# Different measures of linguistic acceptability – not so different after all?

Thomas Weskott and Gisbert Fanselow
University of Potsdam
weskott@uni-potsdam.de

Over the last ten years, empirical methods for the investigation of linguistic acceptability have developed rapidly, both in terms of the number of experimental studies, and in terms of the methodological rigour applied in those studies. This development can be attributed to 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, 1993; Cowart, 1997). In the meantime, a considerable amount of papers has been published in which magnitude estimation data supply the empirical evidence from which conclusions are drawn to a number of theoretical linguistic issues, ranging across a variety of phenomena in different languages (among others: Keller & Alexopoulou, 2001; Keller, 2000, 2003; Featherston, 2005a, b; Sprouse, 2007).

In this talk, we want to pursue the methodological question whether ME data offer linguists an adequate, and the only adequate, source for testing empirical hypotheses. We want to deal with this question from two perspectives, an empirical one, and a theoretical one.

I. The first part of the talk is devoted to presenting a series of experiments in which we compared three experimental methods for eliciting acceptability

judgments, (a) categorial judgments, (b) gradient judgments on a 7-point scale, and (c) magnitude estimations. We hypothesized that all three dependent variables should display the effects of our manipulations, and that the size of the effects in the three measures should be comparable, hence that the three methods are equally adequate.

For all three methods, we used an identical set of materials, and two groups of participants took part either in the categorial and the gradient judgment task (group 1), or in the categorical and the ME task (group2); order of treatment was controlled for by reversing it for half of the groups, respectively.

The experiment consisted of 4 factorial repeated measure subdesigns which tested (1) accusative scrambling (SO vs OS), (2) dative scrambling (SO vs OS), (3) number agreement in discontinuous NPs (number NP1 (sg vs. pl) x number NP2 (sg vs. pl)), and (4) superiority (order (Nom 1st vs. Nom 2nd) x speech act type (direct vs. indirect question)).

Frequencies of categorial judgments were square-root arcsine-transformed. ME judgments were divided by the value assigned to the modulus, and the result was log-transformed. These data were submitted to separate ANOVAs for the 4 designs for participants and items, respectively.

For all subexperiments, we found the same effects in all three types of dependent variables, i.e. the effect of e.g. accusative scrambling in Exp.1 was the same of the same size no matter whether it was calculated for frequencies of categorial judgements, judgments on a 7-point scale, or magnitude estimation values. Moreover, the size of the effects for the three dependent variables (as measured by partial eta square) was in the same range. This holds for all the other three subexperiments.

We take this result to indicate that all three dependent variables are equally

adequate for testing the kind of factorial designs we employed. From this, we go on to conclude that, at least for the constructions we have investigated, it does not matter whether a categorial variable (acceptable/inacceptable), an ordinal variable (7- point scale), or an interval-scaled variable (ME) is used to test two competing hypotheses.

II. Using these findings as a backdrop, the second part of the talk is devoted to the discussion of a few claims made in connection with different experimental measures of acceptability. Among these are the claim that categorial judgments are too coarse to reflect fine- grained linguistic intuitions (Bard et al, 1996; Featherston, 2005a; Keller, 2003). We will also discuss the claim that frequencies of categorial judgments are not amenable to parametric inferential statistic procedures (Bard et al., 1996) and show how this apparent problem can be overcome by applying the right kind of transformations to categorial data. Finally, we review recent findings from measurement theory and psychophysics (Narens, 1996; Ellermeier & Faulhammer, 2000) which are fit to cast some doubt on the assumptions that underlie magnitude estimation experiments.

A discussion of the various scaling and measurement approaches available for the empirical investigation of linguistic judgments and a plea for methodological pluralism conclude the presentation.

References:

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

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

Ellermeier, W. and G. Faulhammer(2000). Empirical evaluation of axioms of

fundamental to Stevens's ratio-scaling approach: I. Loudness production. Perception & Psychophysics, 62(8), 1505-1511.

Featherston, S. (2005a). Magnitude estimation and what it can do for your syntax: some wh-constraints in German. Lingua 115(11), 1525-1440

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

Keller, F. (2000). Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality. PhD Thesis, University of Edinburgh.

Keller, F. (2003). A psychophysic law for linguistic judgments. In: Alterman, R. & D. Kirsh(eds.). Proceedings of the 25th Annual Conference of the Cognitive Science Society. Boston, 652-657.

Keller, F. and T. Alexopoulou (2001). Phonology Competes with Syntax: Experimental Evidence for the Interaction of Word Order and Accent Placement in the Realization of Information Structure. Cognition 79(3), 301-372.

Narens, L. (1996). A theory of ratio magnitude estimation. Journal of Mathematical Psychology, 40, 109-129

Schütze, C.T. (1996). The empirical base of linguistics: Grammaticality judgments and linguistic methodology. Chicago: Chicago University Press.

Sprouse, J. (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, 2007.

Stevens, S.S. (1946). The direct estimation of sensory magnitude -- loudness. American Journal of Psychology, 69, 1-15.