frequency data. This quite reasonably lead to the de facto criterion: what occurs is what is part of the language, and so the specificities of occurrence data fed into theory.

Only more recently have empirically interested linguists treated judgements with equivalent attention. The results are as we illustrated above. It becomes apparent that ooccurrencedata, such as that from frequency studies, has certain characteristics which affect the picture of the grammar that we gain from it. Frequency data disposes us to look only at those structures which actually occur, and can lead us to think that all those structures which do not occur are just 'bad', perhaps 'all equally bad'. Data which necessarily includes a fundamental binary division, between 'occurs' and 'does not

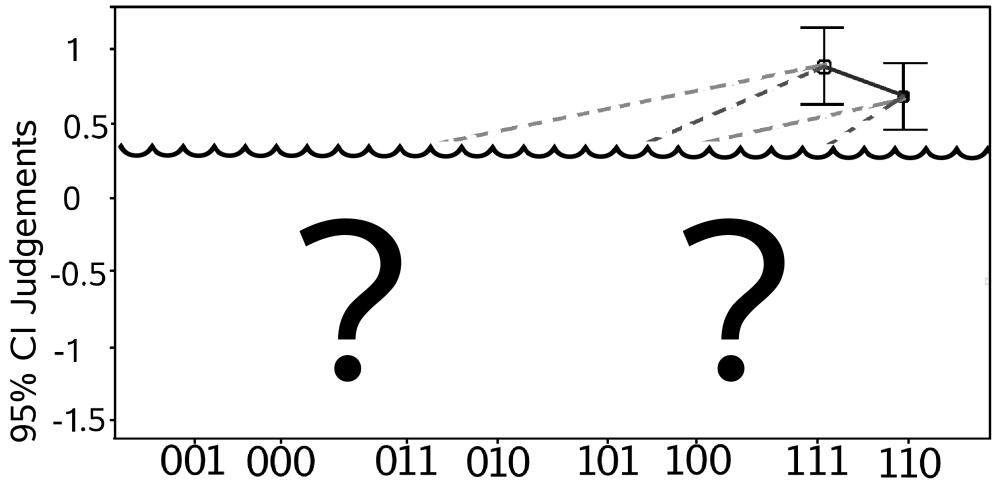

Figure 2: The same data set as in Figure 1, as viewed in frequency data. Measures of occurrence can tell us nothing about structures which do not occur. They only show us the part of the iceberg which projects above the water level.

occur', will naturally tend to lead us to assume that this is a central feature of a the system which produces the data. But this binary opposition is due to a restriction in the range of vision of frequency data, due to what I call the 'iceberg phenomenon'. Consider the data in Figure 2.

Here we present exactly the same data as in Figure 1, but with only those variants which occur presented. The implication is clear: while frequency information can provide us with high quality evidence about what occurs and about the differences between occurring variants, it can tell us little about why certain structures occur and others do not. We can generalize this further and say that frequency data is not the ideal data type for investigating how constraints interact, since normally speaking just the best variant will occur, and the others hardly at all. Frequency data can only reveal much evidence about the interaction of variants when either a) all the violation costs are very small, or b) violation costs of contributory constraints are about equal, so that the variants are still all about as good as each other and can all occur. Our relative judgements on the other hand can access the status of structures which are quite impossible and would never occur. For the architecture of the grammar, this is essential information, as it allows us to quantify the effects of constraints whose violation costs are normally too great for violating structures to occur. Relative judgements, we might say, allow us to see the shape of an iceberg below the water line.[2] When

[2]An example of this second type is the competition between objects and particles for

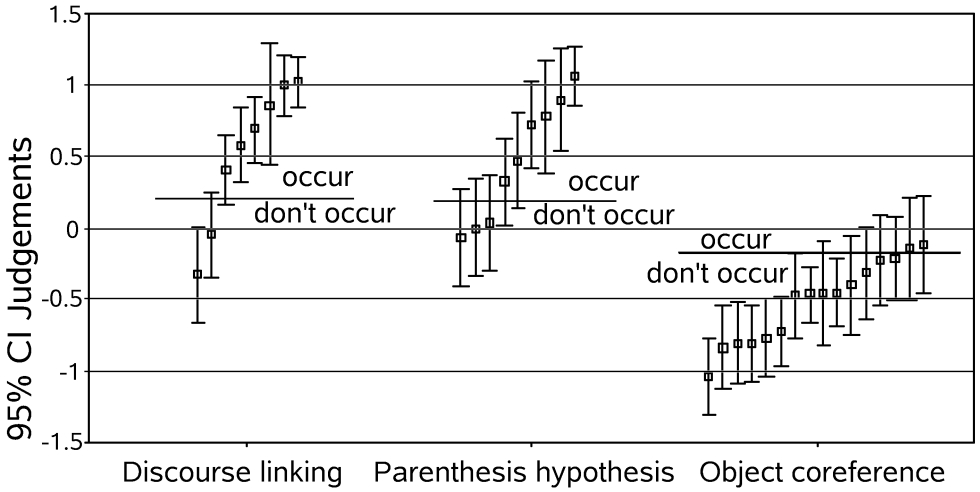

Figure 3: In this study we carried out three sub-studies in one experiment. The results show that the pattern of data we obtain above the occurrence threshold continues below it.

we look below the water line, it turns out that what we see is not what the binary model of well-formedness would lead us to expect. As we see in Figure 3 above, the pattern we see is a continuum, above and below the occurrence boundary. With relative judgements we can measure the violation costs of constraints whose violation costs normally prevent the occurrence of a structure which violates it. Measures of occurrence do not do this.

This therefore is a strength of relative judgements as a data type. It allows us to measure the effects of violations across the whole band of structural well-formedness and across the whole range of violation cost amplitude. This is important, for the patterns we find determine the answers to key questions

---

the position immediately after the verb in English (Wasow 2002). Although both most naturally occur in this position, in sentences in which both are instantiated, one of them must make way. Thus both (i-a) and (i-b) are possible, and so frequency data can tell us about them, and to an extent about the way the causal factors interact.

(i) a. I looked the word up in a dictionary.
b. I looked up the word in a dictionary.

But the equilibrium is very precarious. Even a moderately minor change such as the NP type or 'weight' of the object can make one option impossible. Occurrence data can tell us nothing about the strength of the constraint which causes this.

(ii) a. I looked it up in a dictionary.
b. *I looked up it in a dictionary.

about the way that constraint violations interact in the grammar and the causal factors which result in a particular variant being more well-formed than all other variants. One might summarize thus: if we want to know what structures speakers use, we should gather and interpret occurrence data; but if we want to know why a particular form occurs, then we should gather relative judgements, for these allow us to measure violation costs.

Let us make very clear that this should not be read as a criticism of frequency data per se - all syntactic data types can only show us a partial picture of what happens inside the black box. On the contrary, I am arguing that all empirical findings and assumptions should be checked across data types. We try to confirm our own experimental findings with corpus frequencies whenever possible and would immediately start to look for the error if the frequencies and judgement results were not compatible with each other. The overarching lesson must be that it is always necessary to look at multiple data types, since all data types have their particular gaps and slants, and no single data type provides us with a full picture.

5 *Who that* in embedded wh-questions

One of the very useful features of experimentally obtained judgements is the ability to obtain fine data detail in phenomena which are only marginal in the language. In order to demonstrate this I shall present a small study here which I inserted into a larger experiment as filler material. The issue concerned is the status of *that* after a wh-item in an embedded question. While certain languages permit this (Radford 1988 p486f mentions Canadian French, Middle English, Dutch; Radford 1997 p271, 302 discusses Hibernian English from Belfast citing Henry 1995 p107), standard English normally does not. Bayer (1984) treats this issue at length in Bavarian, where it is possible, comparing it with standard German, which does not allow it. Richter & Sailer (2000) provide a account of this in German in a head-driven phrase structure grammar framework (Pollard & Sag 1994), treating it as lexical variation.

(2) Mary would like to know ...

a. ... who Jackie will take to the party.
b. ... who will take Jackie to the party.
c. *... who that Jackie will take to the party.
d. *... who that will take Jackie to the party.

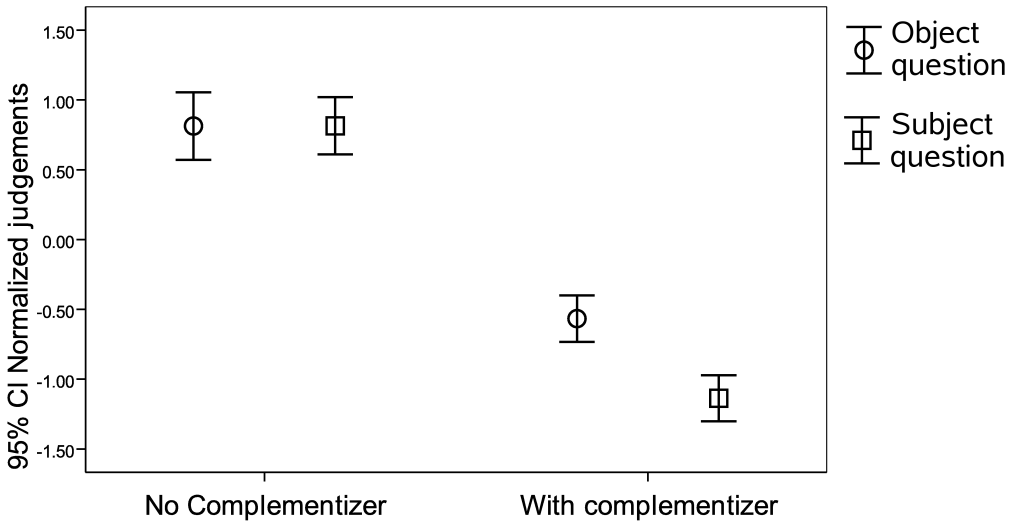

Figure 4: Well-formed embedded wh-questions exhibit no subject - object asymmetry. The versions with *that* after the wh-item show a clear difference.

To my knowledge however, no author has noted a subject - object asymmetry, which is apparent in English. Both (2-a) and (2-b) are fully acceptable, but neither (2-c) nor (2-d) are part of standard English. My own intuitions tell me however that (2-c) is clearly better than (2-d). If this were confirmed, then it would provide useful clues as to the nature of the phenomenon and the most appropriate syntactic analysis. There are known subject - object asymmetries in structures involving extraction over complementizers such as the *that*-trace effect. We may be fairly certain that this represents an aspect of the core syntax since we have replicated it in German (Featherston 2003).

We therefore tested this phenomenon in a small-scale study with just the four conditions in (2), each appearing in eight different lexical forms. Thirty-one informants recruited by flyer at the University of Lancaster took part in our experiment.[3] They logged themselves into to our server and carried out the experiment remotely, which we ran on the software package WebExp. The experiment was carried out using the thermometer judgements procedure detailed above.

The results in Figure 4 show that our participants strongly rejected the versions with *who that*, which is not surprising. But the participants did not perceive these ill-formed structures as being equally bad. The subject extraction is plainly judged even worse than the object extraction, in line with my own intuitions. These are relative judgements, but we included some

[3]The data of one participant was discarded because of doubts about its quality.

fully acceptable and fully unacceptable fillers to act as an absolute standard. The subject extraction was judged just as bad as the clearly unacceptable examples such as *What Mandy want read to her sister next?*, while the object extraction was visibly better. There is no sign of a similar difference between the extractions without *that*.[4]

This result is a simple observation, yet it is linguistically interesting for a number of reasons. First, because it provides useful evidence how we should analyze this structure. This finding immediately reminds us of the *that*-trace effect which produces almost identical results (Cowart 1997, see also Featherston 2003). In exactly those cases when a subject has been 'moved' out of a subordinate clause headed by *that*, that is (2-d) and (3-d), the structure suffers reduced perceived well-formedness.

(3) a. Who does Mary think Jackie will take to the party?
b. Who does Mary think will take Jackie to the party?
c. Who does Mary think that Jackie will take to the party?
d. *Who does Mary think that will take Jackie to the party.

To show how clear the correspondence to the *that*-trace phenomenon is, we include here in 5 the results of Cowart's (1997) work on it in English and our own findings for the nearest equivalent structures in German (Featherston 2003). In each case, the extraction over a complementizer is judged worse than if there is no complementizer, but the subject extraction over a complementizer is worse than the object extraction. This difference is not found in the structures without complementizers.

A fuller analysis of the syntax will have to wait for another occasion, but it will be clear that the resemblance between the otherwise unpredicted findings in the two closely related structures is very persuasive. Whatever aspect of the movement of a subject wh-item over a complementizer proves to be the ultimate cause of the *that*-trace effect, it does seem very likely that this same factor will be the cause of the effect that we find here. This finding is thus a contribution towards an explanation of the still poorly understood interaction of subject wh-items and complementizers; when we can delimit a phenomenon we are closer to explaining it.

Our finding can also help us to select between competing accounts. For instance this evidence would tend to defeat the analysis of the embedded

[4]The key statistical test of this finding is the interaction of the question type (subject, object) and the presence or absence of the complementizer. In the repeated measures analysis of variance the interaction was significant by subjects (F_1 (1,29) = 7.604, p_1 = 0.010) but only approached significance by items (F_2 (1,7) = 3.203, p_2 = 0.117). This was due to an experimental error; there is no doubt that the items analysis will be fully significant in a full-scale study.

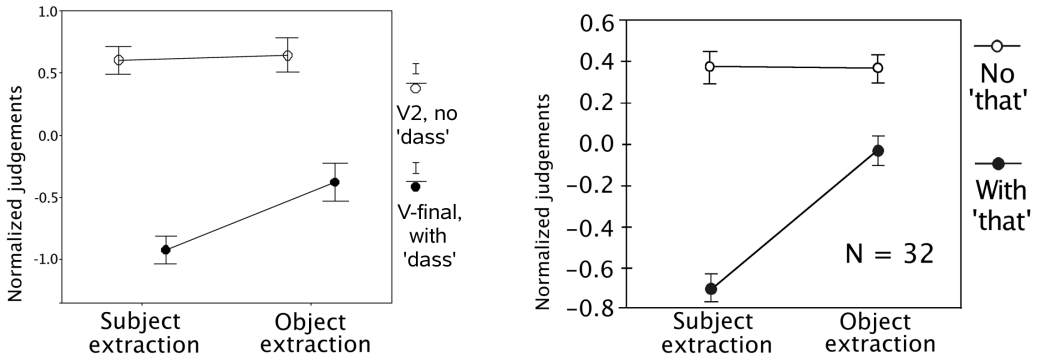

Figure 5: The left-hand picture illustrates the *that*-trace effect in English (adapted from Cowart 1997). The right-chart shows the same pattern of judgements in the German closest equivalents (Featherston 2003).

wh-questions with complementizers (2) in which the *that* is merely a lexical extension of the wh-item. The fact that we find the same effect in the *that*-trace phenomenon, when the two are not adjacent and thus cannot possibly be single lexical unit, would seem to rule this lexical account of wh-item plus *that* out. This claim has been made for German: in the analysis of Richter & Sailer (2000), the combination wh-item plus *dass*, an apparent complementizer, is a single unit. Their claim is that the left periphery of German finite clauses contains just a single position, in contrast with the standard assumption that there are two positions. This account would be undermined, if we were to find the same subject - object asymmetry in German embedded questions as we found in English, since the *that*-trace effect is active in German too 5. We hope to test whether this effect is detectable in German in the near future.

The wider issue here is the relevance and interest of the data type. We hope to have demonstrated that relative judgements can be informative: with a simple study collecting the introspective judgements of naive informants we have here made a useful contribution towards the study of an outstanding issue in the syntax. Judgement studies can be easily targeted at specific questions: we can gather relative judgements of this phenomenon even though the crucial structures are ungrammatical. This feature is a major strength of this type of linguistic evidence. What is more, when we ask about structures which would not occur, informants are readily able to distinguish between varying degrees of ill-formedness. This point is theoretically important. It is widely overlooked in the design of syntactic models that it is insufficient for a grammar to specify what structures are and which are not part of the language, an empirically adequate model must also distinguish different de-

grees of well-formedness and ill-formedness. This has profound implications for the nature of the grammar. Again, we would argue, this data type is showing its value.

6 Conclusions

In this paper I aimed to convince the reader that judgements could be objective, quantifiable data and provide valuable linguistic evidence. For the first proposition, the core of my argument was that judgements are 'hard' if (and only if) we gather them using standard experimental methods, controlling for irrelevant factors whenever possible, and eliminating variability caused by individual differences and lexical specificities by testing groups of informants and using multiple lexical variants of our sentence material. These fairly simple measures raise introspective judgements into an objective scientific measure which is independent of the linguist as experimenter. Results collected with our own *thermometer judgements* method, for example, allow empirical predictions to be made which can be tested and falsified, on the one hand, or replicated, on the other. I therefore see little room for doubt that these judgements are valid linguistic evidence.

One question to which we have only been able to give a partial reply is that of what we measuring. The range of factors which can influence judgements gives the impression that *relative judgements* are measuring something like the effort involved in analyzing and assigning a meaning to an input structure. It would thus appear to be a psycholinguistic construct. While we can be no more accurate than that, we would note that most other linguistic data types suffer from similar indistinctness. Processing speed, for example, as measured in reading times, is clearly a composite of many different factors, some more and some less directly of relevance to linguistics. The occurrence of a structure is similarly not direct evidence of well-formedness any more than non-occurrence is direct evidence of ill-formedness. Judgement data too is no doubt never more than indirect evidence of linguist constructs, but it shares this feature with other data types and it may in fact be more primary evidence than some other types.

In the second part of this paper I tried to make clear why I find relative judgements to be linguistically interesting. The pattern of data which we regularly find is an important criterion. This shows that the interaction of the formants of well-formedness is cumulative (Keller 2000), and that violation costs are consistent and automatic, contrary to the assumptions underlying some grammatical frameworks. The detailed picture of the factors which influence well-formedness provide a very useful tool for the investigation of

why a structure is good or bad. Since this data type makes available information about the quantitive effects of constraints whose violation is normally incompatible with occurrence, it clearly scores here over data types which depend upon occurrence, such as frequency information.

The small study of embedded questions with complementizers illustrates this. We are able to gather data about a structure which is absent from modern standard English, revealing a subject - object asymmetry which, to our knowledge, has never previously been noted. The particular finding is so similar to a known phenomenon (*that*-trace effect) in closely related structure that we find it implausible that the two effects are not driven by the same structural cause. This would exclude the possibility that *who that* is simply a lexical form of *who*, for example. This example study was chosen as an instance of how just a little information can have important theoretical implications, and how data from below the water-line of visibility, from structures which do not in fact occur can provide important linguistic data. Relative judgements can provide such information; they should thus become one part of the linguist's .

References

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

Bayer J. (1984) COMP in Bavarian syntax. *The Linguistic Review 3*, 209-74.

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

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

Henry A. (1995) *Belfast English and Standard English: Dialect Variation and Parameter Setting*. Oxford: Oxford University Press.

Keller F. (2000) Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Dissertation Univ. of Edinburgh.

Laming D. (1986) *Sensory Analysis*. London: Academic Press.

Labov W. (1972) Some principles of linguistic methodology. *Language in Society 1* (1), 97-120.

Labov W. (1975) Empirical foundations of linguistic theory. In: Austerlitz R. *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.

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

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

Poulton E. (1989) *Bias in Quantifying Judgements*. Hove: Lawrence Erlbaum.

Radford A. (1988) *Transformational Grammar*. Cambridge: Cambridge University Press.

Radford A. (1997) *Syntactic Theory and the Structure of English*. Cambridge: Cambridge University Press.

Richter F. & Sailer M. (2000) On the left periphery of German finite sentences. In: Kiss T. & Meurers D. *Constraint-based Approaches to Germanic Syntax*. Stanford: CSLI.

Ringen J. (1977) On evaluating data concerning linguistic intuition. In: Eckmann F. (ed) *Current Themes in Linguistics: Bilingualism, Experimental Linguistics, and Language Typologies*, 145-60. Washington: Hemisphere.

Quirk R. & Svartvik J. (1966) *Investigating Linguistic Acceptability*. London: Mouton.

Sampson G. (1975) The Form of Language. London: Weidenfeld & Nicholson.

Sampson G. (2001) Empirical Linguistics. London: Continuum.

Schütze C. (1996) *The Empirical Basis of Linguistics*. Chicago: University of Chicago Press.

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

Wasow T. (2002) *Postverbal Behviour*. Stanford: CSLI.