1 Background: experimental grammar research

The use of experimental methods in grammar research has gone from strength to strength and has established itself as one of the key ways to investigate linguistic patterning among words, phrases, and clauses up to the sentence level. This is strongly to be welcomed: many linguists have a feeling of unease about the thin ice of weak validity that work in syntactic and semantic theory sometimes skates upon when it is done without reasonable attention to its evidential base. In particular, if linguists can radically disagree about the underlying architecture of the grammar that they are attempting to describe without it being clear who is wrong and who is right, then this is an unmistakable sign that the data basis used is insufficient (either in quantity or quality or both) to uniquely determine the system to be described. With Popper, we can doubt that unfalsifiable claims are any scientific claims at all.

In this paper we take the view that both hypothesis building and hypothesis testing can be improved by the use of more fine-grained data and the use of multiple lexical variants of structures. If linguists employ data sets with proper control of potential confounding factors then the range of analyses they will propose will be more constrained. But it is especially important that the data set permits sufficiently sharp descriptions and predictions to allow clear falsification of hypotheses.

Another issue which is perceived to be problematic is the non-independence of the data source. When linguists give their own judgements and base their theory
development and testing on just this data, then the danger of bias, albeit uncon-
cscious bias, is obvious. We thus need evidence from third parties. The old claim
that something is possible ‘in my idiolect’ will not do either. Most researchers in
grammar will agree that the aim in the field is not just to describe the language
system of one individual. Instead we wish to make far wider generalizations about
what speakers of a language do in general. Collecting data from groups of inform-
mants enables us to make these stronger claims. The use of experimental proce-
dures has thus brought us some way forward, both by hardening up the data basis
of syntax to make wider, falsifiable statements, and by providing a far more exact
picture of what well-formedness tells us about the grammar.

That picture turns out to be very interesting. The two most important new
insights into the nature of the grammar concern the effect that a grammatical con-
straint violation has and how constraint violations interact with each other; neither
of these are as often assumed in the traditional literature. The first observation is
that the violation of a particular constraint triggers a constant cost in terms of per-
ceived well-formedness. A corollary of this is that different violation types have
different amplitudes of cost, so there are stronger and weaker constraints. This
means that an empirically adequate model of grammar must have a parameter of
violation cost strength; put differently, grammatical constraints must have specific
weightings.

The second observation is that multiple constraint violations affect our per-
ception of well-formedness in a cumulative fashion, so that an example sentence
that breaks two rules is worse than one which breaks only one rule. Furthermore,
both the violations appear to affect the violating example sentence independently
and additively, so that their violation costs remain constant.

These two findings are very robust, and can be formulated as in (1) and (2),
which are parts of the Decathlon Model (Featherston 2005, 2011).

(1) Weighting and cumulativity axiom 1

Given a sentence $S$ and a sentence $SA$, which is a minimal pair to $S$ but con-
tains the grammatical constraint violation $A$, then the difference in perceived
acceptability between $S$ and $SA$ will be $a$, the specific violation cost of $A$, all
other things being equal.

(2) Weighting and cumulativity axiom 2

Given the following sentences:
- a sentence $S$, . . .
- a sentence $SA$, which is a minimal pair to $S$ but contains the grammatical
  constraint violation $A$ with the violation cost $a$, . . .
- a sentence $SB$, which is a minimal pair to $S$ but contains the grammatical
  constraint violation $B$, with the violation cost $b$ . . .
- a sentence $SAB$, which is a minimal pair to both $SA$ and $SB$ but contains both grammatical constraint violations $A$ and $B$, . . .

then the difference in perceived acceptability between $S$ and $SAB$ will be the sum of the violation costs $a + b$, all other things being equal.

The differential weightings of constraint violations and the fact of their cumulative effect have become very clear, so that these characteristics are a fundamental component of our understanding of any realistic grammar model (see especially Keller 2000). Interestingly, some syntacticians had in fact come to this view independently by merely considering their own judgements, but these views were not sufficiently influential to cause them to be more widely adopted (e.g. Uszkoreit 1986, Jacobs 1988). It is more surprising that Chomsky’s *Barriers* (1986), which also contains the seeds of cumulativity, did not have this effect either. Perhaps the adoption of the fairly radical design feature competition on the basis of economy into the Minimalist Program, which was presumably also an attempt to match the model to the attested patterns of judgements, blocked the adoption of cumulativity. At any rate, cumulativity was dropped in favour of competition as a means of dealing with gradient data. Perhaps the popularity of Optimality Theory at this time played a role. Competition won all. Wrongly, as it now appears.

But there is one more apparent feature of the grammar which we have not dealt with yet but which follows from the previous findings. Every linguist who has collected data relating to well-formedness will know that robust distinctions are obtained between conditions, but there is no binary opposition between absolutely good and absolutely bad. Instead, we usually find a continuum representing degrees of perceived well-formedness.

This is sometimes due to the nature of the task assigned: most experimental procedures gather data points along a scale in some form or other. This task itself is likely to produce results in the form of a continuum, but this is not only due to the task type; informants generally perceive well-formedness as a gradient phenomenon. This is for example apparent when we consider the results of the speeded judgements task used by the research team Markus Bader and Jana Häussler (e.g. Bader & Häussler 2010). This method also produces gradient data even though participants are only given a binary differentiation task. The gradience comes from the frequencies with which informants choose one or other option, so that an integrated scale combining a (binary) well-formedness judgement and scalar frequency of choice is derived. The effect is not merely task-related, therefore, but seems to be inherent in our perception.

The traditional model of grammatical and ungrammatical structures, sometimes with the addition of a marginal class in between with a question mark, is thus insufficient: we must distinguish not only good and bad, but also good from better, and bad from worse. This seems quite revolutionary today but looking back
at the history of syntactic theory it is less surprising. A re-reading of Chomsky’s *Aspects* (1965) makes it clear that he is aware that this is a simplifying assumption: ‘Obviously, acceptability will be a matter of degree, along several dimensions. [...] it is clear that we can characterize unacceptble sentences only in terms of some ‘global’ property of derivations and the structures they define — a property which is attributable, not to a particular rule, but rather to the way in which rules interact [...]’ and then: ‘Like acceptability, grammaticalness is no doubt a matter of degree (Chomsky 1955, 1957, 1961).’ (Chomsky 1965; 10-11). In the light of this it can appear more surprising that syntax has spent the last thirty or forty years assuming that well-formedness is a binary opposition, especially since firstly, the raw data of gradient well-formedness is available to every speaker of a language, and secondly, the additional information is rich and useful.

This greater degree of differentiation offers much more information and real additional insights into the way that our grammars work, but this comes at a cost in additional complexity: it prevents us from giving a simple categorization of a sentence as well-formed or ill-formed, since experimental judgement studies offer us relative data, rather than absolute data. This can sometimes be a problem: linguists want to be able to make a clear statement about the status of a particular example. Relative data alone does not permit this; we can only say that example A is significantly better or worse than example B. It is for this reason that we developed standard items for inclusion in judgement experiments. We have previously made available standard items for German (Featherston 2009), in this article we supply standard items for experiments on English.

## 2 The uses of standard items

Standard items are sets of filler sentences which should be included in any judgement experiment. They are carefully selected and tested to represent the whole of the accessible scale of syntactic well-formedness. When the results of an experiment are evaluated, the scores given to the standard items should be evaluated in the same way. They thus provide a basis for comparison so that the judgement score of a given sentence can be more nearly given an absolute interpretation. The values given to the standard items thus function as anchor points on the scale of perceived naturalness. We distinguish five such points and label them from A (good) to E (bad).

In order to function as fixed points relative to which comparisons can be made, standard items should ideally represent known values on a familiar scale. The reason for this is that we are better at judging where a new stimulus is located on a scale if we have close and familiar reference points. An example would be that of temperature. If you ask people how warm it is outside, they can usually guess it to within few degrees. People know that below 18°C you need to put something over
your t-shirt, below 12° you need a coat, and below 6° you could put on gloves. If on the other hand we present people with water between 40° and 80° and ask them to estimate its temperature, their guesses will be much less exact. People know that boiling water will scald them, they know that 40° is a nice warm bath, but they have few points of reference between these, so they will judge it much less accurately.

Unfortunately, there is as yet no recognized scale of linguistic well-formedness beyond the division of sentences into those with stars, which are thought of as ‘ungrammatical’, and those without stars, which are thought of as ‘grammatical’. Although this is a start, linguists are notorious for disagreeing about whether specific examples should receive a star or not. Part of the problem is that the categories ‘grammatical’ and ‘ungrammatical’ refer to ranges of values, not just single points, so absolute values which could act as anchors are lacking. The system of standard items provides a multi-point scale of perceived well-formedness, relative to which other examples can be more exactly and definitely located.

In order to be maximally useful, a system of anchor points should have as many scale points as possible, because more points provide more distinctions. One approach to the size of the scale would be to match it to the number of degrees of well-formedness that speakers are able to distinguish. Informally our observation is that individual informants can distinguish perhaps ten different degrees of well-formedness, but the results of groups yield far more distinctions, perhaps as many as twenty (see section 4.2 below)

We do not need standard items that distinguish quite so many points on the scale, however. Practical considerations dictate that the standard items should be as few as possible, since these items must be included in every experiment. Each anchor point must be exemplified by (ideally) three items in order to establish a reasonably reliable value; a scale with ten points would then require thirty sentences to be inserted into every experiment. We therefore decided on five as the optimal number of degrees of well-formedness in the standard items. In practice this gives more than five different degrees on the scale as positions between the points are also identifiable. The use of our standard items thus requires the inclusion of fifteen standard items in every experiment, which seems a reasonable trade-off between desirable detail and undesirable experiment bloating.

The use of standard items in experiments also permits comparisons across experiments. If we include the same fifteen items in every experiment, we can realistically state how good the scores from one experiment were compared to the scores in an entirely different experiment. This is useful and often provides intriguing results, because the correspondences between structure types are not always as expected.

This can be taken further by using the standard items as the basis for a direct calculation of equivalences across experiments. When we analyze the results of
a judgement experiment, it is often useful to normalize the data in order to re-
move 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 exper-
imental judgements into z-scores. This manipulation involves subtracting from 
each score the participant’s mean score and dividing the result by the participant’s 
standard deviation of scores. Thus the scores of each participant have the mean 
value zero and the standard deviation one, which removes a degree of irrelevant 
inter-participant variation.

This method can be built upon to provide a quantified comparison across ex-
periments using the standard items. Instead of using the participant’s mean of 
all scores and standard deviation of all scores, we can use just their mean of the 
standard items and standard deviation of the standard items. This provides a firm 
basis of comparison even across experiments, as long as the standard items were 
included in both and the procedure and context of the two experiments was rea-
sonably similar.

Since the five values on the well-formedness scale are used again and again, 
they thus provide something approaching absolute values of well-formedness. If 
every linguist measured their examples against these standard items, then they 
would exactly become an inter-subjective standard, which would allow compari-
son of well-formedness judgements between any two linguists. Linguist A would 
say: ‘I find this one quite good, a little better than a B, perhaps.’ To which linguist 
B might reply: ‘I think you are being a bit kind there. I’d say it’s more like a B 
minus’. And each would know exactly what degree of well-formedness the other 
meant.

Another way of thinking about the absolute values of the five points is as repre-
sentatives of the degrees of well-formedness that linguists have traditionally used. 
We can think of both A and B grades are grammatical, but C is something like 
the intermediate degree often given a question mark (?), D is worth two question 
marks (??) and E is the fully ungrammatical asterisk (*).

An additional advantage of the use of standard items is that they fill out the 
accessible range of syntactic well-formedness. We should perhaps briefly clarify 
what we mean by this term. Experimental sentences fall within the accessible 
range of well-formedness if the informant knows what they are intended to mean 
and can analyze them within their grammar. Informants are able to judge word 
strings which do not have these qualities, but such examples do not produce evi-
dence relevant to syntactic computation. A couple of examples should make this 
clearer.

(3) a. The seven cows in the field eat grass happily all afternoon.
b. The seven cows in the field eat happily grass all afternoon.
3  THE SELECTION OF STANDARD ITEMS

Example (3-a) can be regarded as fully acceptable and provides an anchor point. Example (3-b) will be judged worse than (3-a) by exactly the violation cost of the requirement for objects to directly follow their subcategorizing verbs. Again the comparison of (3-a) and example (3-c) will show the violation cost of a verb agreement violation. The violation costs are quantifiable because all of these are recognizable faulty structures. But (3-d) consists just of word salad. It is not possible for us to say what is wrong with it or indeed what it is intended to mean and thus how it should be structured. It therefore falls outside the range of well-formedness judgements because it is not a faulty structure whose fault can be uniquely identified. It is no single structure, rather it exhibits a lack of structure. Such examples will show us how bad we perceive word mixtures to be but they can tell us nothing about the effects of specific constraints within the grammar.¹

3  The selection of standard items

We went about choosing the items for English in the same way that we selected the standard items for German, with multiple judgement experiments using the Thermometer Judgements procedure (Featherston 2009). Since we shall report several studies using this technique here, we shall briefly outline this collection method. This is a procedure for obtaining judgements from naive informants with the greatest possible degree of differentiation and the least possible distortion. This method is a development from Magnitude Estimation (Bard et al 1996). It varies from the simple elicitation of standard categorical judgements (‘Is this grammatical or not?’) in several ways. First, informants are asked to provide purely relative judgements: at no point is an absolute criterion of grammaticality applied. Judgements are relative to two reference examples and the informant’s own previous judgements. Second, all judgements are proportional; i.e. subjects are asked to state how much better or worse sentence A is than the reference examples. Next, the scale along which judgements are made is open-ended: subjects can always add an additional higher or lower score. Additionally, the scale has no minimum division: participants can always place a score between two previous

¹It is a moot question whether an example can be too good to be properly judged within the range of accessible structures. There does seem to be an intuition that there is such a thing as a perfect structure, but there still seem to be distinctions within perfectness. One relevant factor is length. Shorter examples, like (i), are judged better than longer ones, like (ii), because they require less processing effort, even though both are perfect.

i. Cows eat grass.
ii. The seven cows in the field happily eat grass all afternoon.
ratings. Last of all, two reference examples, one of which is quite good and the other quite bad, are given the values 20 and 30. These are sufficiently far from zero to avoid the known problem of scale end distortion there (Poulton 1989).

The instructions have the form ‘Look at example A. It is worth 20. Look at example B. It is worth 30. Relative to these reference items, how much would you give this one?’ The result is that subjects are able to produce judgements which distinguish all the differences in well-formedness they perceive with no interference from an imposed scale. This approach can produce judgement data of higher definition than traditional techniques such as a seven-point scale and thus affords us a clearer picture of the factors which affect perceived well-formedness.

In order to select standard items for English we carried out a total of five experiments with the aim of finding sets of appropriate standard sentences. The procedure has two steps: first we collect experimental judgements from English native speakers of example sentences which seem to us to span the full range of naturalness. These items should be evenly distributed across the full range of well-formedness, so that all areas are well represented. When we have a nice set of examples which cover the full range of values, we can divide the range of judgements exhibited into five equal areas. The values located in the middle of these areas become our five degrees of well-formedness.

We carried out several studies in which we elicited judgements using Thermometer Judgements (see above). We repeated this step several times in order to be sure that the distribution of items across the scale of naturalness was even and that it extended as far as possible up and down. We illustrate the approximate range of naturalness in (4) and the three sentences judged best and worst in (5).

(4)  a. The sales assistant went to ask if she could permit a discount.  
    b. You can’t say that to me, who does most of the work round here!  
    c. Who do you doubt that will complete his degree in three years?  
    d. Alice is looking for something which for her to give her brother.  
    e. The Spartans brave stood in line fight for freedom.

(5)  a. There’s a statue in the middle of the square.  
    b. The patient fooled the dentist by pretending to be in pain.  
    c. It’s nice to eat pizza and watch a film.  
    d. Historians wondering what cause is disappear civilization.  
    e. Student must read much book for they become clever.  
    f. Old man he work garden grow many flowers and vegetable.

Between these studies we discarded some items which had a lot of judgement variance as this would make them unsuitable as standard items. A lot of variance suggests that an item has more than one readily accessible reading, so that people are processing it differently. We also excluded some others when there were more
Figure 1: The results of our judgement collection (Exps 1-3) show the distribution of example sentences and the areas which exemplify the five degrees of naturalness.

items than necessary at the same level, and added items in places where there were too few. The final distribution can be seen in Figure 1. For the purposes of establishing a set of standard examples only the vertical axis is the measure that we are interested in, so the bumps in the line do not matter. The continuum in Figure 1 thus represents the range of well-formedness which is accessible to speakers (or perhaps to this methodology).

The second step in this process is to find good exemplars of the five values. These standard items should be fairly consistently assigned the same or similar scores, so that we can be sure speakers agree about them, and they should be approximately half-way between the items selected for the adjacent values above and below. We next tested which of these are most reliably assigned to their group.

To do this we carried out a new type of experiment. We first choose six possible candidate examples for each naturalness value, on the basis of their scores so far. Participants in this experiment have the naturalness groups presented to them and are instructed that their task is to assign examples to the correct group. They have the opportunity to practise this in a practice stage. From the six sentences from each of the five groups we select one example sentence; these become the comparison items for the groups. Participants then see the other twenty-five examples one after another and are instructed to assign each of them to the group whose naturalness level the example best matches.

This experiment took place in three forms. These varied only in which example sentence from the group was used as comparison item for that group. An example may make this clearer. If we label the six candidate examples for a given naturalness level $a$ to $f$, in the first version of the experiment, the informants re-
ceived example $a$ as the reference item and were presented with examples $b$ to $f$ to assign to the groups. In the second version $b$ was the reference example, and $a, c, d, e, f$ were to be assigned, and so on. The results give us a good idea how reliably the examples are matched to their groups.

The results of this experiment are presented in Figure 2. In order to make the accuracy of the assignments to groups visually accessible, we must treat the standard values A to E as numerical values 5 to 1. We then encode the group assignments as numerical values and calculate a mean value: an example from well-formedness group 1 (= A) should have the mean value 1.0, an example from group 3 (= C) should have the mean value 3.0. The squares mark the expected - or rather intended - value if an item is a perfect representative for a naturalness group. The error bars thus give a good idea of how an example was assigned; if the mean value is further from the square, the item was assigned more often wrongly. If the error bar is long, there was less consistency in assignment. In an ideal world, the means of the error bars would all lie exactly on the rows of squares; in fact there is a degree of error. In particular, the best examples tend to be rated worse than the 5.0 value, while the worst are rated better than the 1.0. This is perhaps inevitable, given that errors on these items can only occur in the one direction.

It will be noted that almost no items are always correctly assigned. The fact that there is a degree of random noise in human language performance is reflected in the Decathlon Model (Featherston 2005). This has sometimes been interpreted as unreliability in the data type (for extensive discussion see Schütze 1996), but while individual single judgements are noisy, groups of informants can consistently give judgements with very fine distinctions (Featherston 2007). The noise in individual judgements partly explains why linguists provided contrasting judgements in the days of the armchair linguist.

We chose those items which are most consistently assigned correctly to become members of our final set of standard items. The full set of examples tested is in the appendix, but the final set of standard items is listed here in (6). The sentence numbers are those given in Figure 2.

(6) Standard items for naturalness value A

6: The patient fooled the dentist by pretending to be in pain.
50: There’s a statue in the middle of the square.
53: The winter is very harsh in the North.

Standard items for naturalness value B

13: Before every lesson the teacher must prepare their materials.
15: Jack doesn’t boast about his being elected chairman.
21: Jack cleaned his motorbike with which cleaning cloth?
Figure 2: The results of our group assignment experiment. The squares mark the expected results, the error bars show the mean values and 95% confidence intervals of the assignments.

**Standard items for naturalness value C**
24: Hannah hates but Linda loves eating popcorn in the cinema.
16: Most people like very much a cup of tea in the morning.
32: The striker must have fouled deliberately the goalkeeper.

**Standard items for naturalness value D**
44: Who did he whisper that had unfairly condemned the prisoner?
63: The old fisherman took her pipe out of mouth and began story.
30: Which professor did you claim that the student really admires him?

**Standard items for naturalness value E**
65: Historians wondering what cause is disappear civilization.
60: Old man he work garden grow many flowers and vegetable.
61: Student must read much book for they become clever.

We would recommend the use of these standard items in all studies which involve measuring well-formedness, especially when it is beneficial to be able to make reference to the approximation to an absolute standard of acceptability which they provide. The use of standard items in studies of syntax has amply demonstrated its usefulness in studies on German (Featherston & Winkler 2009), so we are pleased
to offer an English set for general use.

It is best to include all fifteen items in any and every study, randomized among all the other example sentences being investigated. As such they additionally offer a basic set of fillers which distract participants from the point of the investigation. These items will generally be found to produce a continuum of five roughly equally-spaced points, though there are a number of factors which can distort this. For example, if the well-formedness values of the experimental sentences fall disproportionately often between two specific values, then the scores of these values will tend to be pushed further apart. We can imagine that the participants are trying to ‘make space’ for the many judgements. For more details of known distortion effects, see Poulton (1989). Note however that the simple use of standard items in an experiment will help reduce the general level of distortion by providing values across the full range of accessible well-formedness values.

4 Example study: Binding into adverbial clauses

In order to show how the standard items work in practice we shall present a recent experiment using them. It comes from our series of studies on the syntactic integration of adverbial clauses. As the experiment is fairly complex, we will present it in two parts, for clarity of exposition. We reflect on the relevance of the standard items after each part.

4.1 Example experiment part I

There has recently been considerable interest in the question how dependent clauses are attached to matrix clauses. Can the full range of these apparently diverse relationships be captured with the mechanisms and categories of standard models of syntax? To investigate this question we have been looking at examples of adverbial subordinate clauses where the identical surface form can have two different grammatical and interpretive statuses. For example, the adverbial while clause in (7-a) is a temporal specification and it modifies the matrix clause event. The superficially identical while clause in (7-b) on the other hand expresses a separate proposition which contrasts with the first.

\[
\text{(7) a. While Jean runs in the park in the morning she mentally prepares her lectures.} \\
\text{b. While Jean runs in the the park in the morning, she mentally relaxes in the evening.}
\]

These two types of subordinate clause are dubbed Central Adverbial Clauses (CACs) and Peripheral Adverbial Clauses (PACs) by Haegeman (2003). They are frequently referred to as integrated and non-integrated clause types (e.g. Reis 1997). The assumption is that their integration status derives from the fact that
they are attached to the syntactic tree of the matrix clause at different points, though there is no consensus about what that point might be (Reis 1997, Frey 2011). Our experiment uses variable binding, a fairly standard test of constituency, to investigate whether or not the two clause types behave as if they were a part of the core clause, perhaps the VP structure. If they are, then they would be c-commanded from the subject position in the matrix clause and thus variable binding could occur from that position into the clause. This therefore is the criterion we use.

While many authors have used quantifier variable binding as evidence, the sorts of structures which have been used and the choice of quantifier has been very varied. This study was part of our research programme aiming to find out if all quantifiers and context structures produced the same results. In order to do this, the experiment contrasted four quantifiers: nobody, no NP, hardly anyone, and an NP. Three of these quantifiers were negative, because it is often felt that negative quantifiers produce the most valid data (cf. Pauly 2013). We also varied the position of the quantifier and main clause: it could either precede or follow the adverbial clause. Since the aim was to investigate whether this structural test could yield worthwhile data about the difference between CACs and PACs, we contrast these two adverbial clause types in our experiment.

We may thus distinguish three parameters in the 4 x 2 x 2 design:

- Quantifier: nobody, hardly anyone, no NP, an NP
- Clause order: matrix > adverbial, adverbial > matrix
- Adverbial clause type: CAC, PAC

The materials had the following form. In these sentences the quantifier was always in the matrix clause, the variable always in the adverbial clause.²

(8) Temporal adverbial clause: CAC
   a. Matrix clause > adverbial clause (Mq Av)
      {Nobody/hardly anyone/no Italian} wants to go to church while they still have sins on their conscience.
   b. Adverbial clause > matrix clause (Av Mq)
      While they still have sins on their conscience, {nobody/hardly anyone/no Italian} wants to go to church.

(9) Adversative adverbial clause: PAC

²Note that the examples with a positive existential quantifier are omitted here reasons of space. They require the insertion of a negative particle into the clause in order for the sentences to make sense. So *No Italian wants to go to church … becomes An Italian doesn’t want to go to church …*
a. Matrix clause $>\$ adverbial clause (Mq Av)
   \{Nobody/hardly anyone/no Italian\} wants to go to church while they readily claim to believe in God.

b. Adverbial clause $>\$ matrix clause (Av Mq)
   While they readily claim to believe in God, \{nobody/hardly anyone/no Italian\} wants to go to church.

This was a judgement study which used the Thermometer Judgements method as above and which was made accessible on-line on our server at Tübingen University in four counterbalanced versions. The 36 participants were collected by advertising the study in Linguistics departments in universities in the UK. Three prizes of £50 were distributed by lottery among the participants.

The results are illustrated in Figure 3. The error bars show the means and 95% confidence intervals of the normalized judgement scores by condition.\(^3\) The conditions are grouped by the type of quantifier on the x axis. Within each group, the pair on the left (with square mean markers) are PAC conditions and the pair on the right (round mean markers) are CACs. Within the pairs, the left-hand error bar (with filled mean marker) shows the matrix $>\$ adverbial clause order, while the right-hand error bar (with empty mean marker) show the adverbial $>\$ matrix clause order. On the right-hand end of the chart we see the standard items, whose mean points are marked with diamond shapes.

What therefore do these results show us? Above all, the PACs behave differently to the CACs, which was the core question in the experiment. Consistently across the different quantifiers, the PAC pairs of bars with squares are always judged clearly weaker than the CAC pairs of bars with circles. Within these pairs, the relationship between the filled and hollow markers is also different: while for the PACS, the bars with filled markers are never better than the bars with hollow markers, the reverse is true for the CACS.

These results support several observations. First, that our informants permit variable binding differentially into integrated and non-integrated clauses, which would support the claim that they have different attachment points, if we assume that binding is dependent on a structural relation such as c-command. Since the CACs are judged consistently better, it seems natural to assume that these subordinated clause types are embedded below the quantifier in the matrix clause. This does not appear to be the case in the PACs, which are judged clearly worse.

Second, that the CACs and PACs respond differently to the order of the clauses. This is an effect that we have observed in other experiments contrasting temporal and adversative while: there is an ordering preference asymmetry. Temporal while seems to be preferred following its main clause, adversative while preceded-

---

\(^3\)The judgement scores are normalized to z-scores in order to reduce the inter-judger variation. Each person’s scores then have the mean value zero and the standard deviation one.
ing it. Given some thought, this effect appears plausible. Temporal *while* merely modifies a main clause event, adding the specification when it takes place. It is essentially subordinate information, which does not require the salient preceding position. Adversative *while*, on the other hand, structures the discourse and tells us how to interpret that which follows, namely as a contrasting pair of facts. It is credible that it eases the processing of the discourse if this information is provided early on. Relating this to our examples in (7) above, this would mean that the preferred versions are as in (10). If this interpretation is correct, it is further evidence that informants process CACs and PACs differently, in this case in terms that we might think of as relating to information structure.

(10) a. CAC: *while* clause preferred following

Jean mentally prepares her lectures while she runs in the park in the morning.

b. PAC: *while* clause preferred preceding

While Jean runs in the park in the morning, she mentally relaxes in the evening.

---

4Interestingly, this effect seem to be specific to English. It does not visibly appear in German with the equivalent connector *während*.
There are further distinctions in the data, in particular across the quantifiers, which we shall mention though we cannot here discuss them in detail (von Wietersheim 2016 for discussion). Roughly, we may say that the data from the different quantifiers behaves in a parallel but not identical manner, so if our aim is to distinguish CACs from PACs, we may use any of these quantifiers. In fact there is a tendency for the differences to reduce as the semantic content of the quantifier becomes less exclusively negative. *Nobody* produces the most polarized results, *hardly anyone* slightly less so, and *no NP* less again. *An NP* is similar to this last one.

To see this, we look first at the results for quantifier *nobody*. The condition predicted to best permit variable binding is $\text{Mq Av CAC}$, which is the first in (8), and in fact it is judged better than the equivalents with the quantifer phrases *hardly anyone* and *no NP*. But the minimally different condition with the adverbial clause preceding the matrix clause $\text{Av Mq CAC}$ is relatively worse for *nobody* than for the other quantifiers. The PAC conditions too are worse for the quantifier *nobody*. Nevertheless, the basic patterns are the same across the quantifiers: *nobody* produces not different results, but more extreme results.

One way of accounting for this effect, which we have found for German negative quantifiers too, would be in terms of pragmatic preferences on antecedents. It has been observed at least since Sag & Pollard (1992) that specific, non-abstract, and animate antecedents are preferred in binding constructions. The effect observed here may be related to this, but unfortunately we cannot discuss this further here, since it takes us too far from our main aim in this paper.

### 4.2 Standard items I

We are reporting this study here as an example of the use of the standard items, so it is to this that we now turn. How do they behave in this study? The standard items are on the right-hand side of Figure 3. We should first note that they form a pleasingly regular pattern - our attempts to produce a five equally-spaced reference points have thus apparently been successful. There are also no experimental conditions which lie beyond the standard items, so we have no reason to think that our scale does not cover the whole range of accessible values.

What therefore is the added value of the scale? One advantage of having the scale is to demonstrate that the differences between the conditions are fairly minor: nothing is as bad as the C value on the scale and nothing is as good as A. This restricted range of the results only becomes apparent when we have an external comparison scale which shows the full range available.\(^5\)

\(^5\)We might note in this context that result charts like figures 3 and 4 also demonstrate the fine granularity of differences detectable by this methodology. The lengths of the error bars show the degree of accuracy of group judgements; the standard items show the accessible range of perceived well-formedness that these judgements can be given over. We might estimate on the basis of this data that group judgements can distinguish 15 to 20 different degrees of well-formedness, a
4.3 Example experiment part II

We mentioned above that this experiment had some more conditions which we have so far omitted for clarity of exposition. All the conditions up to now had the quantifier in the matrix clause, but in fact we also tested some conditions in which the quantifier was in the subordinate clause. In our experiments on binding as a diagnostic of clause integration (e.g. von Wietersheim 2016) we normally include this additional condition, but it is not easy to achieve with negative quantifiers because the relocation of the quantifier relocates the negation too, and this changes the meaning radically. For example, changing the position of the negative quantifier in (11) alters the meaning completely, so that neither (11-b) nor (11-c) can be minimal pairs to (11-a).

(11) a. No student lies in bed when they could go to the library.
   b. They lie in bed when no student could go to the library.
   c. When no student could go to the library, they lie in bed.

For this reason we do not test this additional condition with the negative quantifiers, but only with the existential quantifier an NP. The full set of conditions with the existential quantifier thus looks like this - (12), (13)

(12) Temporal adverbial clause: CAC
   a. Matrix clause + quantifier > adverbial clause + variable (Mq Av)
      An Italian does not want to go to church while they still have sins on their conscience.
   b. Adverbial clause + variable > matrix clause + quantifier (Av Mq)
      While they still have sins on their conscience, an Italian does not want to go to church.
   c. Matrix clause + variable > adverbial clause + quantifier (Mv Aq)
      They don’t want to go to church while an Italian still has sins on their conscience.
   d. Adverbial clause + quantifier > matrix clause + variable (Aq Mv)
      While an Italian still has sins on their conscience, they don’t want to go to church.

(13) Adversative adverbial clause: PAC
   a. Matrix clause + quantifier > adverbial clause + variable (Mq Av)
      An Italian does not want to go to church, while they readily claim to believe in God.
   b. Adverbial clause + variable > matrix clause + quantifier (Av Mq)
      While they readily claim to believe in God, an Italian does not want to
go to church.

c. Matrix clause + variable > adverbial clause + quantifier (Mv Aq)
   They don’t want to go to church, while an Italian happily claims to
   believe in God.

d. Adverbial clause + quantifier > matrix clause + variable (Aq Mv)
   While an Italian happily claims to believe in God, they don’t want to go
   to church.

The full result set is illustrated in figure 4. The only changes are to the group of the *an NP* quantifier.

The additional conditions with the quantifier in the subordinate clause pattern rather differently to those we have looked at so far. Their error bars have triangles and rectangles, filled and hollow. The *Mv Aq PAC* and *Mv Aq CAC* conditions in which the quantifier is not only in the subordinate clause but also follows the variable (filled triangle and rectangle), are the worst of all by some way, below standard item C. The cases where the quantifier precedes the variable (hollow triangle and rectangle) are much better, roughly at level B, about the same as the PAC conditions (squares). We would highlight this as illustrating the importance of linear ordering, when - and only when - the parser is attempting to pragmatically repair failed binding. If we look at the CAC conditions across the quantifiers, we see
that the order $MqAv$, in which the quantifier precedes the variable, is judged more natural than the inverse order. But crucially the difference is not all that large. The error bars with filled and hollow mean markers are between the A and B standards items. It is easier for us to process sentences with quantifiers preceding variables, but syntactic binding is quite possible in the other direction too.

This is very different from the conditions that we have newly introduced, where syntactic binding is quite impossible. In these conditions the linear order of quantifier and potentially bound variable makes a very large difference: the error bars with filled and hollow triangles and rectangles are far apart. This finding is an indication that these examples are being dealt with as best as possible by a process of pragmatic repair. For syntactic binding, what counts is the hierarchical relationship; for pragmatic repair processing, factors such as linear order play a large role too.

4.4 Standard items II

These additional examples are intended to show further aspects of the usefulness of the scale of standard items. We have been able to label degrees of well-formedness in absolute terms by using the standard values. We have also been able to show varying degrees of difference between conditions. Being able to recognize degrees of perceived well-formedness in absolute terms permits us to put forward arguments such as the distinction between syntactic and pragmatic effects that we advance here.

5 Standard items in semantic studies?

A question often raised about the use of standard items is their applicability to studies concerning factors other than syntactic well-formedness. These items were originally developed for use within the field of syntax and the range of well-formedness that they cover is that which is accessible to syntactic intuitions (Featherston 2009). Can they therefore sensibly be used in studies where the variable at issue is one of semantic coherence or pragmatic felicity? Can we map intuitions of coherence and semantic feasibility onto the same scale? This is an interesting question, because it seems fairly uncontroversial that syntactic well-formedness (often termed, for better or worse: ‘grammaticality’) is not the same as semantic felicity. In many cases the two are clearly distinguishable: there is nothing wrong in strictly syntactic terms with ‘a married bachelor/spinster’, ‘I am in the process of loving you’, or ‘the old man died the oak tree’, but their (potential) meanings do not fit our world. These examples would thus appear to contain semantic mismatches. On the other hand, we can readily make sense of ‘You should can swim after all these lessons’, ‘I’m afraid I seen’t the problem’, and the question ‘Go you home now?’ but we cannot phrase them that way in current
English. These therefore would seem to be problems of the syntactic form, not the semantic or pragmatic content. Since this is the case, it must be clear that the two types of *can* in principle be distinguished. Given that they can in some cases be distinguished, they cannot be the same thing. It might well therefore be expected that syntactic well-formedness and semantic coherence and/or felicity should not be mapped onto the same scale.

Nevertheless, we argue here that it can indeed be useful to include the set of standard items into experiments which address issues of interpretation. One reason is simple: the boundary between form and interpretation is highly permeable, so that the two depend on each other. Work on judgement studies has made it clear that introspective judgements are sensitive to both: a well-formedness judgement is always a judgement of a structure *in* a certain interpretation. Speakers are quite incapable of giving judgements of strings that they do not understand, and judgements of ambiguous structures with multiple roughly equally accessible readings will show more variability. The problem is thus less acute than it might at first glance appear. In fact the study reported above is not purely syntactic, but rather addresses an issue with both syntactic and semantic features. Binding, particularly variable binding, appears to have syntactic structural preconditions, but it relates to an aspect of interpretation. So the example study we have advanced here is itself partly a semantic one, but the standard items produce the usual tidy pattern, which would seem to confirm that the participants in the experiment had no problems judging them.

Nevertheless, experience has shown that there is a difference between syntactic and semantic studies in the use of the standard items, and it lies in the absolute values that the standards represent. In semantic studies, the well-formedness judgements rarely descend to the lower values on the scale - just as we observed in the example study reported above. This might appear puzzling, but we have a suggestion why this might be the case: whereas a syntactic violation produces a form which is impossible, a semantically ill-formed example can generally (perhaps: always?) be repaired by changing our perspective on the world or imagining a figurative interpretation. For example, ‘a married bachelor’ could be a husband with a BA, and ‘a married spinster’ could be a wife who has characteristics that are conventionally associated with spinsters. Similarly, the statement ‘I am in the process of loving you’ could be uttered by a over-rational robot such as Star Wars’ C-3PO. Semantic failures can thus often be saved by an act of imagination. For this reason, we hypothesize, these examples tend not to be judged as entirely unacceptable on the standard scale. The absolute values of the scale therefore might require re-anchoring when we use them in semantic studies.

This, however, is the only aspect of their usefulness which gets lost; all the other advantages remain. We therefore strongly recommend their use in studies of both types, syntactic and semantic. We hope we have persuaded you, dear reader,
to do so.

6 Appendix

The candidate examples we tested were the following. The numbers refer to the sentence numbers in Figure 2 above.

Candidates for standard item group A
6: The patient fooled the dentist by pretending to be in pain.
50: There’s a statue in the middle of the square.
53: The winter is very harsh in the North.
55: The tired teacher drank a cup of tea in the staff room.
52: It’s nice to eat pizza and watch a film.
51: A hammer is a useful thing in a tool box.

Candidates for standard item group B
54: The best thing about a sister is you can borrow her clothes.
13: Before every lesson the teacher must prepare their materials.
23: What I need to know is which witness identified which defendant.
15: Jack doesn’t boast about his being elected chairman.
21: Jack cleaned his motorbike with which cleaning cloth?
25: A thousand dollars are a lot of money to pay for a flight.

Candidates for standard item group C
26: What I want to know is which exam which student failed.
24: Hannah hates but Linda loves eating popcorn in the cinema.
16: Most people like very much a cup of tea in the morning.
27: You can’t say that to me, who does most of the work here!
32: The striker must have fouled deliberately the goalkeeper.
18: The estate agent revealed him the whole plan to build a gas works.

Candidates for standard item group D
37: That boy had run away from home, who we saw in the bus.
44: Who did he whisper that had unfairly condemned the prisoner?
63: The old fisherman took her pipe out of mouth and began story.
39: In southern Spain heavy rain falling only in the mountains.
30: Which professor did you claim that the student really admires him?
47: The author looked in the dictionary up the word.

Candidates for standard item group E
64: Crossing the wild seas Odysseus was meet some monster.
29: Who did you wonder when Michael will introduce to his parents?
62: The Spartans brave stood in line fight for freedom.
7 References


Frey, Werner (2011) Peripheral adverbial clauses, their licensing and the prefieId in German. In Eva Breindl, Gisella Ferraresi & Anna Volodina (eds.), *Satzverknüpfung – Zur Interaktion von Form, Bedeutung und Diskursfunktion,* 41-77. Berlin: de Gruyter


65: Historians wondering what cause is disappear civilization.
60: Old man he work garden grow many flowers and vegetable.
61: Student must read much book for they become clever.


