How stable are your grammaticality judgments and exactly how does error affect them?

Stefan Grondelaers (Radboud University Nijmegen) & Dirk Speelman (University of Leuven)

Although grammaticality judgments have been widely criticized for their unreliability and instability (see Schütze 1996), few studies have seriously taken them to the task. In an attempt to evaluate the socio-syntactic use of intuitions, we report a rating experiment which was designed to replicate 3 confirmed determinants of the distribution of existential er “there” in Dutch constructions such as In de asbak lag (er) een hagelkorrel “In the ashtray there was a hailstone”. On the basis of “robust” corpus and laboratory evidence, we had previously demonstrated that er is an expectancy monitor inserted to facilitate the processing of unpredictable subjects such as hailstone. Crucially, er’s distribution could be fully predicted on the basis of 7 low-predictability factors (Grondelaers et al. 2009).

In order to test the validity of grammaticality judgments, we entered the 3 most robust low predictability factors in a rating experiment designed to neutralize or control all the known sources of error as best as possible. To check the stability of the ratings, an identical copy of the original questionnaire was administered to the same participants three weeks later. Regression analysis demonstrated that the ratings were valid – to the extent that er improved intuited grammaticality in the predicted direction – but also that all the manipulated effects were tainted by several sources of error.

In this paper we zoom in on the actual effect of three frequently cited but poorly investigated error sources. Separate regression analyses on the ratings of participants who were ignorant of the research purpose and participants who were not demonstrated that the latter did not tailor their intuitions in accordance with it, but displayed rule-based behaviour which significantly underspecified the complexity of er’s distribution. Another set of analyses indicated that ratings are highly unstable, both within the experiment as from moment to moment. Interestingly, the quality of ratings deteriorates as the same stimulus appears later in the questionnaire, but it improves when the same participants rate the same stimuli three weeks later. All in all, our data demonstrate that the use of introspective data should be subject to extreme design care and advanced prior knowledge of the investigated variable.

References