Scaling Acceptability – Different Measures, Same Results

Keywords: Categorial vs. Gradient Judgements, Magnitude Estimation

The last ten years have seen a rapid growth of empirical investigation into linguistic acceptability. This development can be traced back to two influences: the methodological critique of the common practice of introspective judgements put forward by Schütze (1996) and Cowart (1997), but also to the advent of an experimental alternative to this practice, which became available by the extension of the magnitude estimation (ME) technique (Stevens, 1956) to linguistic acceptability judgments (Bard, Robertson & Sorace, 1996; Cowart, 1997). Papers which make use of this method deal with a variety of phenomena in different languages (among others: Keller & Alexopoulou, 2001; Keller, 2000, 2003; Featherston, 2005a, b; Sprouse, 2007).

In our talk, we want to stress the logical and methodological independence of these two developments. That is, we agree that formal linguistics is in need of careful empirical validation of its claims. But we want to raise the question whether ME data offer linguists the only adequate source for testing empirical hypotheses. The answer we propose to give to this latter question is in the negative. We want to argue for this on the basis of two claims: (1) ME does not contain more information than ordinal or categorial measures when conventional two-level factor designs are employed in an experiment. And (2): even in experiments with a gradient factor (e.g., a three-level markedness factor), ME does not give us more information, even if we look at measures of variance accounted-for (i.e., effect size as measured by partial η^2).

Both claims are backed up by experimental studies of acceptability in which we investigate several German word order constructions. Our first study shows that, in a conventional two-level factorial design (German SO vs. OS word order), ME judgments contain no more information than categorial or Likert-scale judgments. In our second study, we found that in a three-level factorial design, Likert-scale judgments show the same gradient decrease in acceptability as ME judgments do. In both studies, participants took part in the same experiments (consisting of several subdesigns) twice, once for each task. Order of task treatment was controlled for by reversing it for half of the groups, respectively. Frequencies of categorial judgments were square-root arcsine-transformed. ME judgments were treated as usually. All data were z-transformed for comparability reasons. These data were submitted to separate ANOVAs for the different subdesigns. We hypothesized that the factor TASK should have no main effect, and that it should not enter into interactions with the experimental factors of our subdesigns. In addition, we predicted that the variation that we can account for is not produced by the different measures (i.e, by the TASK factor), but by our experimental factors.

Both predictions are clearly borne out by our data. In both studies, the ANOVAs computed on the transformed data showed no effect of or interaction with the TASK factor, and all three measures showed comparable effect sizes as measured by partial η^2. The variation in our data hence was produced by our experimental factors, and not by the different tasks.

From this we conclude that all three measures we used contain the same amount of information in the kind of factorial designs we have studied. A fortiori, ME judgments do not contain more information than the other two measures. We hope that our results help to separate the desideratum of empirical validation of linguistic hypotheses from the question of choice of a particular method such as ME.

References

Bard, Ellen Gurman, Dan Robertson & Antonella Sorace (1996). Magnitude estimation of linguistic acceptability. *Language 72(1)*, 32–68.

Cowart, Wayne (1997). *Experimental Syntax.* Thousand Oaks: Sage Publications.

Featherston, Sam (2005a). Magnitude estimation and what it can do for your syntax: some wh-constraints in German. *Lingua 115(11)*, 1425–1440.

Featherston, Sam (2005b). That-trace in German. *Lingua 115(9)*, 1277–1302.

Keller, Frank (2000). *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality.* Ph. D. thesis, University of Edinburgh.

Keller, Frank (2003). A psychophysic law for linguistic judgments. *Proceedings of the 25th Annual Conference of the Cognitive Science Society.* Boston, 652-657.

Sprouse, Jon (2007). *Experimental Syntax: what does it get you?* Talk presented at the 20th Annual CUNY Conference on Human Sentence Processing. San Diego, CA, March 29-31.

Stevens, S.S. (1956). The direct estimation of sensory magnitude – loudness. *American Journal of Psychology 69*, 1–15.