The relationship between judgement data and frequency data in syntactic well-formedness: The Decathlon Model[*]

Sam Featherston

SFB441 Linguistic Data Structures, University of Tübingen, Germany

sam.featherston@uni-tuebingen.de

1 Introduction

Introspective judgements, frequency data, and processing studies all provide clues about the genesis, storage and use of linguistic constraints and structures. However the evidence from these separate data sources is often contradictory, which results in researchers who favour different data types having rather different ideas about the nature of linguistic structures and the mental architecture embedding them. For example, syntacticians making use of corpus data often favour probabilistic grammars (eg Manning, 2003), since this feature of language is very evident in this data type, while those relying on introspective data will tend to declarative grammatical models as a response to the intuition of absolute grammaticality (eg Pollard and Sag, 1994). A third group lays weight on processing evidence and finds support for the-grammar-is-the-parser approaches (eg Phillips, 1996). Exactly how these separate findings relate to each other is problematic and tends to lead syntacticians to inhabit different schools, depending in part on their chief data source. Needless to say, this division constitutes a major block on progress in syntax: a first step in describing a grammar must be agreement on what it means for a given structure to be part of the language.

2 The psycholinguistics of data types

The resolution of this complex of relationships is in essence a psycholinguistic task, which has however received little attention from psycholinguists, perhaps because it
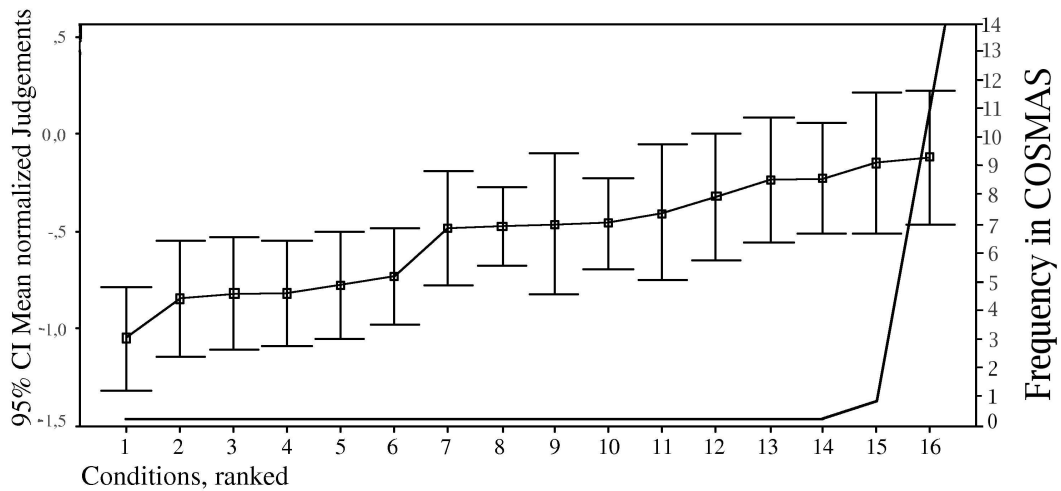
Figure 1: The contrasting data patterns of relative judgements and corpus frequency.

requires a wider, almost inter-disciplinary perspective in order to appreciate its importance. We have investigated this issue by looking at the same syntactic phenomena in two data types: experimentally obtained judgements and corpus frequencies. In a series of studies we have applied the technique of magnitude estimation of well-formedness (Bard et al., 1996; Keller, 2000), an experimental methodology which detects grammatical regularities with unsurpassed detail and reveals new insights into effects and interactions at the sentence level. We compare these results with corpus data (COSMAS, Institut für Deutsche Sprache, Mannheim) on the same constructions in order to determine the similarities and differences between the data types. This leads to insights into the relationship of the data types and additionally into the nature of the constructs of well-formedness underlying them.

We illustrate the different patterns obtained in figure 1. Of the sixteen different possible syntactic instantiations of the object coreference structure (see Featherston, 2000, for further details) only two appear in the 531 million word COSMAS corpus. The most frequent form appears 14 times, the next one only once, the rest not at all. Judgements of the same data set reveal a very different picture. The same two structures are "best", but there are differences between all the structures, right down to the very worst ones, furthermore these differences are systematic. It will be clear that the two data types are at least in part measuring different things.

We combine our insights into a partial model of the relationship of the varying data types – our Decathlon Model – and relate this to Levelt's (1989) well-known model of the sub-systems of production. Essentially this requires us to distinguish two syntax formulation modules: Constraint Application and Output Selection, which have very different roles and procedural characteristics. Constraint Application tests structures
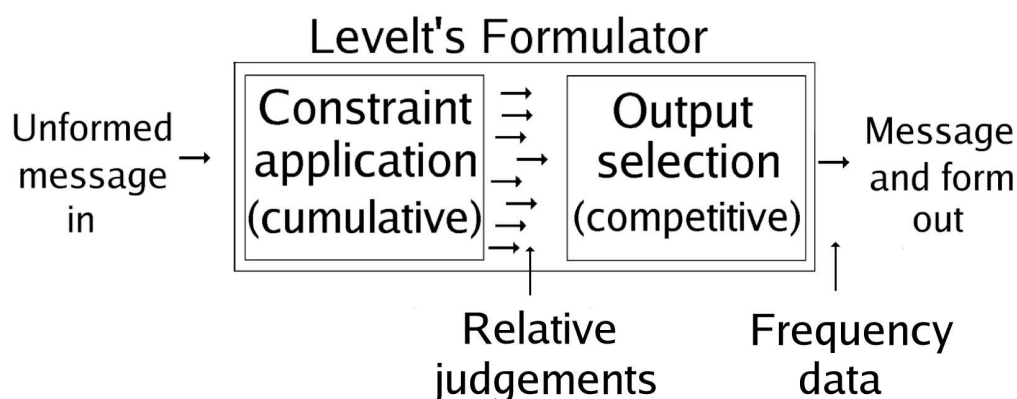
Figure 2: The Decathlon Model, located in Levelt's Formulator

for conformity with constraints and applies violation costs to violating structures. The testing is blind and exceptionless, the violation costs are cumulative: it has no probabilistic element nor any role for competition since it has no selection function. This task is carried out by the module Output Selection, which is competitive and probabilistic. These two stages show a striking similarity to the process of performance assessment in the athletic discipline decathlon: all competitors take part in all sub-disciplines but are essentially competing only against themselves or against the clock in each of these. Performance relative to other athletes comes only at the stage when the medal winners are to be determined, and takes place on the basis of the summed scores of each candidate in the full set of sub-disciplines. No athlete is excluded for poor performance in any sub-discipline: his or her chances of winning a medal may be reduced by coming last in the 100 metres, but there is no necessary connection. Indeed a candidate who performed third best in every sub-discipline would stand an excellent chance of winning the gold medal.

The association of syntactic forms to semantic contents happens in a very parallel manner. The Constraint Application module is where grammatical (and non-grammatical) constraints on syntactic forms are applied to candidate encodings of the message. Note that one must conceive of this process as functioning incrementally, but we shall abstract from this and many other similar details here. All candidate structures are subjected to all constraints, and all constraint violations are penalized with a violation cost of an amplitude specific to the constraint violated. Note that this directly contradicts the thesis of *violability* in OT (Optimality Theory; Prince and Smolenksy, 1993), which involves constraints having no effect upon the output (=failing to apply) in certain circumstances. In the Decathlon Model, all constraints always apply, but constraints are *survivable*, which means that there is no constraint whose violation necessarily prevents the violating structure from appearing in linguistic production. Within the model

this is trivial, since no differentiation of successful and unsuccessful candidates takes place in the Constraint Application module; all candidates are output with a cumulative well-formedness weighting. The Output Selection module chooses which form will be output, using the well-formedness ratings as its criterion. Just as three medals are awarded in the decathlon, the probabilistic mode of operation of the Output Selection module ensures that not only optimal candidates are produced, but that output frequency will closely reflect well-formedness.

The Decathlon Model provides insights into several vexed questions of data and theory. First, it gives an account of the variation in outcome of studies using different data types. Judgements (especially relative judgements, see below) tap into the output of the Constraint Application module, while frequency data additionally reflects the effects of the Ouptut Selection module. Second, it provides a psychologically realistic explanation of why even strongly suboptimal structures occasionally appear in the linguistic output, if a large enough sample is looked at. Third, it offers a solution to the problem why grammars with radically different architectures (eg OT based on competition vs HPSG based on declarative licensing) can both attain some degree of descriptive adequacy. Neither of them reflects all, but both capture some aspects of the nature of the grammar.

3 Relative judgements and categorical judgements

Analysis of our experimental results suggest that introspective judgements data are responsive to several factors, but that these can be at least partially teased apart. Subjects explicitly asked for relative judgements produce a true continuum of well-formedness, which contradicts the general assumption that there is, at least underlyingly, such a thing as categorical grammaticality. The undeniable intution that certain structures are, or are not, absolutely grammatical leads us to the distinction of two syntax-relevant well-formedness constructs within judgements: relative grammaticality and categorical grammaticality. Let us note here that both of these must in turn be distinguished from syntax-irrelevant factors such as felicity, a distinction which itself must be at least potentially operationalizable, if it is to be a meaningful term. We offer an approach to this problem which distinguishes between syntax-relevant and syntax-irrelevant effects by the exclusion of known irrelevant factors, but the inclusion of all others.

Judgements are affected, we argue, by (at least) two syntax-relevant factors, cumulative well-formedness (at base a reflex of computational complexity, we would argue) and frequency, the former playing a larger role in relative grammaticality and the latter producing the intuition of categorical (un)grammaticality. These different components may be identified by varying parameters of the judgement task: subjects instructed to give relative judgements respond to a greater degree to the cumulative degree of constraint satisfaction, those asked for categorical judgements reflect frequency to a greater extent. It follows that for syntacticians interested in identifying the formants

of grammaticality, that is, the individual rules which make up the grammar, relative judgements must be the data type of choice.

These findings have important implications for grammatical models, which, if confirmed in further work, could motivate major changes in syntactic theory. First, it appears that judgement data, but specifically relative judgement data is preferable to frequency data for research into the finer detail of grammatical systems, since it provides closer access to it, without the additional filter of output selection. Second, it suggests that many models of grammar would benefit from more attention to the psycholinguistic status of their assumptions. For example, declarative grammars need to consider what sanctions apply to structures containing unlicensed structures. Third, and most generally, much work in syntax should pay more attention to the evidential base on which it founds its arguments and should build in support for the reality of gradient grammaticality. We ourselves are developing a five-step abstraction from gradience which we hope will permit current syntactic practices to adapt to gradience with a few realistic practical methodological steps (Featherston, 2003).

References

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

Featherston, S. (2000). Coreferential objects in German: Experimental evidence on reflexivity. *Linguistische Berichte*, **192**:457–484.

Featherston, S. (2003). Judgements as the data base of syntax: Problems and some possible solutions. Talk given at Potsdam University 30.11.03. URL: `http://www.sfb441.uni-tuebingen.de/~sam/papers/Pots03.handout.pdf`.

Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Ph.D. thesis, University of Edinburgh.

Levelt, W. (1989). *Speaking: From Intention to Articulation*. MIT Press, Cambridge, Massachusetts.

Manning, C. (2003). Probabilistic syntax. In R. Bod, J. Hay, and S. Jannedy, eds., *Probabilistic Linguistics*, pp. 289–341. MIT Press, Cambridge, Massachusetts.

Phillips, C. (1996). *Order and Structure*. Ph.D. thesis, MIT.

Pollard, C. and I. Sag (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago, Chicago.

Prince, A. and P. Smolenksy (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report No.2, Center for Cognitive Science, Rutgers University.