# Can we build a grammar on the basis of judgements?

Sam Featherston
University of Tübingen

5th July 2019

## 1 Introduction

This paper is intended as a contribution to the debate on introspective judgements as an appropriate basis for a grammar. I think traditional practice within the field of grammatical theory has been and in its unreformed form still is inadequate, and that more efforts are consequently needed to enable syntax to develop and make advances. Proposals for new analyses need to be more than speculation, we need to have a clear cycle of hypothesis generation, testing, and rejection for progress to be made. This is not new: more people have made this point than me, but combating dataphobia in the field of syntax is like losing weight or dealing with climate change: it is not something that happens instantaneously, it is a long-term process. The pressure needs to be kept up or else people will sink into complacency. It is partly because of this that I perceive with dismay a sort of counter-current of data quality scepticism going around. Some linguists feel themselves encouraged to pursue grammatical study with little regard to the quality of the evidential base, because some papers have been published which seem to allow this or encourage it. The aim of this paper is to respond to this trend.

One of the ways that I will do this is by discussing a couple of papers which are frequently referred to on this topic, namely Sprouse & Almeida (2012) and Sprouse, Schütze & Almeida (2013) (here: SA12 and SSA13). These are excellent articles which have been conscientiously composed and which constitute very real contributions to the field. I have no argument with their contents, but I do wish to take issue with some aspects of their interpretation. Mostly my conflict is with conclusions which are drawn from these papers but not made within the papers, though sometimes the texts do seem to invite strong claims. I will argue that these papers do not in fact legitimize data-light theory building, by taking a closer look at what the studies reported in these papers actually show. My view is that

these papers do not, on closer inspection, justify armchair linguistics, as they are sometimes taken to do.[1] In fact, I shall argue that they yield fairly strong evidence that the single judgements of the individual linguist do not form an adequate basis for further theory construction.

There is something that I need to clarify before we get under way. The title of this article asks whether we can build 'a grammar' on the basis of introspection, but in reality that is just a short form of the real question; what is fundamentally at stake is whether we can build a good grammar, a better grammar than we have up to now. Perhaps we should talk about *the* grammar, the grammar which is uniquely specified by the data. This is one way that this article asks a rather different question to those addressed in SA12 and SSA13. While these texts test whether the judgements employed in the linguistics literature were adequate as the basis for the grammar so far, I would seek to query whether they are sufficient for us to make progress in grammar research. If they are not, then I would see these judgements as sub-optimal, even if they have been a sufficient basis to get us to the point where we now are. If you are driving the Paris - Dakar rally, you need a vehicle that will get you first down through France and then across the Sahara. The fact that you can do the first 1000km on tarmac in a family car doesn't change this. Schütze (1995) relates this ambition to Chomsky's (1965) claim that linguists' own judgements are adequate for the 'clear cases'. It was true then, but theory development is quite rightly trying to move beyond the clear cases and requires more finely grained distinctions in order to continue.

I think a forward-looking perspective is necessary because our knowledge and understanding about syntax is still very sketchy. One of the clearest signs of this is that a range of different grammar models exist, with clearly divergent underlying architectures. If linguists cannot agree whether the form of the grammar is overgenerate and filter, declarative, generate and economize, or winner takes all, I cannot see that we can yet be satisfied with our knowledge. Worse still, data-light linguistics offers us little in the way of infrastructure to support progress. In armchair linguistics, analyses come and sometimes go, but others remain in spite of having little empirical support. The issue is that claims which are only weakly linked to the data basis make few testable predictions and are thus impervious to the standard scientific tests of corroboration or falsification. Experimental data types are more exact and thus go hand in hand with a more rigorous data-driven academic praxis. It's tough on linguists, because you can be proved wrong, but that is science.

In the following I first present the sort of evidence that persuades me that we should beware of armchair judgements. In this section I show some experimental

[1] Phillips (2009) talks about *armchair linguists*, and I will use the term too, contrasting 'armchair' judgements and 'experimental' judgements, because I find these terms expressive.

work done by myself and my colleagues but also refer to some of the results of SSA13 which seem to me to confirm what I am suggesting. I then move on to discuss why some people seem to think that the papers SA12 and SSA13 validate armchair judgements. Here I highlight three factors which lead me to interpret them more cautiously. I will finish up by arguing in favour of quantified judgements, because only these support more complex grammar models.

## 2 The quality of armchair judgements

There is an extensive literature on judgements which debates whether they are a suitable basis for linguistics (see extensive citations in Schütze 1995, Featherston 2007, SA12, SSA13). We can perhaps identify four main problematic features of introspective judgements as they have been traditionally used in syntactic literature.

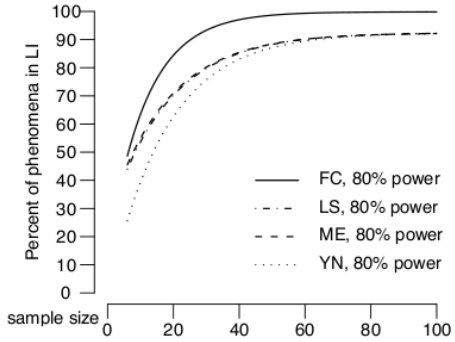
1. single judgements are coarse-grained
2. single judgements have a high noise component
3. conflict of interest of linguist as data source and interpreter
4. judgement data patterns don't match theoretical predictions

I shall chiefly address the first two here. My basic point is that these reproaches are justified in the case of armchair judgements, but less so for experimental judgements.

### 2.1 Armchair judgements are less sensitive

In order to support this claim I shall revisit the findings of a paper often held to support the value of informal judgements - SSA13. This paper is an exceptionally detailed and well performed investigation of the quality of judgements reported in *Linguistic Inquiry*. The authors have tested whether a sample of the intuitions reported can be replicated using experimental techniques, applying a range of different statistical tests. The results are scrupulously reported, which means that the reader can build their own picture of the findings, or indeed re-use their findings as I do here. Figure 1 – their Figure 9 – shows the success rates in replicating judgements from the *Linguistic Inquiry* sample in graphic and tabular form.

Figure 9: The percentage of phenomena (y-axis) in *Linguistic Inquiry* (2001-2010) that would be detectable with 80% power by each task (lines) as a function of the sample size (x-axis) assuming only 1 judgment per participant per condition.




| Sample | FC | YN | ME | LS |
|---|---|---|---|---|
| 6 | 35% | 15% | 29% | 28% |
| 10 | 69% | 39% | 58% | 57% |
| 15 | 78% | 58% | 70% | 69% |
| 20 | 83% | 64% | 74% | 74% |
| 25 | 89% | 72% | 77% | 76% |
| 30 | 91% | 75% | 79% | 79% |
| 35 | 94% | 77% | 81% | 81% |
| 40 | 95% | 80% | 82% | 82% |

Figure 1: The percentages of judgement replications in Sprouse, Schütze & Almeida (2013), on the left in graphic form and and the right in tabular form.

The authors understandably highlight the very high success rates of the experimental methods in capturing the judgements from *Linguistic Inquiry*. With sufficient participants, the forced choice method reaches a 95% replication rate, while the other methods (yes/no question, magnitude estimation, seven-point rating scale) all manage 80% or above. But here's the rub: it requires 40 informants to reach that sort of level. Below 25 participants the success rate drops off steeply. When we look at sample size 6, the lowest participant number given, the success rate is between 15% and 35%. The figures for single judgements are not given, but we can see where the lines in the graphic are tending. So this carefully processed data gives us a clear message. We need 20 or ideally more indpendent judgements to achieve satisfactory power. Fewer than 10 judgements must give a real cause for concern. Judgements by a single person are an inadequate basis for any firm conclusions.

Now we need not take these figures too exactly. All linguists who actually use judgements as a data type (unlike Labov 1996) would agree that even a single person's intuitions can under certain circumstances make quite fine distinctions. But one aspect of this data set is beyond discussion: the sharp downward curve in the success rate with fewer participants. Fewer informants undoubtedly produce less powerful results. So this is one clear way in which experimental judgements are a much better basis for grammar development than armchair judgements. Another way in which experimental judgements with multiple informants can produce better data concerns the revelation of finer differences. We present an example of this in the next section.

## 2.2 Armchair judgements are noisy

One reason why armchair judgements are less sensitive is that introspective judgements contain a relatively large noise component. I have discussed this in the past (e.g. Featherston 2007), but any linguist who has gathered judgements from groups and looked at them comparatively knows this. We will illustrate this using the patterns of data that we obtain from our standard experimental items which instantiate the so-called cardinal well-formedness values (Featherston 2009, Gerbrich et al. in press). We have developed a set of 15 sentences which are reliably judged at five different levels of well-formedness, labelled A to E, with three examples at each level. These provide a useful fixed external comparison set which allows us first, to obtain something approaching absolute well-formedness values from studies collecting relative judgements, and second, to make meaningful comparisons across experiments. The standard items provide a useful reference set here because they have been tested many times in lots of different experiments. Their levels of well-formedness are established.

In Figure 2 we show the results of the complete set, judged by 32 informants. These rating were gathered in a experiment testing aspects of NP movement using the thermometer judgements method with a two-stage practice phase (Featherston 2009). Native speaker informants were recruited via the Prolific experiment participant portal and paid £2 for their participation. There were 8 versions of the experiment, with 4 participants per version. Each person saw a total of 55 sentences in a pseudorandomized order, 15 of which were these standard items. The scores were normalized to z-scores to remove variation in the use of the scale between participants. This procedure aligns each person's scores to give them a mean value of zero and a standard deviation of 1. The y-axis on the chart thus shows these converted scores, and the zero value is the mean of all judgements. Higher scores represent greater perceived acceptability. The results are displayed as error bars showing the means and 95% confidence intervals of these normalized ratings. We choose this chart type to show that the groups clearly distinguish the cardinal well-formedness levels; the error bars of the confidence intervals do not overlap.

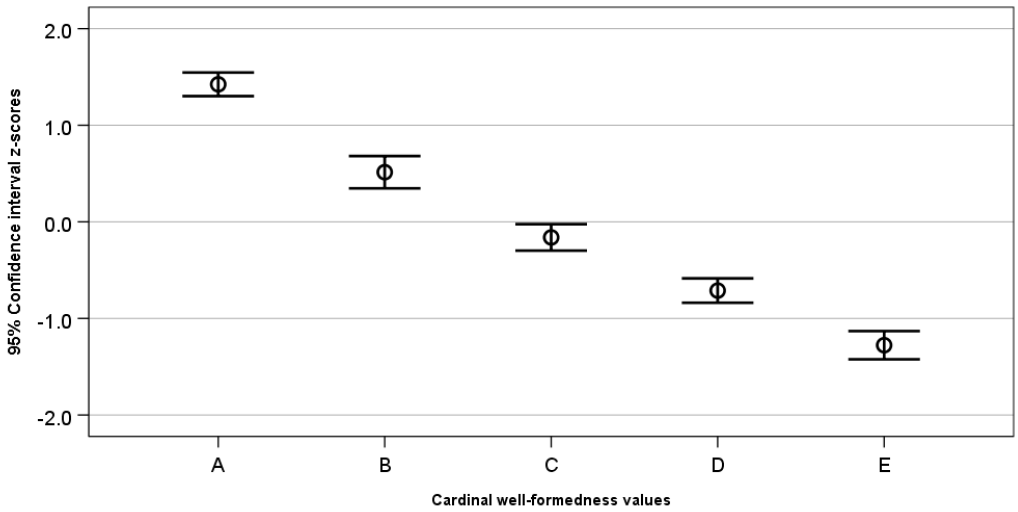



Figure 2: Mean results of the standard items instantiating the five cardinal well-formedness values A-E from Gerbrich et al (in press), Featherston (in press). Each error bar represents 96 data points.

But now look at Figure 3, which shows the same conditions as 2, but split up so that each participant's ratings are shown separately. Note that the judgements of the individual informants are shown as bars, with each bar representing the mean of three judgements. The participants' results are arranged by the experiment version that they saw, but the standard items were identical across the versions, so it is no surprise that these do not differ.

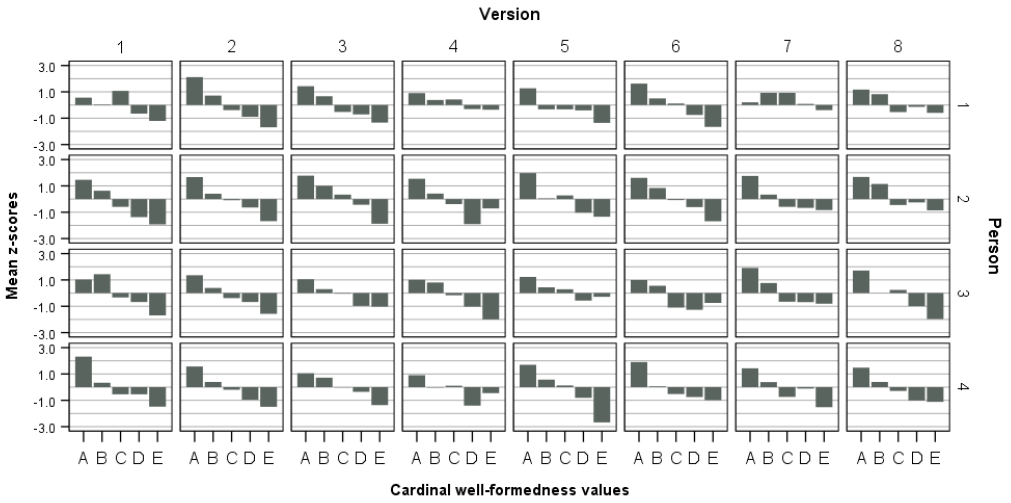



Figure 3: Mean judgements of each participant of the three standard items instantiating each of the five cardinal well-formedness values A-E. The labels A-E on the baseline apply to all charts vertically above them.

Most people most of the time get the five values in the right order, but sometimes they don't. In fact 19 out of 32 get them all in the correct order (counting ties as success), so 13 don't. Person one of version one in the top left-hand corner is a case in point: the bars A, B, D, and E (see labels in baseline) are in the right order, but the three judgements making up value C have a higher mean value than those of A and B.

The noise in an individual's judgements is plain to see even in these means of three judgements, but single judgements are naturally even noisier. Figure 4 shows each single judgement as a bar from the participants who did versions 1-3; this chart is thus an expansion of the first three columns of chart 3 above.

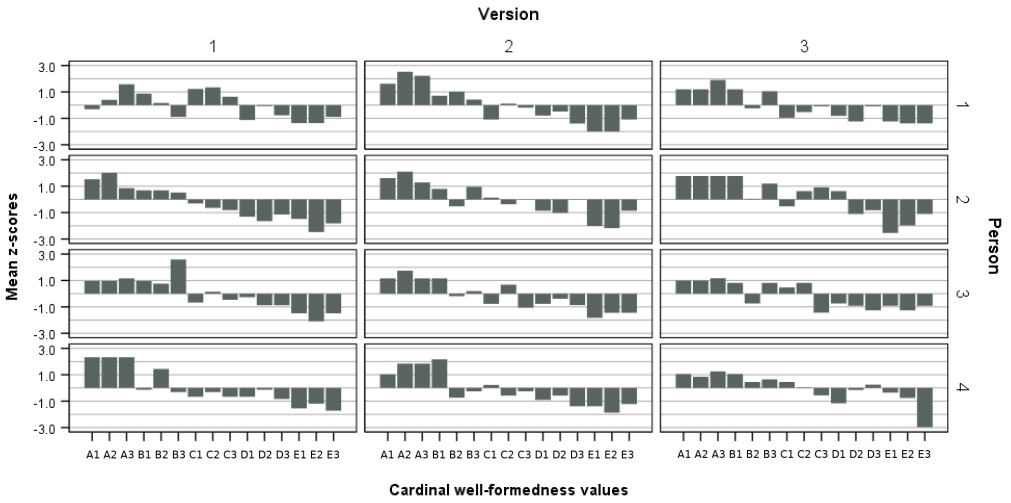



Figure 4: Single judgements of the standard items by the participants from versions 1-3. This chart distinguishes the three individual judgements of each well-formedness level.

The finding is unambiguous: while the group judgements produce clear and robust distinctions, the single judgements are noisy. This is consistent with the finding of very low statistical power reported for the sample size of 6 judgements in SSA13. While group results show a clear patterns, even the means of three are still quite noisy, and the individual judgements show the trend in the data but not much more.

To summarize, we thus have converging evidence that individual judgements offer only poor quality evidence, but judgements from groups are much more reliable. Since one of the key factors differentiating armchair judgements from experimental judgements is the use of informant groups, this would seem to suggest that experimental judgements are preferable as a source of evidence. This seems obvious to me, but I meet people who are persuaded that the papers SA12 and SSA13 showed that most armchair judgements are correct. In the next section we will examine how these contrasting views can come about.

## 3 Reassessing SA12 and SSA13

We mentioned in the introduction that we would contrast our findings above with those of SA12 and SSA13, which are thought to present a much more positive picture of the success of armchair judgements. In order to see why people might interpret the results this way, we will first need to look at their studies in more detail. Let me repeat that I do not in any way wish to criticize or devalue this work. I only wish to clarify what it shows so that it is not misinterpreted.

SA12 wished to establish whether the quality of armchair judgements had led to erroneous assumptions in the field of linguistics. This is a useful approach: rather than addressing the *ability* of individuals to make accurate judgements, they investigate whether *in practice* this has caused erroneous judgements to be entertained in the field. In order to do this they took 365 judgements from the introductory textbook *Core Syntax*, Adger (2003). They examined the judgements in the book and broke them down into different types, and focusing on those which could most easily be tested. They looked at judgements of paired examples, where the claim is that one example is good and the other bad, and existence judgements, where the claim is either that a particular sentence type is grammatical, thus a part of the language, or ungrammatical, and thus not part of the language. They tested 219 of the first type using magnitude estimation (Stevens 1975) and 250 of the second type by giving a yes/no binary choice. A total of 240 informants carried out the magnitude estimation experiment, 200 did the forced choice experiment.

The core finding for the magnitude estimation experiment was that 3 out of 115 contrasts were not replicated (max. 2.6% non-replication rate), for the yes/no task it was that 5 out of 250 did not replicate (max. 2% non-replication rate). The authors therefore claim that they see 'no reason to favor formal methods over traditional methods solely out of a concern about false positives'. (SA12; 634). It is this conclusion which has led some people in the field to believe that armchair linguistics have been validated. There are however a number of reasons why one might consider this belief to be hasty. I will address them in turn here.

### 3.1 What counts as success?

The first reason why we might relativize this interpretation of the results is the *relative* criterion that the authors choose for test for replication. For the magnitude estimation experiment, they state:

> '[...] we will define replication as the simple detection of a significant difference in the correct direction between the conditions in a phenomenon.'(SA12; 615)

And for the existence example with a yes/no test:

> '[...] we will define replication as the observation of significantly more yes-responses than no-responses for the sentences that were reported as grammatical by Adger, and as the reverse (more no-responses than yes-responses) for the sentences that were reported as ungrammatical by Adger.'(SA12; 615)

They discuss this choice at some length and explicitly admit that other criteria might have been applied. But the reasoning is clear:

> 'First and foremost, we believe that the simple detection of a difference in the predicted direction is closest to the intent of Adger (2003).'(SA12; 616)

I would suggest that an alternative interpretation is possible. Adger (2003) is an introduction to the central aspects of Generative Grammar. The model of syntax and well-formedness that I would expect such a textbook to apply is one where the successful grammar generates all and only the sentences which are part of the language, and no others. In this model therefore, well-formedness ('grammaticality') is *absolute*, by which we mean that every sentence has a value on a two-pointed scale, with no possible further gradations, but well-formedness is also *inherent*, which means that this value is independent of any structural relationship or comparison with other sentence. Looking back at Adger (2003) we can see support for exactly this position, since the text uses almost exclusively the contrast of starred and unstarred examples and not intermediate symbols (as SA12 themselves note). In fact the text is admirably explicit on this point. It distinguishes three reasons for unacceptability (Adger 2003; 1.1.2); an example can be unacceptable because it is:

- hard to parse: *I looked the number that you picked out with a needle [...] up.*
- implausible: *The amoeba coughed.*
- or else it is simply not licensed by the grammar: *By is eaten monkey banana that the being*

He distinguishes between unacceptable for performance reasons and ungrammatical in the technical sense:

> 'Remember that we assign a star to a sentence if we think that the explanation for its unacceptability is that it does not conform to the requirements of the grammar of the language under discussion.'(Adger 2003; 5)

This leads me to conclude that he is dealing with the classic model of grammaticality. Note also that (un)grammaticality is structurally driven:

> '[. . . ] we assume that speakers can't assign a structure to the particular string in question at all.'(Adger 2003; 5)

Ungrammaticality is absolute and categorical:

> 'We cannot make the [ungrammatical] sentence any better [. . . ].'(Adger 2003; 3)

In the light of these very clear statements in Adger (2003), it seems to me that the most obvious interpretation of Adger's assumptions is that grammatical sentences should be judged to be acceptable, and ungrammatical sentences should be judged to be unacceptable, and it is this criterion that we shall pursue here. This contrasts with the interpretation in SA12, which is that Adger only assumes a relative difference between grammatical and ungrammatical. We can illustrate the contrasting predictions in Figure 5.

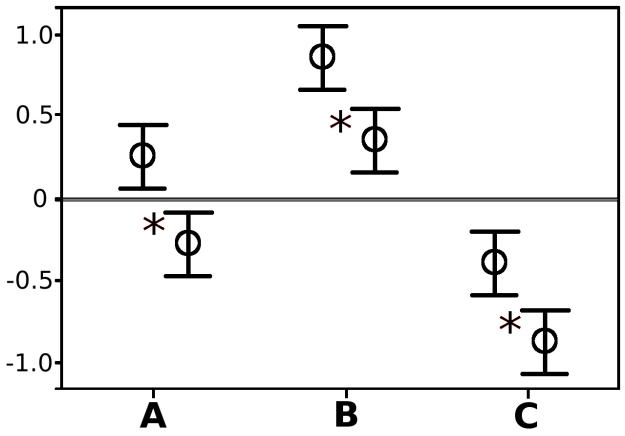



Figure 5: This chart illustrates the differences in interpretations of Adger's (2003) model of well-formedness. The pairs of error bars represent paired examples in SA12's data set. The unstarred bars represent examples Adger (2003) judges grammatical, the starred bars represent their ungrammatical comparison examples. The question is whether pairs B and C would count as replications.

In Figure 5 we see three example pairs and their hypothetical experimental results shown as error bars, the y-axis represents normalized judgements, with higher values representing greater perceived acceptability. On SA12's criterion for replication, all three pairs count as successful replications, because only a relative difference is required between the pairs; there is no requirement for an absolute level of acceptability. If we adopt my own interpretation of Adger's model of grammaticality as the criterion, pair A is a replication, but pairs B and C are not. This model of grammaticality would demand that the grammatical example be in the acceptable range and the ungrammatical one in the unacceptable range for successful replication.

Since SA12 gather continuous data in their magnitude estimation studies, there is no simple division between the acceptable and unacceptable ranges. But SA12's magnitude estimation study tested exactly equal numbers of examples claimed to be grammatical and ungrammatical (109 of each, plus one with a question mark which we ignore). We can therefore, exceptionally, treat the mean value of the

normalized judgements as a meaningful threshold. This is of course only an approximation, but as long as we keep the approximation in mind it seems legitimate. This adjustment to the success criterion results in 15 of the 115 contrasts in the magnitude estimation experiment failing to be replicated, which is a 13% failure rate, instead of the 'maximum' 2.6% failure rate that SA12 report.

To their credit, the authors of SSA13 report a similar calculation to this on the equivalent data sets from their *Linguistic Inquiry* sample. They look at the normalized data from the magnitude estimation task and the seven point scale task, which produce results on a continuous scale and examine how many sentences fall into the 'wrong' half, that is, how many examples judged well-formed are given ratings like those of the ill-formed group and how many examples judged ill-formed fall among the well-formed group. They do not fix a threshold on the basis of an external criterion but look for the threshold which minimizes the number of examples on the wrong side.

They find that the fewest possible items on the wrong side is 28 in the magnitude estimation data, and 27 in the seven-point scale data. These figures represent 9.8% and 9.5% of the total data included in the analysis. Applying this minimization approach to the magnitude estimation data in SA12, we find the smallest number of errant data points with the threshold set at -0.105, which yields a minimum number of 13 failures, which is an 8.8% failure rate. Now in fact I don't think that the exact numbers are very important. But I think it is important that it is clear that even subtly different assumptions may radically affect the result.

To summarize, if we assume a classic generative model of well-formedness with a grammatical/ungrammatical dichotomy, these tests of armchair judgements are producing minimum failure rates of 8.8% (SA12) and 9.5% and 9.8% (SSA13). This is one of the reasons that I think the results in these papers are misinterpreted when they are taken as giving armchair linguistics a blanket clean bill of health.

### 3.2 Do binary oppositions make a grammar?

In this section I would like to highlight another reason why I think the implications of SA12 and SSA13 need to be given a nuanced interpretation. The authors of SA12 and SSA13 are addressing the very specific question about whether judgements in published material are erroneous, and for this they need an easily applicable test for successful replication. But the issue in this wider debate is whether different types of judgements can form an appropriate basis for a grammar. Whether a data type is reliable is only one criterion, the amount of relevant information contained in the data type is another. Whether a judgement type is adequate thus also depends on what model of grammar and well-formedness we assume or strive for.

Relative judgements yield evidence for the existence of a difference between minimal pairs and thus for the existence of the factor that distinguishes them, but

relative judgements offer us little more. They do not directly tell us anything about the size of the effect, nor about where the individual items might be located relative to a scale of well-formedness (as SSA13; 225 note themselves). They are therefore fairly restricted in their information content. We can see this in the results of SA12 in the multi-example contrasts. While most of the examples from Adger (2003) they tested were pairwise comparisons or existence examples ('this is/is not possible in the language'), there were also 11 example sets consisting of multiple examples. The authors note that the failure rate in these examples was 3 out of 11 cases, which is 27%, far higher than in the pairwise tests; they discuss the finding at some length.

I present here an example of these non-replications to show how this occurred. This set of examples was designed to illustrate the well-known superiority and discourse linking effects (Chomsky 1973, Pesetsky 1987). Table 1 below gives the sentence codes from SA12, which include the example judgements (g=grammatical, *=ungrammatical), the sentence itself, the magnitude estimation score, and an abbreviation for the syntactic condition (my addition).

| Code | Sentence | mean | structure type |
|---|---|---|---|
| 9.120.g | Who poisoned who? | 0.11 | wh-subj wh-obj |
| 9.120.* | Who did who poison? | -0.54 | wh-obj wh-subj |
| 9.124.g | Which poet wrote which poem? | 0.40 | wx-subj wx-obj |
| 9.125.g | Which poem did which poet write? | -0.10 | wx-obj wx-subj |

Table 1: Table showing results of magnitude estimation experiment on superiority and discourse linking from SA12. *wh* indicates a bare wh-item, *wx* indicates a discourse linked wh-phrase.

We illustrate these scores in Figure 6 below, because the graphic presentation allows us to take in the situation at a glance.

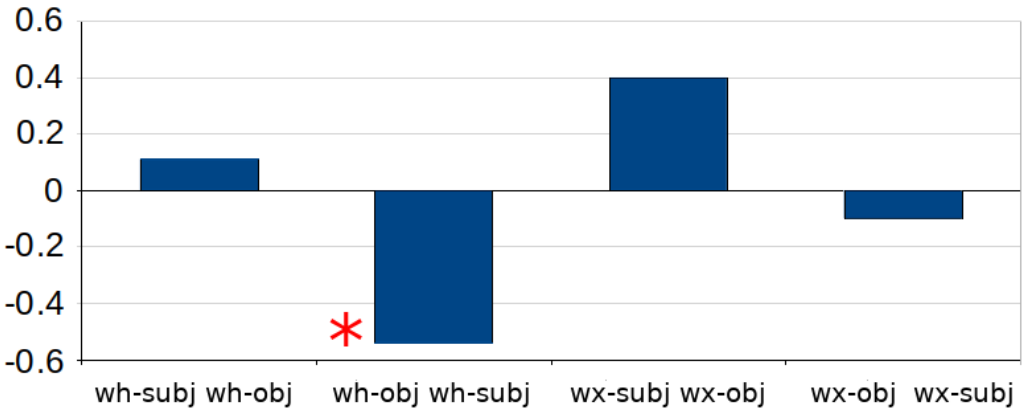



Figure 6: This chart shows the magnitude estimation scores of the four sentences making up the illustration of superiority and discourse linking from SA12.

Traditionally (e.g. Pesetsky 1987), this pattern has been thought of as consisting of three good examples and one bad example, and Adger (2003) follows this, so SA12 with their relative replication criterion consequently test it as three 'better' and one 'worse' examples. But the experimental scores show one good (0.40), two medium (0.11, -0,10), and one bad condition (-0.54), and these three levels are roughly equidistant.

The conception of well-formedness as a mere relative distinction that SA12 apply has difficulties dealing with such a pattern. When it finds a difference, it attributes the status 'well-formed' to the better one and 'ill-formed' to the worse one. That is all it can do. So it naturally encounters problems when applied to a data set like this one with more than two levels, as superiority and discourse linking standardly show (e.g. Featherston 2005). The results will probably show us that the two medium conditions are different to both the bad condition *and* the good condition). At that point this model crashes because these intermediate bars are assigned both the status 'well-formed' and 'ill-formed'. A purely relative well-formedness model as applied in SA12 and SSA13 simply does not make the distinctions required to deal with this sort of data pattern. Using a more sophisticated well-formedness model would capture this data set, but it would set higher demands on what counts as descriptive success, which would make it more difficult to achieve with only armchair judgements.

In fact we can derive a generalization from this example. We find this multi-level data set because there are two things going on: there is both the superiority constraint and the separate discourse linking effect, which the data reveals to be two independent factors. SA12's relative model can capture the existence of individual constraints, but it cannot reliably capture the relations between constraints. Since a grammar is surely more than just a collection of isolated non-interacting

constraints, then this is a real drawback. Constraint interaction is a key component of a grammar. We see this in Adger (2003), chapter 2 where the basic sentence *The pig grunts* is varied in various ways to show how underlying features determine grammaticality. Linguistic data sets do not consist only of binary pairs; they must contain multiple contrasts to show the full picture - Table 2.

| | | |
|---|---|---|
| The pig grunts | The sheep bleats | *The pig grunteds |
| The pigs grunt | *The sheeps bleat | The pigs grunted |
| *The pig grunt | The sheep bleat | The pig grunted |
| *The pigs grunts | *The sheeps bleats | *The pigs grunteds |

Table 2: Table showing variations of the form of the basic sentence *The pig grunts*, illustrating how multiple factors interact to produce complex well-formedness patterns (based on Adger (2003) chapter 2).

No set of four, much less any pair, is enough to gain a full picture. Only comparisons in multiple parameters can capture how factors interact to produce the complex grammatical patterns of even such apparently simple phenomena as subject-verb agreement. Chomsky (1965) underlines this too:

> '[. . . ] it is clear that we can characterize unacceptable sentences only in terms of some 'global' property of derivations and the structures they define – a property that is attributable, not to a particular rule, but rather to the way in which rules interrelate in a derivation.' (Aspects; 12)

The question that this article asks is whether we can use judgements to build a grammar. The purely relative judgements instantiated in the replication criterion in SA12 and SSA13 yield pairwise distinctions, but they do not contain enough information to serve as a basis for a grammar, since this consists not only of a list of individual effects, but also how these interact with each other. It follows that, even if it were the case that armchair judgements reliably made relative distinctions, this would still not demonstrate that these judgements can function as a sufficient basis on which to build a grammar. This test cannot therefore answer this question.

### 3.3 Do armchair judgements make too many distinctions?

One more general reason that I do not think that the studies in SA12 and SSA13 can be seen as legitimizing armchair linguistics relates to the sort of test that these papers apply. Whenever these texts report replication success, the authors are careful to note that this only concerns the *false positives*, and they discuss the implications of this in some detail. In spite of this, I repeatedly meet colleagues

who think that these papers demonstrate the validity of armchair judgements conclusively. It is thus apparent that not all linguists grasp what the restriction to false positives means. So let me restate it.

Armchair judgements and experimental judgements are more or less *the same data type*, only for experimental judgements we take more trouble with design and procedure: we ask more people, we use more different lexical forms of the structures, we may use a more precise scale. These factors give us more exact information. It is therefore quite implausible that armchair judgements should produce *more detail* than experimental judgements. But the test for false positives is asking whether the armchair judgements in the literature contain *too many* distinctions, that is, distinctions which cannot be replicated.[2]

Now this reproach of being too powerful, producing too much differentiation, is one which is fairly commonly made towards experimental syntacticians and experimental data, by practitioners of armchair linguistics. The idea is often that some differences revealed by experiments are not grammatically relevant, which is perhaps not without some justification. It is the other way round with armchair judgements, their major disadvantage – in my opinion – is that they produce insufficient detail. But the false positives test does not address this problem, it should not therefore be interpreted as validating armchair linguistics.

In a court of law, we must swear to tell 'the truth, the whole truth and nothing but the truth'. The false positives test addresses whether judgements are telling nothing but the truth. But it doesn't establish whether they are telling the whole truth. A validation of armchair judgements would require that too. To continue this metaphor: armchair judgements are like a short-sighted witness who had forgotten his glasses on the day in question. They don't even *see* the whole truth. Whether they are prone to exaggeration (=false positives) is not our major concern.

## 4 Towards a better grammar

We have seen in section 2 that there is robust evidence that armchair judgements are able to make fewer distinctions than experimental judgements and that they are noisy. In section 3, we looked at the papers SA12 and SSA13 which address some aspects of the accuracy of judgements in the literature. They are interesting and valuable work, but the implications should not be overstated. They

[2]It is interesting to consider what the source of these additional distinctions might be, since they cannot derive from the data type itself. One possibility is that they are instances of linguists generalizing their own noisy judgements. We saw above in 4 that people giving judgements sometimes make idiosyncratic distinctions that are not replicated in the group results. Another possibility is that distinctions could derive from a particular lexical instantiation of a minimal pair. Armchair judgements usually contain lexical variants, but experimental judgements control for this by systematically testing structural contrasts with multiple lexical items.

certainly do not provide a blanket validation of armchair judgements – nor do their authors claim this.

In this final section I should like to sketch out how I think that our field should move on from armchair judgements. Let us note here that the authors of SA12 and SSA13 are quite aware of this alternative approach and address it in these papers, but their focus is on testing the data basis of work done in the past rather than designing the tools for the future.

I think syntactic theory needs to look forward, but it also needs to be true to its roots. In order to make progress the field needs to continue its move towards becoming evidence-based. Claims need to be based upon a more solid data base than just a single person's judgements, since these are so noisy and subjective as to be unfalsifiable. This step will enable us to develop a clearer research cycle of hypothesis testing and rejection. This in turn will help clear the field of its legacy of theoretical fossils and allow their replacement. Armchair judgements are insufficiently finely grained and determinate to support this process.

But this does not demand that we abandon judgements as our basic data type. One factor in the success of the generative enterprise has been that it seeks to account for our intuitions of well-formedness, which we can access at any time. This both makes it into a field of study with psychological relevance and grounds it in our personal experience. We can build on this and stay true to our roots by building grammars as models of well-formedness judgements, in line with Chomsky's suggestion:

> '[...] there is no way to avoid the traditional assumption that the speaker-hearer's linguistic intuition is the ultimate standard that determines the accuracy of any proposed grammar [...]' (Chomsky 1965; 21)

But if we are going to take judgements as our ultimate standard, we need to faithfully reflect what we see in the judgement data, not abstract from it. We cannot pick and choose what aspects of judgement data we will adopt in our analyses, unless we can show that a particular factor is unrelated to the linguistic stimulus, and if we argue this, then we should systematically control for this factor. But our aim should be to account for the whole of the remainder.

### 4.1 Crime and punishment

So what does that mean in practice? In terms of the metaphor of telling 'the whole truth', what is the whole truth? Let us look again at the magnitude estimation results for superiority and discourse linking from SA12 in Figure 6 (which we repeat here as Figure 7 for convenience) to remind ourselves what the primary data looks like.

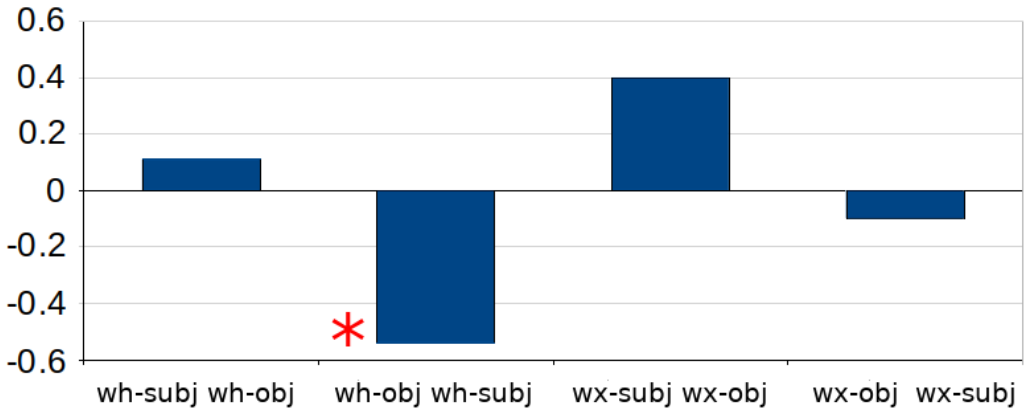



Figure 7: This chart shows the magnitude estimation scores of the four sentences making up the illustration of superiority and discourse linking from SA12.

When we look at such results, we observe again and again that the basic driver of the patterns we see are quantifiable effects, not just rules which are broken or not broken.

The traditional (e.g. Pesetsky 1987) conception of this phenomenon is that there is a rule violation for the inversion of subject and object wh-items, which is however nullified if the wh-items are wh-phrases of the form *which poet* and link into the discourse. But that is not what the data shows. Instead we observe that there is a cost to inverting the order of subject and object wh-items, which is largely independent of the wh-item type: this superiority effect just makes the structure seem less acceptable by a roughly constant amount. But there is a second effect in this data too, which is discourse linking. The observation is that informants judge these multiple wh-questions better if the wh-items involved are full wh-phrases of the form *which poet* or *which poem*, rather than simple wh-pronouns such as *who* or *what*. This effect too causes a difference in perceived well-formedness.

But here is the big point: the two effects are independent of each other, but they add up. The best condition is that with no superiority violation and discourse-linked wh-items *wx-subj wx-obj*, which takes the form *Which poet wrote which poem?*. The worst condition is *wh-obj wh-subj*, in which both factors are negative. The two conditions in the middle each have one bad and one good factor. We cannot describe the status of these examples without reference to both these factors, but we cannot capture their interaction without quantifying their effects. This is all the more necessary because these two factors do not have the same 'strength': the superiority effect is larger than the wh-item type effect. Our grammar has to specify both what the restrictions are but also what happens if they are violated.

We find this sort of pattern not only in this data set, but systematically. Here is another set of four items that SA12 tested from Adger (2003).

| Code | Sentence | Mean | Structure type |
|---|---|---|---|
| 10.90.g | That Peter loved Amber was obvious. | 0.06 | ClsSubj |
| 10.91.g | It was obvious that Peter loved Amber. | 1.02 | itSubj |
| 10.93.* | Who was that Peter loved obvious? | -1.04 | ClsSubj ex |
| 10.92.g | Who was it obvious that Peter loved? | -0.39 | itSubj ex |

Table 3: Table showing results of magnitude estimation experiment on sentential subjects and extraction from SA12. *ClsSubj* indicates clausal subject, *itSubj* indicates an expletive subject, *ex* indicates extraction.

We illustrate these results in a bar chart too, as this makes the pattern more immediately discernible - Figure 8.

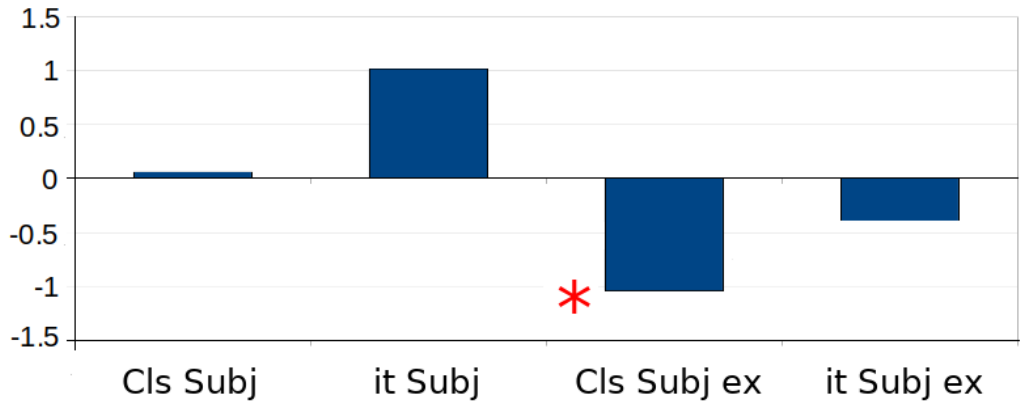



Figure 8: This chart shows the magnitude estimation scores from SA12 of the four sentences from Adger (2003) showing the interaction of sentential subjects and extraction.

Adger (2003) uses these examples to present a sentential subject island constraint; so the set should consist of three good examples and one starred one. But here too we can see that there are two effects in operation which independently affect the perceived well-formedness of the example sentences and produce multiple levels of acceptability. Having a clausal subject entails a cost in acceptability; the extraction also comes at a cost. As in the superiority examples above, the effects add up. Again, we cannot model this without quantifying the effects, because you can't add categorical differences.[3]

[3]Let us note here that there are other examples where effects seem to have more complex interactions, super-additive for example (cf. Sprouse & Hornstein 2013). There are also ceiling and floor effects. But for our purposes it suffices if the contraints interact in some way.

We therefore see two strong reasons to use a model of well-formedness which quantifies effects. Firstly, the observed patterns of data require this information. The patterns of judgements show (amongst others) additive effects, and addition entails quantification. Since we have committed ourselves to native speaker intuitions as the ultimate criterion we should therefore include this information. Second, approaches which do not use this information are making things unnecessarily difficult for themselves. In both the discourse linking case and in the subject island case, the traditional assumption is that the grammar must capture a rule infringement that only applies in specific circumstances: superiority only with a certain wh-item type, extraction only from clausal subjects. The data shows these assumptions to be questionable: on the face of it, both apparent island constraints look like cumulative effects which apply across the board (cf. papers in Sprouse & Hornstein 2013, especially Phillips 2013). This matters, because it means that syntacticians have been looking for something that doesn't exist (cf. Hofmeister & Sag 2010).

There are thus both theoretical and practical reasons to gather quantifiable data, and this will contribute towards the production of more detailed and more empirically grounded grammar models. The use of experimental methods will assist us in this, since these enable us to obtain more detailed results. This does not necessarily mean that only fully-fledged experimental data can be useful, however, as we shall suggest below.

## 4.2 Gathering quantified judgements

Setting ourselves higher goals in terms of data quality and descriptive detail brings challenges. While experiments on syntactic phenomena are becoming commonplace, the numerical results are still of restricted comparability. We can look at the size of an effect in numerical form and compare it to the size of another effect within the same experiment, but we cannot compare it to any other value, because there is no common scale. This limits the value of the quantification: for example, Savage (1970) argues that 'measurement' requires a scale of units.

The way forward we will suggest here is to use the cardinal well-formedness values we briefly introduced in section 2.2 above. We have developed 15 sentences which are reliably judged at five roughly equidistant points along the scale of perceived syntactic well-formedness. These five points are labelled A-E, and are designed to provide the same sort of comparison points that the cardinal vowels do for vowel quality. The highest value, A, is very natural and familiar. The lowest value, E, represents a level of ill-formedness which is as unnatural as can be achieved while still maintaining analyzability. We provide one example of each here in the text, the full set is in the appendix.

*Cardinal well-formedness value A*

The patient fooled the dentist by pretending to be in pain.

*Cardinal well-formedness value B*
Before every lesson the teacher must prepare their materials.

*Cardinal well-formedness value C*
Most people like very much a cup of tea in the morning.

*Cardinal well-formedness value D*
Who did he whisper that had unfairly condemned the prisoner?

*Cardinal well-formedness value E*
Historians wondering what cause is disappear civilization.

Although this was no part of the origin design, Gerbrich et al. (in press) suggest that the five values can be associated with the fairly standard annotations of degrees of acceptability used in the literature. Values A and B are fully acceptable and thus unannotated, but value C can be thought of as anchoring the value question mark (*?*), level D two question marks (*??*), and level E the asterisk (*). Not every linguist might use these annotations for exactly these values, but the association of the cardinal values with these conventional degrees of well-formedness seems useful.

We include the fifteen items in every experimental study, where they offer the advantage that they provide an external set of comparison points. This is useful because experimental judgements are often gathered on a continuous scale. This provides much more detailed information, but it does not in itself provide any statement about how good in absolute terms the sentences are. Syntacticians often wish to know not only what relative differences there are between items, but also where on the scale of well-formedness they lie. The standard items provide this information by providing anchor points along the scale of perceived well-formedness. We therefore include the scores of the standard items in the results graph of experiments.

A further advantage is that they permit us to compare results directly across studies. This can be done informally by just looking at the results of two experiments relative to the five cardinal values. But they can also be used to create a quantified comparison across studies. When we analyze the results of a judgement experiment, it is useful to normalize the data in order to remove the variation in the use of the scale by the participants: some people give better scores, others worse; some utilize a wider spread of scores, other use only a narrow range. In order to compensate for this, researchers often transform experimental judgements into z-scores. This manipulation involves subtracting from each score from the participant's mean score and dividing the result by the participant's standard deviation of scores. The normalized scores of each participant then have the mean value

zero and the standard deviation one, which removes a degree of inter-participant variation.

Now instead of using the participant's mean and standard deviation of *all* scores as the basis for normalization, we can use just their mean and standard deviation of the standard items as the basis. Each individual's scores will thus be expressed relative to that person's ratings of the standard items. This provides a directly comparable quantification of judgements even across experiments, as long as the standard items were included in both and the procedure and context of the two experiments is reasonably similar. Some caution in such comparisons is of course required, because many factors can distort judgement studies (Poulton 1989).

We chiefly use this technique to compare experimental results, but it can have advantages for informal judgements too. The availability of local anchor values can for example permit more exact judgements. SSA13 discuss how the forced choice methodology produces finer differentiation than the other methods they tested because it involves the discrimination of difference between two stimuli, rather than the placing of one stimulus on a scale. It is known that local comparison points, that is, those with very similar well-formedness values, provide even better support for fine judgements (e.g. Laming 1997); this is also the reason for the intermediate cardinal vowels, e.g. half-open and half-closed. The standard items provide local comparison points along the full range of syntactic judgements so that there is always a close comparison point.

A linguist wishing to evaluate an example X will (effectively but perhaps unconsciously) perform a series of forced choice judgements of the example relative to the standard examples. If she determines that the well-formedness of X falls between the B value and the C value, she can additionally consider whether X is nearer to B or to C, and perhaps decide that it is closer to B and therefore assign it the rating B-. The standard items thus allow us to import the advantages in exactness of comparative judgements into an anchored scale. This is an example of the way that improved data quality does not necessarily have to come from full-scale experimental studies. If the field collectively adopted cardinal well-formedness values there would be a marked improvement in exactness.

The scale also makes judgements more communicable. Our linguist can thus test whether her judgements correspond with those of another person, and they can debate whether a B- or a C+ is more appropriate. They can then report their conclusions to a third party, and the judgements will still be meaningful because the anchor points are accessible to any speaker of the language. When they give an example in a paper, they should also give their rating using the five cardinal values A-E, additionally distinguished by plus and minus signs. So I would give the examples in Table 1 above the ratings here in (1).

(1) a. (B-) Who poisoned who?
b. (D+) Who did who poison?
c. (B+) Which poet wrote which poem?
d. (C+) Which poem did which poet write?

This gives both information about the absolute ratings of the examples and about the amplitude of the factors which differentiate them. The addition or other interaction of different constraints becomes visible, which, as we have argued above, is an essential component of a grammar. Even if these values do not have the exactness of experimental data, they still go some way towards allowing the field of syntax to make progress. On the one hand, they are explicit and falsifiable, because they are related to the standard items. If another linguist wishes to contest these judgements, they can. If it turns out that example (1-c) is generally judged to be worse than the B level items, then my rating here is in trouble. This is already one big improvement from the traditional situation where a paper could assert or assume that a particular example was grammatical, but that claim would be so vague as to be incontestable. I think this will have a knock-on effect: if I know that someone can easily demonstrate that my claim is wrong, I will take much more care when choosing the judgement that I give and will ask some other people beforehand.

## 5 Summing up

This paper had three parts. In the first instance, I wished to highlight some facts which lead me to treat an individual's judgements with caution. The contrast between the group results and the individual results that we saw in section 2 show the apparent random variation present in single judgements and even in the means of three judgements. This finding was supported by the work in SSA13 on the power of different judgement methods, which show little power for small informant sample sizes. But the answer does not always have to be experiments: there are such things as careful judgements. I too use judgements and have some confidence in them, and I suspect that I could manage more consistently to approach the group data of the cardinal well-formedness levels in Figure 2 than some of individuals in Figure 4 do. But even I, who have been working on and with judgements for 25 years, do not believe that I can achieve anything like the quality of results available from groups judging carefully constructing sets of materials. It helps too, that I know what the data pattern of judgements look like: I know that I am looking for values on a continuum. Some linguists in the past have tied themselves in knots trying to fit multiple levels into a two-level categorical schema. The failure to recognize the interactions in the superiority and subject island data sets in Figure 6 and Figure 8 are a testament to this.

The second part addressed the papers SA12 and SSA13, which have been

pointed out to me as providing definitive validation for armchair judgements. These are high-quality papers, but I nevertheless do not see them as providing a blanket validation for armchair judgements; nor do the texts themselves make this claim. I make three arguments to support my position. First, if we change the criterion for replication to reflect the traditional model of categorical well-formedness, which is at least arguably what Adger (2003) assumes, then the replication failure rates rise considerably, rising to a minimum value of a little under 10%. This strikes me as being more a cause for concern than a cause for satisfaction, but this is naturally a matter of individual interpretation.

Much more significantly, the errors tested for are only the false positives, the claims of distinctions in the literature where none were shown to exist in finer-grained data. But making *too many distinctions* is not the chief problem that armchair judgements have: their main weakness is that they capture *too few* of the distinctions which finer-grained judgements reveal to be present. This poor definition and indeterminacy of the data basis has been a real brake on progress in syntax for decades, a matter which is of vastly more significance than whether some unsupported claims are made in the literature.

My final point in this section also relates to the well-formedness criterion adopted in SA12 and SSA13, but addresses it from the perspective of the question that this paper raises, whether judgement data can form the basis for a grammar. I therefore discuss how much and which information we need in order to construct a grammar, contrasting this with the only weakly informative pairwise distinctions. I suggest that a grammar is more than just a rule list, rather it must also specify how rules interrelate to produce sometimes complex patterns of data. Taking two of the multi-item example sets in SA12 as examples, I argue that the development of the grammar requires the quantification of rule violation costs, because only this will permit us to capture they way that rules interact.

In the final section I look forward. Even if it were the case that armchair judgements had proved the ideal data basis so far, it does not follow that they are ideal for further development. The genie is out of the bottle: we have seen all the extra detail which experimental judgements contain, the radically different analyses that it supports, we cannot pretend it is not there. We need to sort out how to deal with the new information accessible. We must seek ways of capturing it, recording it, and using it to build grammars. As an example of this I return to the five cardinal well-formedness values and their instantiation in standard items. I seek to show that this simple device permits even individuals to give more exact, communicable judgements, which are furthermore inter-subjectively anchored. They thus permit judgements to be assigned something approaching absolute values, and allow the inclusion of important information about the strength of factors.

We should finish by answering the question in the title. We can build grammars on the basis of judgements. But the quality of the data base is a limiting factor on

the quality of the grammar. We have in the past used relatively coarse judgements to build grammars, the quality therefore necessarily suffered. We now have easy access to a range of ways to improve our data basis and consequently our models of grammar. I see no reason for nostalgia.

## Appendix

These standard items anchor the five cardinal well-formedness values. The German examples from Featherston (2009) are well-established, the English examples here (from Gerbrich et al. in press) are still undergoing beta testing.

*Naturalness value A*
The patient fooled the dentist by pretending to be in pain.
There's a statue in the middle of the square.
The winter is very harsh in the North.

*Naturalness value B*
Before every lesson the teacher must prepare their materials.
Jack doesn't boast about his being elected chairman.
John cleaned his motorbike with which cleaning cloth?

*Naturalness value C*
Anna loves, but Linda hates, eating popcorn at the cinema.
Most people like very much a cup of tea in the morning.
The striker fouled deliberately the goalkeeper.

*Naturalness value D*
Who did he whisper that had unfairly condemned the prisoner?
The old fisherman took her pipe out of mouth and began story.
Which professor did you claim that the student really admires him?

*Naturalness value E*
Historians wondering what cause is disappear civilization.
Old man he work garden grow many flowers and vegetable.
Student must read much book for they become clever.

# 6 References

Adger, David (2003) *Core Syntax: A Minimalist Approach*. Oxford: Oxford University Press.

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

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

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

Featherston, Sam (2007) Data in generative grammar: The stick and the carrot. *Theoretical Linguistics 33* (3), 269-318.

Featherston, Sam (2009) A scale for measuring well-formedness: Why linguistics needs boiling and freezing points. In: Sam Featherston & Susanne Winkler (eds.) *The Fruits of Empirical Linguistics. Vol.1 Process*, 47-74. Berlin: Mouton de Gruyter.

Gerbrich, Hannah, Vivian Schreier & Sam Featherston (to appear) Standard items for English judgement studies: Syntax and semantics. To appear in: Sam Featherston, Robin Hörnig, Sophie von Wietersheim, & Susanne Winkler (eds.), Experiments in Focus: Information Structure and Processing. Berlin: Mouton de Gruyter.

Hofmeister, Phillip & Ivan Sag (2010) Cognitive constraints and island effects. *Language 86* (2), 366-415.

Labov, William (1996) When intuitions fail. In: Lisa McNair, Kora Singer, Lise Dolbrin & Michelle Aucon (eds.) *Papers from the Parasession on Theory and Data in Linguistics. Chicago Linguistics Society 32*, 77-106.

Laming, Donald (1997) *The Measurement of Sensation*. London: Oxford University Press.

Pesetsky, David (1987) Wh-in-situ: Movement and unselective binding. In: Eric Reuland & Alice ter Meulen (eds.), *The Representation of (In)definiteness*, 98-129. Cambridge, Massachusetts: MIT Press

Phillips, Colin (2009) Should we impeach armchair linguists? In Shoichi Iwasaki, Hajime Hoji, Patricia Clancy, & Sung-Ock Sohn (eds.) *Japanese/Korean Linguistics 17*, 49-64. Stanford: CSLI Publications.

Phillips, Colin (2013) On the nature of island constraints. I: Language processing and reductionist accounts. In: Jon Sprouse & Norbert Hornstein (eds.),

*Experimental Syntax and Island Effects*, 64-108. Cambridge: Cambridge University Press.

Poulton, Eustace (1989) *Bias in Quantifying Judgements*. Hove: Erlbaum.

Savage, C. Wade (1970) *The Measurement of Sensation*. Berkeley: University of California Press

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

Sprouse, Jon & Norbert Hornstein (eds.) (2013) *Experimental Syntax and Island Effects.* Cambridge: Cambridge University Press, 2013.

Sprouse, Jon & Diogo Almeida (2012) Assessing the reliability of textbook data in syntax: Adger's "Core Syntax". *Journal of Linguistics, 48* (3), 609–652.

Sprouse, Jon, Carson T. Schütze, & Diogo Almeida (2013) A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua, 134*, 219–248.

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