Lab or Armchair? The Benefits of Formal Acceptability Judgements

(Extended Abstract)

keywords: computational and experimental approaches, syntax, acceptability judgements, introspection, English

Jana Häussler[1] & Tom Juzek[2]

[1] University of Potsdam (jana.haeussler@uni-potsdam.de)
[2] University of Oxford (tom.juzek@googlemail.com)

Introduction For decades, informal ways of obtaining acceptability judgements dominated syntactic theory, with introspection being the most common. We use the term introspection to refer to acceptability judgements given by the author herself/himself and indicated by means of diacritics. Introspection in this sense belongs to the informal methods. These typically involve only few participants (in the extreme case only one, as with introspection) and often do not adhere to experimental standards, e.g. several lexicalizations per construction, randomisation, and distribution of items over lists in such a way that each participants sees each item in only one condition. Of course, acceptability judgements given by participants in an experiment are also introspective, but they are independent of the author. Furthermore, they adhere to common experimental standards.

There are several possible disadvantages of informal methods (cf. Schütze, 1996), among them are the lack of a common scale and an increased effect of judgement errors. As to the lack of a common scale: The number and types of diacritics that linguists use to express degrees of grammaticality vary across authors and papers. As a consequence, the meaning of any given diacritic can vary, even for authors applying the same number and the same types of diacritics. This scale bias can hamper comparisons across papers. As to judgement errors: Whenever a speaker, trained or naive, makes a judgement, this judgement is subject to performance-"noise" (memory limitations, distractors, priming effects, etc.). If the same sentence is rated on multiple times, it is likely that the ratings will slightly vary. This fluctuation is amplified in informal methods, due to the low number of subjects.

Although the field witnessed an increase of formal methods in the past 20 years, the predominant way of data collection is still informal. This conflict fueled the debate around the adequacy of introspection, and a quantitative comparison of formal and informal acceptability judgements was long missing. Sprouse et al. (2013) filled this gap and presented quantitative results that suggest that informal and formal judgements concur to a large extent. However, Sprouse et al. focused on pairwise comparisons, i.e. they compared marked constructions to their unmarked counterparts and checked *for each pair* whether informal and formal results agreed. Though pairs play an important role in syntactic theory, pairwise comparisons do not, in our view, reflect best how syntactic research is conducted. First, a majority of papers in our corpus (see below) involve more than two levels of acceptability, which indicates that they consider more than just pairs, as they are concerned with degrees of acceptability. An example for such gradience (from Kluender, 1992; cited in Hofmeister and Sag, 2010) is the following ordered quintuple in (1a-e).

(1a) This is the paper that we really need to find someone who understands.
(1b) This is the paper that we really need to find a linguist who understands.
(1c) This is the paper that we really need to find the linguist who understands.
(1d) This is the paper that we really need to find his advisor, who understands.
(1e) This is the paper that we really need to find John, who understands.

Kluender establishes the following ordering: (1a) ≥ (1b) ≥ (1c) ≥ (1d) ≥ (1e), where "≥" signifies "equally or more acceptable than". Though such an ordering can be established by pairwise comparisons of the type "x is better/worse than y", the underlying scale must be a continuum. (1a) to (1e) show gradience within a construction. A similar point can be made across constructions. For instance compare the baseline sentence in (2a) to the weak island violation in (2b) vs the strong island violation in (2c) (from Szabolcsi, 2006; her diacritics).

(2a) Which topic do you think that I talked about?
(2b) *?Which topic did John ask who was talking about?
(2c) *Which topic did you leave because Mary talked about?

A comparison of informal and formal methods should take these points into account and our aim is to provide exactly such a comparison. Consequently, we designed an experiment, in which we randomly sampled constructions from the literature and then compared them at large (i.e. beyond pairs).

LI Corpus To ensure a random sampling procedure, we created a corpus of informal acceptability judgements, based on all papers in *Linguistic Inquiry* (LI) from 2001 to 2010. This approach is similar to Sprouse et al. (2013). We created the corpus as follows: First, we excluded papers that were written by non-native speakers of American English (based on biographic information about the authors). Then, we extracted any kind of informal judgement concerning US-English. We did not include any items coming from experiments. In a final step, we categorised all extracted items, resulting in 2619 standard acceptability judgements, of which 2539 are testable in an experiment (non-testable items included ambiguities, use of offensive language, etc). Of the 2619 standard acceptability judgements, about a fifth (486 items) came from authors whose introspective judgements were *binary* (items marked with "*" vs not marked by any diacritic, in the following "OK"), 2133 items came from authors whose introspective judgements were gradient (by including some form of "?", e.g. "*?", "??", etc.). Note that we use "*" and "OK" only to refer to *introspective* judgements from LI papers.

Experiment From the 2539 testable data points, we randomly selected 2 × 100 sentences, one set from papers in which judgements were binary and another set from papers in which judgements were gradient. We decided to extract these two sets because we could not anticipate whether or not binary and gradient items behave differently. For both the "binary" and the "gradient" set, 50 items were *-sentences and 50 were OK-sentences (we focus on endpoints, since any result for endpoints will be most informative). Items were selected independently of potentially present counterparts. I.e. if a *-sentence got selected, we did not extract its OK-counterpart (if present) nor construct it (if absent); similarly for selected OK-sentences: We did neither extract nor construct their *-counterparts. So, rather than testing pairs of sentences, we tested single sentences.

We did so in an online acceptability judgement task, using a gradient scale (as the vast majority of items in our LI corpus was judged gradiently). The scale ranged from 1 ("fully unnatural") to 7 ("fully natural"). If scale bias and judgement errors were not an issue, one would expect OK-items and *-items to form distinct classes. In particular, OK-items and *-items from papers distinguishing more than these two

levels of acceptability should cluster at the endpoints of the 7-point scale used in the experiment. For items of medium acceptability, the corresponding authors would have used a diacritic expressing medium acceptability. In other words, given that the authors used more than two levels, one could expect experimental ratings to reflect this space between OK and *.

We used Amazon Mechanical Turk to recruit 80 subjects for two sessions (2 × 40 subjects). Exclusion criteria were: Returning incomplete results, being non-native speakers of American English, having extreme reaction times, or failing on "booby trap" items. We excluded 15 subjects, resulting in a total of 6500 data points (65 subjects × 100 sentences). For our analysis, we chose a point-biserial correlation measure, which suits our data structure best (the introspective judgements are binomial data, the z-scores of the online ratings are interval data; cf. Jackson, 2011).

Ratings and Results Fig. 1a shows the online ratings for the 50 *-items and 50 OK-items, taken from articles with binary introspective judgements. On the x-axis, the items are ordered by their online rating, which is given on the y-axis. Items *-marked in the corresponding LI article are given in red, items unmarked in the LI article are given in green. Fig. 1b shows the same for items from authors whose judgements were gradient. For illustrative reasons, results are given in their original 7-point ratings (the analyses are done with z-scores). As one would expect, the *-items tend to get low experimental ratings (means: 3.1 for *-items from papers with binary judgements and 3.4 for *-items from papers with gradient judgements) and the OK-items tend to get high ratings (means: 5.6 and 5.2). However, we also see a substantial amount of overlap. Ratings for *-items from the binary set range from 1.5 to 6.2 , OK-items from the same set range from 3.1 to 7.0 ; for the gradient set the ranges overlap as well: they range from 1.3 to 5.3 for *-items (mean 3.4) and from 2.6 to 6.9 for OK-items . Moreover, for gradiently judged items, the ratings do not cluster at the two endpoints, contrary to what one might expect. As a consequence, the point-biserial correlations are far from being perfect: For items from authors with binary judgements, the coefficient is 0.746; for items from authors with gradient judgements, it is 0.655.

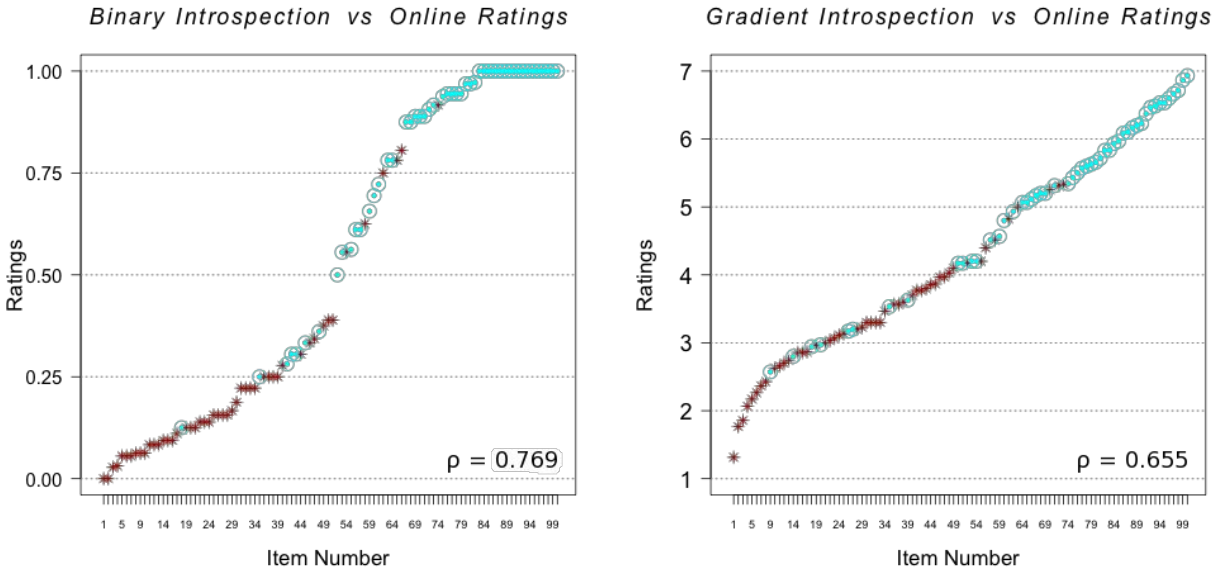

Fig. 1a (left) shows the ratings for items that come from LI articles with binary judgements. Fig. 1b (right) shows the ratings for items that come from articles with gradient judgements. Items (x-axis) are in an ascending order by their corresponding online rating (y-axis), including both *-items (dark red asterisks) and OK-items (green circles).

Discussion and Conclusion The rather moderate correlation coefficients, even for the gradient set, suggest that there is a non-trivial mismatch between introspective judgements and experimental results. In principle, both the LI authors and the participants will have made "good" judgements. However, both informal and formal judgements will be affected by scale biases (i.e. interspeaker differences in the application of the scale) and judgement errors. But the impact on the formal results is reduced thanks to averaging over a high number of participants and normalizing the individual ratings (i.e. using z-scores). For introspection, though, scale biases and judgement errors are more pronounced due to a low N. The consequences are potentially damaging, specifically for syntactic theory building. Increasing N, i.e. resorting to experimental methods, is an effective way to reduce the effects of scale biases and judgement errors. On a theoretical level, this line of argumentation has already been laid out by Schütze (1996). The present study provides quantitative support for Schütze's arguments and, in our view, shows that formal methods, and acceptability judgement tasks in particularly, have their place in syntactic theory.

REFERENCES

- Hofmeister, P., Sag, I. (2010). Cognitive constraints on syntactic islands. *Language*, 86, 366-415.
- Jackson, S. (2011). *Research Methods and Statistics: A Critical Thinking Approach*. Wadsworth: Cengage.
- Kluender, R., (1992). Deriving islands constraints from principles of predication. In Helen Goodluck and Michael Rochemont (eds.), *Island constraints: Theory, acquisition and processing*, 223-58. Dordrecht: Kluwer.
- Schütze, C. T. (1996). *The empirical base of linguistics*. Chicago: Chicago University Press.
- Szabolcsi, A. 2006. Strong and weak islands. In Martin Everaert and Henk van Riemsdijk (eds.), *The Blackwell Companion to Syntax*, Vol. IV, 479-532. Oxford: Blackwell.
- Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua*, 134, 219-248.