

The Decathlon Model of Empirical Syntax

Sam Featherston, Tübingen University

Abstract

This article summarizes the findings of some of our studies of the data base of syntactic theory, contrasting the characteristics of frequency data and judgement data. Examination of frequency data reveals that the factors affecting its production interact competitively and probabilistically. This contrasts strongly with the patterns observed in judgement data, which point to a system in which violations of constraints produce negative weightings on form/meaning pairs. Since both data types are the result of human linguistic processing, we present a model of the architecture that such a system might have in order to produce such contrasting data. This Decathlon Model has two modules: Constraint Application and Output Selection. The first is blind, exceptionless and applies violation costs cumulatively (Keller 2000), the second is competitive and probabilistic. This constrains frameworks of syntactic explanation: an empirically adequate grammar must include gradient well-formedness, specify constraint violation costs, and distinguish between the application of rules and the selection of outputs.

In this paper reports our investigations into the data base of syntactic theory, specifically addressing the similarities and differences between corpus data and judgements and sketching the implications for the construct of grammaticality and the architecture of the grammar which our findings have. The motivation for these studies was a dissatisfaction with the state of affairs in syntax, when, for example, two syntacticians can look at the same phenomenon and come up with widely differing analyses of what is going on. Another disappointment is the lack of any real forward movement in theory: alternative analyses seem to succeed each more due to fashion than due to falsification. We might say that syntactic description, let alone syntactic explanation, is underdetermined by its data base.

In fact most data feeding into syntactic theory has significant flaws: it is fuzzy, it reflects multiple factors, only some of which are relevant to theory, and perhaps worst of all, these factors are poorly understood and not clearly

distinguishable (Schütze 1996). Judgements have been particularly criticized as a data type, partly because of their inherent qualities, but partly for the way that they have been used (eg Labov 1996). One problem is that, faced with the impreciseness of judgement intuitions, researchers have idealized the data type to a very great degree, reducing the scale to a binary opposition, with marginal values as unclear cases. More problematic, they have failed to establish any clear process or criterion for idealization, so that the distinction between data relevant and irrelevant to syntax is close to being arbitrary. Second, researchers have tended to assume that any individual's judgements are a sufficient basis for theoretical work, ignoring the conflict of interest between the linguist as informant and the linguist as analyst, and positing that any difference in judgements implied divergent idio-grammars. In part as a response to this situation, some syntacticians have widened their perspectives and sought other data sources, such as corpus frequencies and processing studies, which has tended to split the field even more and furthered the development of schools of syntax, who have neither a common formalism nor a common data base which might serve mutual rapprochement.

In part as a result of this diversification of data types, a range of different grammar architectures have arisen. Generative syntacticians most commonly still use judgements, and assume a "live rail" grammar, in which any infringement of a grammatical rule causes a structure to be excluded absolutely. Those interested in competition models such as Optimality Theory (OT, Prince & Smolensky 1993) will tend to use frequency data and allow some idealization, while those favouring probabilistic models will tend to take a more fine-grained approach to frequency data, and account for the variants using probabilities (eg Manning 2003). Those taking a "grammar-is-the-parser" position will naturally place most faith in processing studies. Since they have no common data base, no common definition of what it means for a structure (not) to be part of a language, and widely divergent views of the architecture of the grammar, these schools have little basis for dialogue.

Our view is that a way forward on this problem lies in the more detailed study of data types and their characteristics. When we have a more detailed understanding of the factors which each reflects and in what proportions, and which parts of human linguistic processing each reflects, then we shall be in a better position to judge the evidence of each for syntax. This should allow us to establish a well-founded procedure for idealization for each. Additionally, when we know each data type better, we shall be able to establish the similarities and differences among them, and be more able to weigh up the strengths and weaknesses of each. In the following we first sketch the sort of studies we

have undertaken, outline the broad picture of their results, and then move on to the implications for the nature of the grammar of these findings. Note that this article aims to provide an overview of our results and our interpretation of their implications, space does not permit discussion individual experimental or corpus studies (see Featherston 2002a, 2002b, 2003a, 2003b, 2005).

1 Our studies

We have carried out a number of studies comparing frequency and judgements, aiming firstly to clarify these issues of data and data type, secondly to clear up outstanding questions in the syntax, and thirdly to clarify the nature of the grammar. We have performed experiments on German and English, and have addressed a range of syntactic structures, among others island constraints, reflexives, reciprocals, word order, parentheses and echo questions.

Our frequency data for German is drawn the COSMAS corpus of German (IDS, Mannheim), and for English, the British National Corpus (Oxford), but we are actively experimenting with the use of web search engines for our studies (following Keller et al 2002). We have generally elicited our judgement data using the magnitude estimation procedure (Bard et al 1996). This procedure has three main differences to standard judgement elicitation. First, only relative judgements are gathered: subjects are asked whether example A is better or worse than structure B, and by how much. No absolute criterion of well-formedness is used but judgements are expressed in numerical form (we lay out the importance of the distinction between relative and categorical judgements in 5 below). Second, to anchor the scale, subjects give judgements relative to a reference item and to their own previous judgements. Third, there is no imposed scale; no top or bottom limit nor minimum division between scores. The net result is to allow informants to express all the differences they perceive, with no interference from the scale. An important point to bear in mind when considering this data is that these are not categorical judgements, in which informants are asked to decide whether a structure is or is not acceptable, but rather relative judgements, in which they express their intuitions about “better” and “worse”.

While magnitude estimation has proved itself very effective, we have now moved on to developments of this approach which avoid prescribing that subjects should produce a magnitude scale, in the light of our own findings that they are do not in fact do so, even when instructed to (see also Poulton 1989). In fact, informants produce something close to a linear scale whether their instructions point them towards a magnitude scale or towards a linear scale (see

figure 1 below).¹ This fact will be important when we look at the patterns of judgement data below.

2 Rethinking abstraction

We noted above that the lack of any common approach to deciding what data is and is not relevant to syntax is one factor hindering the development of a common data base of syntactic theory. Our studies have only underlined this lack. For example, we found that the superiority effect, which disallows in-situ wh-subjects in multiple wh-questions, was regarded as grammatical in nature, but the identical data pattern exhibited by non-wh-NPs was thought to be merely stylistic. Cross-linguistically, superiority was analyzed as grammatical in English but merely a markedness effect in German, in spite of producing almost exactly the same results. Our studies on binding and reference showed that much of the differentiation between binding possibilities among complements could be attributed to surface factors, such as heaviness, and required no recourse to grammatical factors such as reconstruction.

In short, it has become clear that syntacticians are in fact not able to distinguish between syntax-relevant and syntax-irrelevant effects, but chiefly rely on two criteria. The first is tradition: if it has once been assumed that a phenomenon is or is not syntactic, then this assignment is generally adhered to. The second factor is categoricity. Since the scale of well-formedness has been abstracted to a dichotomy, any syntax-relevant effect must necessarily cause categorical ill-formedness. Any factor which does not exclude a violating structure absolutely is often rated as mere markedness (cf Bresnan et al 2001, Müller 1999, Aissen & Bresnan's *Stochastic Generalization* 2002). The result has been that research has focused on harder constraints, and much effort has been made trying to formulate these hard constraints so as to account for the cumulative effects of weaker constraints (Keller 2000, see 3 and figure 1), which are invisible in a dichotomous grammaticality model.

We use the following procedure to abstract syntax-relevant information from experimental results, and I suggest it be adopted more widely.

- (1) a. *Rule 1*
An effect is only syntax-relevant if it cannot be accounted for by known performance factors (heaviness, information structure, frequency, processing complexity, phonotactics, plausibility...).
- b. *Rule 2*
Any effect is syntax-relevant if it cannot be thus accounted for.

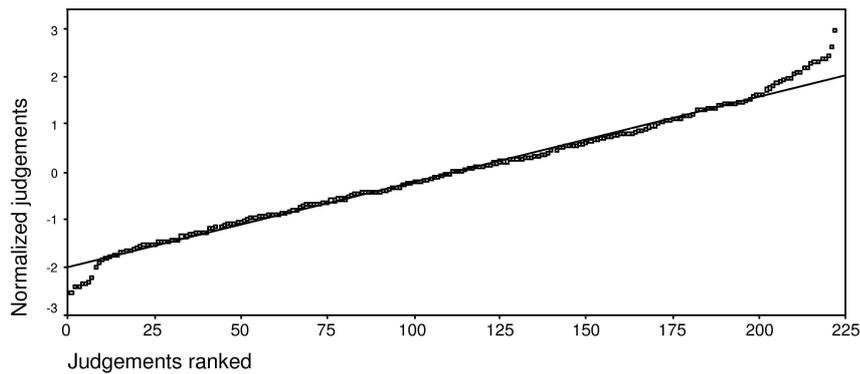


Figure 1: Judgements elicited under controlled conditions produce a linear pattern of well- formedness.

This approach can of course be applied flexibly: if one feels that a given finding is of syntactic interest, even though it could plausibly be accounted for by other factors, one is free to argue the case. But this approach should be the default, deviations from which should require justification. Let us note that this no doubt makes life more difficult for the syntactician, since syntax-relevance cannot simply be assumed, but its aim and no doubt its effect will be to strengthen the discipline by making its arguments more convincing, because they are better grounded.

3 Judgement data and frequency data

Our experimental and corpus studies of the same syntactic phenomena in the two data types revealed that the two data types showed consistently divergent patterns. In particular, judgements reveal a very different pattern to that often assumed. Firstly, well-formedness is a continuum. Figure 1 shows the results from a magnitude estimation experiment of about a thousand data points. On it judgements are mapped in ascending order on the x axis, while their well-formedness values (normalized by subject) are shown on the y axis, with up representing “more natural” judgements. Informants show no sign of using categorical well- formedness, when given the option not to. There is absolutely no bunching of judgements at the top or the bottom, and indeed the outliers show exactly the opposite tendency (see endnote 1).

For the further features we turn to figure 2. Frank Keller (2000) has dealt with this type of data and pointed out many of its features, especially the

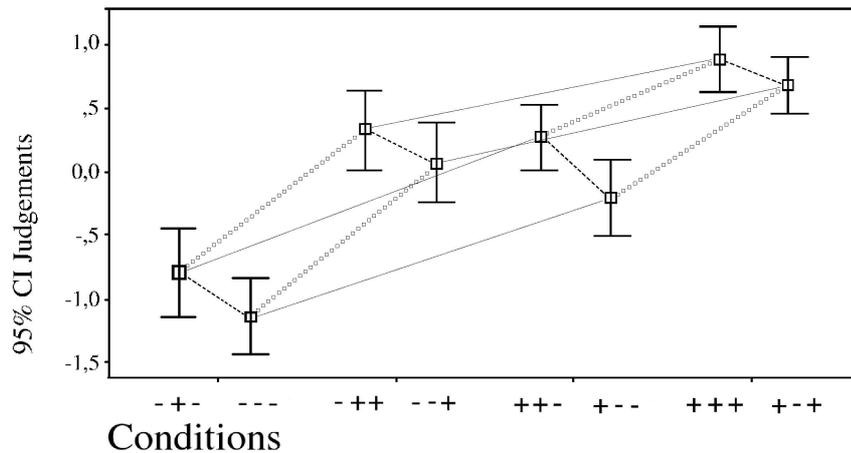


Figure 2: Judgements reveal that violation costs are quantifiable, cumulative, and applied blindly, but survivable.

quality of *cumulativity*. This figure shows the results obtained by testing eight structures in a 2x2x2 experimental design looking at object coreference. The graph shows the conditions on the x axis and the mean judged well-formedness on the y axis, with higher scores again reflecting “more natural” judgements. The error bars show the mean values and 95% confidence intervals of the scores for each condition. Each pair of error bars linked by a line is a minimal pair differing only in one condition, in each case one of the pair violates a constraint that the other fulfils.

It is evident that each factor differentiating a minimal pair has a consistent effect upon the judgements: the relationship between the scores assigned to each pair is the same. We can thus conclude that each linguistic factor has a fixed effect upon well-formedness. It is also clear that the starting point before the application of the additional constraint is irrelevant: constraint application is blind. A third important point is that there is such a thing as a quantifiable constraint violation cost. The pairs linked by the short broken line differ in their ratings only weakly, which shows that that particular constraint has a relatively small violation cost. The other two have greater violation costs, but neither of them could be said to be a “hard” constraint, that is, one whose violation automatically drops the violating structure to the bottom of the scale, wherever it started from. Experience in such experiments has shown that there is in fact no such thing. The effect of a violation is only ever

ihn _i ihm _i	(“him.ACC him.DAT”)	0 hits
ihm _i ihn _i	(“him.DAT him.ACC”)	0 hits
ihm _i sich _i	(“him.DAT REFL.ACC”)	0 hits
ihn _i sich _i	(“him.ACC REFL.DAT”)	1 hit
ihn _i ihm _i selbst	(“him.ACC him.DAT SELF”)	0 hits
ihm _i ihn _i selbst	(“him.DAT him.ACC SELF”)	0 hits
ihm _i sich _i selbst	(“him.DAT REFL.ACC SELF”)	0 hits
ihn _i sich _i selbst	(“him.ACC REFL.DAT SELF”)	14 hits

Table 1: Data from COSMAS, IDS, Mannheim (531 million word forms)

to make a structure worse, by an identifiable amount; no constraint violation makes a structure so bad that it cannot be made worse by an addition violation, and no violation excludes a structure from being part of the language.

We refer to this quality of linguistic constraints as *survivability*, which is best understood in contrast to the OT concept of *violability*. Violability means that under certain circumstances, constraints have no effect on the output, that is, they fail to apply. This is in part necessary because the only effect that a constraint can have in OT is to exclude categorically the violating structures. Our survivability means that all constraints always apply, exceptionlessly, and a given violation always has the same effect – there is no probabilistic element at all. The effect of a constraint violation is to cause a structure to be judged worse, but no violation in itself excludes a structure.

In the light of the finding that violation costs as measured in judgements of well-formedness are cumulative, survivable and blindly applied, on the one hand, but not directly related to output frequency (see below) on the other, it seems reasonable to assume that these violation costs, and hence well-formedness as measured by judgements, are related to *computational workload*. This raises real questions about what psycholinguistically plausible mechanism might allow us to convert cognitive workload into judgements, and why we have such an ability. We touch on these questions in section 5 below, but space prevents an extended discussion here.

Frequency data reveals a very different pattern. Table 1 contains a typical data pattern of a frequency study, here too looking at the different lexical instantiations of coreferent direct and indirect object structures. The important point here is the distribution of forms found in the corpus: one structure is found fourteen times, another one is found once, but none of the others appear at all. Frequency is a function of linguistic production, and so the

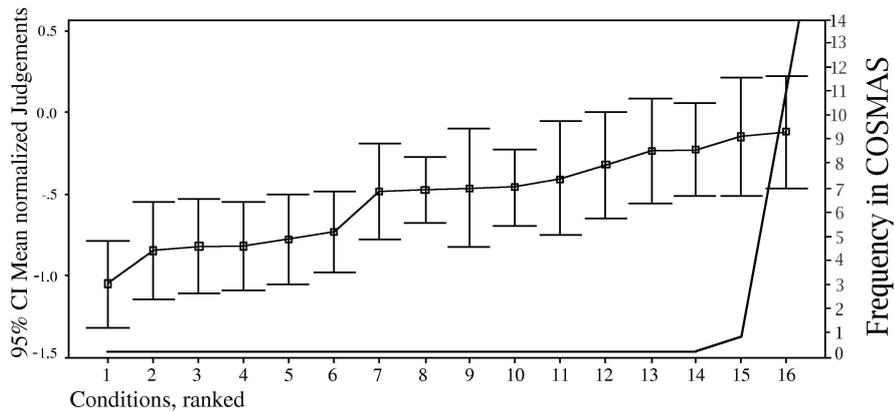


Figure 3: The contrast between frequency and judgement data on the same phenomenon.

minimal unit which it measures is occurrence or non-occurrence, a binary opposition, since a given expression in any given utterance either occurs or does not occur. Unlike our judgements, which are made up of numerical scores, frequency data is made up of the sums of many individual data points, each representing a decision to produce or not to produce a given form. It is thus not surprising to find that frequency data shows evidence of a competitive interaction of candidate forms, which would seem to indicate that the “best” structure of a comparison set usually wins through to be produced. Interestingly, slightly less “good” alternatives are sometimes produced, which would suggest that the competition for output functions probabilistically.

Figure 3 allows us to compare the two data patterns directly, as it superimposes the two different measures of the same sixteen structures on a single graph. The error bars show the mean normalized judgements obtained for the sixteen structures tested (left-hand scale). These can be seen to increase steadily from the very bottom to the very top, while the frequencies, represented by the line without error bars, creep across the bottom at zero, and only rise sharply at the right-hand end (see the right-hand scale). The comparison of these two measurements of the same structures brings the contrast of the data patterns into sharp focus. The first point to notice is their similarity: the same structures come top in both data types. The highest frequency structure is judged best and the next highest is judged second best, which makes it seem likely that the two data types are at least in part measuring

the same underlying factor. But we should also note the key difference: the judgement data demonstrates that at least some part of the human linguistic computation mechanism is sensitive to differences among structures which are so bad that they would never be produced. Since this is the case, it is plain that the two data types are also in part *not* measuring the same factor. For a fuller understanding of the nature of the evidence provided by frequencies and judgements it would be necessary to specify what process differentiates the two types of results. We attempt to do this in our Decathlon Model of grammaticality and the position of the grammar.

4 The Decathlon Model

The name of this model derives from the athletic discipline of the decathlon. In this event, competitors take part in ten different sub-disciplines, and their performances are converted into a numerical form. The sum of these scores decides who wins the medals. But the scores are calculated not on their *relative* performance in the sub-disciplines, but in their *absolute* performances, which means that whether an athlete comes first, second, or third in the 100 metres is of no significance, what matters is that they perform at their personal best. In a sense therefore, they are not so much competing against each other at this stage as against themselves. Competition *between* competitors takes place at the second stage, where the ten numerical scores are totalled, and the highest scorer takes the gold. Something similar seems to us to be happening in human linguistic processing, as will become clear in this section.

The Decathlon Model is at once an outline architecture of a constraint satisfaction model of grammar and at the same time an account of the differences between data types. Our finding that gradience reflects a real psychological phenomenon related to constraint violation cost (see section 3) demands that the architecture of syntax reflect this reality, which current models do not do. An empirically adequate and psychologically real grammar must have the following features: quantifiable violation costs, a continuum of well-formedness, and survivable constraints (ie no constraint violation necessarily results in the categorical ungrammaticality of the violating structure); all this to account for our judgement data. It must also generate output competitively and probabilistically so as to reflect the data patterns observed in frequencies.

The obvious way to achieve this is for our syntax model to distinguish between a grammatical module which applies syntactic constraints and another which selects output. Our Decathlon Model thus has a Constraint Application module, which applies constraints, assigns violation costs, and outputs

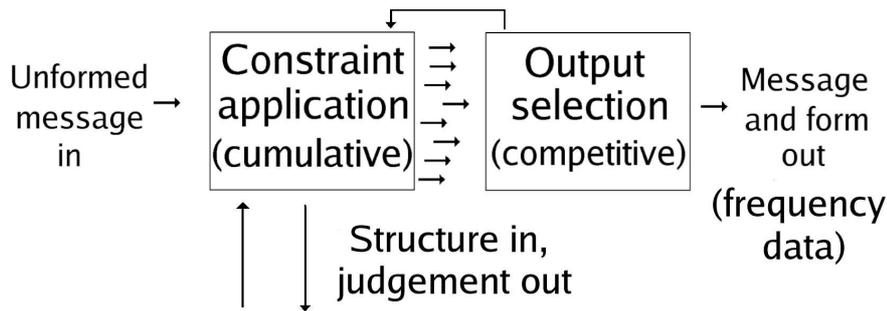


Figure 4: The Decathlon Model of the grammar and grammaticality.

form/meaning pairs, weighted with violation costs. We know certain things about the internal functioning of this module: constraints are applied blindly and exceptionlessly, and violation costs are cumulative. We may think of this module as containing the grammar, though it also contains the other factors which affect well-formedness in judgements.

The second module, Output Selection, functions quite differently. Its task is to select from the possible form/meaning pairs the form which is to be output (in production processing) or the interpretation to be assigned to an input (in receptive processing), and exclude the others. It functions competitively and selects the best candidate on the basis of the weightings assigned by the Constraint Application module, excluding all others. This selection occurs probabilistically however, which accounts for occasional production of sub-optimal versions: rare but documented counter-examples are thus no threat to grammatical generalizations.

In figure 4 we see the processing patterns which generate frequency data and judgements. In production we assume that an unformed message is delivered for formulation in the Constraint Application module. Incrementally, perhaps phrase by phrase, candidates for the linguistic representation together with their weightings are proposed to the Output Selection function, which selects the best, or one of the best. The arrows exiting the left-hand module show the candidate continuations of the structure passing to the selection module, their weightings represented by their offset positions.² Sometimes two continuations will be roughly equally good: *She turned the light off* vs *She turned off the light*, in which case both will occur.

Receptive processing is similar, except that the system is incrementally attempting to assign a structure to the part of the message string already re-

ceived as well as assigning a meaning to it on the basis of the structure so far assigned. Giving judgements is a little different. The example is input processed as usual to determine its structure and meaning, but instead of returning the output of the selection module, relative judgements consist of returning the output of the Constraint Application function, that is, the weightings it assigns. This of course requires the claim that the output of this module can be consciously accessed, as well as merely passed on as usual for selection. The capacity to be aware of fine-grained cognitive workload is not something which we might have predicted for ourselves, but it is nevertheless not implausible, since we are certainly aware of more coarse-grained thinking effort. The difference between frequency measures and relative judgements can therefore be attributed to them being the outputs of two different modules of linguistic processing, which are independently motivated.

This model has a number of explanatory advantages. First, it is firmly based on the primary data of syntax. It accounts for the differences in outcome patterns between data types, an outstanding question in linguistics. Frequency data reflects the output of the Output Selection module, which is (necessarily, since we produce only one form of an utterance) competitive. Since this module uses the weightings output by the Constraint Application module, we account for the fact that judgements and frequencies agree in identifying the same forms as optimal. These weightings are themselves functionally motivated by their identification with computational complexity, an explanatorily economical association, since we know of the existence of workload effects from other sources, and any other approach is left with the additional task of identifying what other factor judgements are a reflection of.

Second, it ties the grammar in to evidence from human linguistic processing. It is consensual that syntactic processing operates on-line, incrementally, and applies information from multiple sources in order to take decisions. It has often been suggested that the processor consists of a rule component and a decision component which prunes less optimal interpretations or outputs, often phrased as the distinction between *structure building* and *structure checking* (see Featherston 2001 for discussion of parser types). Our model is immediately compatible with such evidence from processing as garden paths and the fault-tolerant quality of syntactic processing. We understand faulty utterances and occasionally produce structures we know to be ill-formed – this model's distinction of constraint application and output selection permits these well-documented characteristics to be accounted for.

A third strength is that it provides some explanation of the wide variation in grammar architectures that we find competing in linguistic theory. Most

of these capture a part of the fuller picture that we have sketched: generative grammar has traditionally concerned itself with rules with strong violation costs. Since it is generally true that breaking any given heavily-sanctioned rule makes it unlikely the offending structure will occur, generative grammar has an *live rail* model of grammar, in which any infringement excludes a structure categorically. This has led researchers in this framework to assume that less strongly sanctioned factors cannot be “grammatical” in nature, since they do not conform to this expectation that any broken rule forces instant exclusion. These perspectives on theory structure and data reinforce each other. OT by contrast is entirely committed to competition, motivated by the insight that it is generally the best of any competing set of structural alternatives which is produced. This too reflects a real aspect of the empirical data: the process of selecting a form to produce necessarily results in a competitive interaction and thus blocking effects. But this confuses production with the grammar, which is also the cause of the problems that OT has in defining and generating comparison sets. In the Decathlon Model, competition is not confined to structural alternatives, since it is not a part of the grammar but part of the processor. Probabilistic grammars too are partly right: there is indeed a probabilistic component in the linguistic processor, although our relative judgements demonstrate that it too is not part of the grammar.

Fourthly and lastly, our view of the position of the grammar in the wider picture of syntactic processing allows the syntax to cover a much wider range of phenomena. Such issues as linguistic variation and language acquisition can be well accounted for in a model with exceptionless constraint application but a parameter of violation cost strength. For example, Aissen & Bresnan’s (2002) *Stochastic Generalization* (see also Bresnan et al 2001) notes that similar constraints may be found cross-linguistically, but appear to be grammatical and categorical in one language while being mere statistical tendencies in another. We have a ready account of these facts: the same factors exist across languages, but their violation costs vary, due to the interaction of constraints (for the superiority effect in German and English as an example of this, see Featherston 2005). Not only the differences between languages, but also regional, sociolinguistic and even idiomatic variation can be encoded as differences in violation cost amplitude, since these differences are not dichotomous but points on a continuum for each constraint. The learning of these violation strengths can be seen as a part of the acquisition of syntax.

Our model thus offers a far wider view of the linguistic environment than most approaches to syntax. In this it bears a resemblance to the syntax of the sixties and seventies, when questions about the position of grammar in more

general cognitive functioning were a standard issue for syntacticians. More recently they have tended to see their role as developing grammars within a psycholinguistic framework which in the meantime has become not merely a consensus, but rather a part of the set of basic assumptions of syntactic theory. The investigation of syntactic analyses within this conceptual space has become the role of the syntactician, rather than the questioning of the shape and extent of the space itself. In our work we have aimed to re-open this debate, and revisit these assumptions in the light of the new data available.

4.1 Wellformedness does not trigger occurrence

Our model is also supported by data from the interaction of well-formedness and occurrence. The standard assumption is that the functions of constraint application and output selection are not to be distinguished, and that they take place with the same module. In generative grammar, this would predict that any structure which is generated and does not violate any constraint on structure is grammatical and may be produced, while in OT the last candidate remaining, the only well-formed one, is produced. Both of these thus assume that production depends directly on the grammar, and that well-formedness determines occurrence. The Decathlon Model however claims that production competition determines output, so that no single level of well-formedness triggers occurrence in the output. In the light of this, consider the results of the experiment in figure 5.

This figure shows the results of a single experiment which contained three separate sets of structural contrasts, with their mean judgements indicated by error bars as before, arranged in ascending order of well-formedness, by structural contrast. Each group is thus competing to represent semantic contents. This is clearest in the set on the right-hand side, where all are competing for a single content, whereas the middle group are competing for three different contents, and the left-hand group are competing for four contents. In each set, those structures which actually occur in the COSMAS I corpus (IDS, Mannheim) are above the line, while those which do not are below the line. It is striking that those occurring appear in a solid block, from the top of the group. This alone is evidence of competition for production, based on the same weighting information which we can access as judgements. However, notice that the best two structures from the right-hand group, which are those which occur in the language, are nevertheless judged *worse* than some of the lower structural alternatives in the other groups, which do not occur. The implication is clear: occurrence is not directly dependent upon well-formedness,

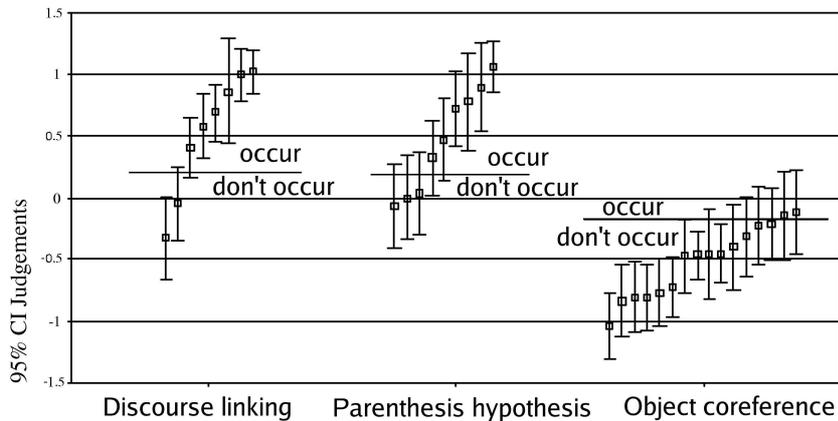


Figure 5: The mismatch of well-formedness and occurrence: Production is competitive.

but rather upon a competition function based on these weightings. This finding supports the distinction of the grammar and the production function, as in the Decathlon Model, but it is not compatible with an architecture in which these two are merged.

5 Categorical judgements and relative judgements

This insight into human linguistic processing offers an account of another outstanding question: Why do judgements, elicited under strictly controlled conditions, show that informants, given a free choice of scale, do not use a binary division or end points which might represent “fully grammatical” and “fully ungrammatical”? Our solution to this quandary is to distinguish the *categorical judgements* commonly used in syntactic work from the *relative judgements* obtained from our experimental studies.

Our assumption of this dissociation is based upon several pieces of evidence. The strongest evidence for the reality of categorical judgements is quite simply our intuition that there are such things as “full grammaticality” (= “I would expect to hear this”) and “full ungrammaticality” (= “I would never expect to hear this”). Every speaker seems to have this, and neither its reality nor its relevance can be doubted: any naive informant, given a binary choice whether an example is good or bad, can immediately make sense of the question. It seems likely that the existence of this intuition is the reason for

the standard linguistic assumption of dichotomous grammaticality. On the other hand, the results of carefully controlled experimental studies such as our own demonstrate conclusively that relative judgements exist too. Further evidence for the distinction is offered by a frequent comment in judgements of sub-optimal structures: “I would never say it, but it is better than the other one”. The frequency of this type of reaction suggests that this intuition too is common to all speakers. With this response the informant is giving both types of judgement information: a categorical judgement and a relative one. This typical comment also gives us a clue about the difference between the two types: categorical judgements concern occurrence, while relative judgements reflect computational cost.

Let us take these in turn. The categorical judgement, we argue, is an expression of the likelihood that a structure *is good enough to occur in practice*. As such it is probably dependent on two factors: firstly, our internal corpus of the language, itself made up of the effects of language exposure, which feeds information into every process which makes use of frequency. The question that the informant is internally answering (at least sometimes consciously) is: “Have I heard structures like this?” The second influence is our Output Selection function. The internal question here is: “Would this structure be produced or is there a better alternative which would block it?”. Categorical judgements thus reflect occurrence, and produce an essentially binary output in the same way that other occurrence-based data types do; a structure either does or does not occur.

The relative judgement, on the other hand, reflects the cognitive workload in processing the form and semantic content of the structure, and relating the two. It reflects the function of the Constraint Application module, and consists of its standard output of a candidate form-and-meaning pair with an assigned weighting. This provides an account of why relative judgements can distinguish between sets of structures which are all seriously ill-formed and none of which would ever occur. Such data cannot possibly reflect occurrence or frequency, since this is consistent across all such structures, but they nevertheless differ in computational workload.

6 The nature of well-formedness

If further work confirms our view that relative judgements reflect computational load and categorical judgements reflect possible occurrence, then a number of important conclusions for the architecture and nature of the grammar would follow. First, the restrictions on linguistic structure which lin-

guists observe are ultimately functionally motivated, since they relate to the factor ease of use, and are also ultimately *emergent*, in that the factors which drive the division into better and worse structures are themselves value-free. It should be clear that this conception of the cognitive roots of grammar has little in common with approaches more generally associated with the label *emergent* (eg Bybee & Hopper 2001), which use the factor frequency as the causal factor in the emergence of structure. Our ambition here is to account for (among other things) structural frequencies, not use them as explanations.

We are associative and competition-driven thinkers: put differently, we are lazy thinkers, and we therefore prefer computationally easier processing tasks. But the processing of every word and every syntactic relation comes with a cost: this can be readily seen in judgement studies, where longer sentences are systematically judged worse than shorter sentences (more words mean more computational load), and examples with pronouns are generally judged worse than those with full NPs (more complex reference means more computational load in comprehension). There is of course nothing “wrong” with longer sentences and pronouns: the interpretation of computational load as “badness” comes only at the stage when linguistic processing has to deal with structural alternatives, and at this stage, forms which incur higher processing costs are dispreferred. Thus longer sentences and pronouns are computationally more costly, but necessary, whereas forms for which a more economical structural alternative existence may only be only a little more costly, but this little difference makes them dispreferred.

This analysis of the nature of judged well-formedness accounts for cumulativeness, violation costs, survivability etc, but at the same time goes some way to explaining why there is evidence for the innateness of grammatical restrictions (architecture-related factors are by their nature innate), and it does this within a psycholinguistically and empirically motivated framework.

7 Implications for data types and their relation to theory

It seems fair to state that a fundamental assumption underlying the use of frequencies as a source of evidence for syntax is that “good” structures are produced, and thus found in corpuses, while “bad” structures are not produced and thus not found. In a second step we might generalize that there is an assumption that better structures are produced more often than less good structures. These assumptions are confirmed by our findings, but they reveal that this is not the full picture: frequencies correlate with well-formedness in judgements among the very “best” structures, but provide no information

about “poorer” candidates, because they undifferentiatedly do not occur.

Or rather, they do not occur in the size of corpus to which we have access. If we are right in our suggestion that output competition is probabilistic, then, in a big enough corpus, we should find not only the best and second-best candidates but also the third and fourth and so on. The fact that linguists are always finding structures in corpuses which they had assumed to be categorically excluded must strongly support this suggestion. It would follow that frequency measures and judgement data are *mathematically* related, since we could predict the score of a given item in a comparison set on the basis of the set’s scores from the other data type. They are not *practically* related, however, since the corpus size required would increase exponentially as we proceeded down the order of preferredness.

It follows from our arguments here that the data type of choice for syntax must be relative judgements. Frequency measures give us the same information as relative judgements about the best (couple of) structural alternatives in each comparison set, but they give us no information about any of the others. Since the interaction of linguistic constraints is demonstrably cumulative, this is a severe disadvantage, especially as it tends to make linguists interpret relative restrictions on structure as absolute restrictions. Put briefly: if you want to know *what* people say, choose frequencies, but if you want to know *why*, you are better off with relative judgements.

8 Implications for syntactic theory

These new, but empirically founded perspectives on data types, and their implications for the nature of grammaticality and the structure of the grammar are in some ways revolutionary, since they require a number of conventional assumptions to be abandoned or revised. Much can remain unchanged, however, since linguists in the past, on the basis of the much more partial data they had available, often nevertheless correctly identified characteristics of the data set. For example, with only individual’s (categorical) judgements and without the immediate access to corpuses that we have now, the abstraction to an essentially binary model of grammaticality was a reasonable step, which has in many ways served the field well. Our findings also raise the clear possibility that many linguistic constraints are indeed innate, to the extent that they related to the architecture of the mind and those structural variants that it finds easier and less easy to process.

On the other hand, our findings should make it clear to every syntactician that the current model of syntax has significant weaknesses. We can well un-

derstand how these came about, but that cannot be a reason not to move on. In fact the necessary reformulation of syntactic theory requires only two major steps. Syntax must also recognize that production processing has a role in deciding what linguistic forms are produced, and that occurrence only indirectly reflects well-formedness. This entails output selection and the grammar are two separate processes, and we must decide which of these two we are modelling. Three different positions are possible: we can look specifically at the grammar, that is the constraint application module, and disregard processing factors, in which case we should use data types which exclude the effects of occurrence as far as possible, ie relative judgements, and refine our theory to more accurately reflect the attested data patterns. Others will be more interested by the processing system: there is an extensive literature on sentence processing and numerous data-near models already. The third approach is to look at the cumulative effect on output of linguistic constraints and output selection. This is what many syntacticians are currently doing, but the mismatch in data pattern between frequencies and relative judgements reveals it to be looking at two heterogenous objects at the same time, and treating them as one. Now this is naturally an interesting and worthwhile field of study in its own right, one closely related to traditional descriptive linguistics, in which the occurring patterns of a language are the issue, rather than the underlying causes of these patterns. Frequency measures will be the natural data of choice for this study.

The differentiation of constraint application and output selection should bring about a major improvement in the empirical adequacy of syntax models, for the division of these two functions resolves at a stroke many of the inconsistencies which obscure the nature of the interaction of linguistic constraints. Syntactic theory will be far closer to the data, and hypotheses about the grammar will be far more constrained, surely a welcome development.

Having cleared the picture by factoring out the competitive effects of output selection, we can take a look at the module containing the grammar, which we have called Constraint Application. The second major step in revision of theory applies here, and consists of the specification of constraint violation costs. Each violation must have a quantified cost, since there are stronger and weaker violation constraints. The introduction of this parameter should alone will bring about many of the changes in architecture which are necessary to adjust current theory to gradient grammaticality, as is demonstrated to be necessary in work such as Keller (2000) and Featherston (2005). As soon as violation costs are accepted as a real variable, the other adjustments (survivability of constraint violations, cumulativity of violation costs, orthogonality

of violation cost strength and grammaticality/acceptability distinction) follow automatically.

These then are the lessons which we argue that syntax theory needs to draw from the closer inspection of its data base. First, we must redraw the boundary between grammar and processor so as to distinguish between the effects of linguistic constraints, and the effects of our production system. Second, we must add the additional parameter of violation cost to our models of syntax. Not words and rules, therefore, are the basic components of the grammar, but words, rules and sanctions.

Notes

This work was carried out within subproject A3 *Suboptimal Syntactic Structures* of the *SFB441 Linguistic Data Structures* supported by the Deutsche Forschungsgemeinschaft. Thanks are due to Wolfgang Sternefeld, project leader, Frank Keller for *WebExp* and many members of the SFB441. All errors are mine.

- 1 Contrary to the assumptions of magnitude estimation, there is no sign of magnitude scaling in 1, which would predict that the absolute differences in the upper half of the results set should be larger than those on the lower half, producing an upwardly curving distribution. Note that the upward curve at the extreme top end and the downward curve at the extreme bottom end of the results together make up only 3% of the data set and may be safely disregarded as outliers; while the upward curve might be taken as evidence of a magnitude scale, its downward equivalent conclusively demonstrates that it is no such thing.
- 2 Since we posit that the weightings assigned to the candidates are related to the computational workload they require, it is tempting to specify that the mechanism by which these weightings are taken into account by the selection process is temporal. That is, it is the quickest to compute and therefore the first arrival which is selected. This is of course pure speculation, however.

9 References

- Aissen J. & Bresnan J. 2002 Categoricality and variation in syntax: The Stochastic Generalization. Talk at Potsdam Gradience Conference, 22.2.2002
- Bard E., Robertson D. & Sorace A. 1996 Magnitude estimation of linguistic acceptability. *Language* 72 (1), 32-68
- Boersma P. & Hayes B. 2001 Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32, 45-86
- Bresnan J., Dingare & Manning C. 2001 Soft constraints mirror hard constraints: Voice and person in English and Lummit. In: Butt M. & King T.

- (eds) *Proceedings of LFG01 Conference*, 13-32. Stanford: CSLI
- Bybee J. & Hopper P. 2001 *Frequency and the Emergence of Linguistic Structure*. Amsterdam: Benjamins
- Featherston S. 2001 *Empty Categories in Sentence Processing*. Amsterdam: Benjamins
- Featherston S. 2002a Coreferential objects in German: Experimental evidence on reflexivity. *Linguistische Berichte* 192, 457-484
- Featherston S. 2002b Magnitude estimation and what it can do for your syntax. To appear in a special edition of *Lingua* on data in linguistics; editor Robert Borsley, Essex
- Featherston S. 2003a Bridge verbs and V2 verbs: The same thing in spades? Submitted to *Zeitschrift fr Sprachwissenschaft*
- Featherston S. 2003b That-trace in German. To appear in *Lingua*
- Featherston S. 2005 Universals and grammaticality: Wh-constraints in German and English. *Linguistics* 43, (4)
- Keller F. 2000 Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. PhD Thesis, Edinburgh
- Keller F., Lapata M. & Ourioupina O. 2002 Using the Web to Overcome Data Sparseness. In: Hajič J. & Matsumoto Y. (eds) *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 230-237. Philadelphia
- Labov W. 1996 When intuitions fail. In: McNair L., Singer K., Dolbrin L. & Aucon M. *Papers from the Parasession on Theory and Data in Linguistics. Chicago Linguistics Society* 32, 77-106
- Manning C. 2003 Probabilistic syntax. In: Bod R., Hay J., & Jannedy S. (eds), *Probabilistic Linguistics*, 289-341. Cambridge, MA: MIT Press
- Müller G. 1999 Optimality, Markedness, and Word Order in German. *Linguistics* 37, 777-818
- Poulton E.C. 1989 *Bias in Quantifying Judgments*. Hove & London: Erlbaum
- Prince A. & Smolensky P. 1993 *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report No.2, Center for Cognitive Science, Rutgers University
- Schütze C. 1996 *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press