Is there variation in the reliability of theoretical syntax data across languages?

Two data-gathering methods are commonly used in syntax: (i) non-quantitative processes lacking formal data collection protocols or statistical analysis; and (ii) experimental (formal) methods, as used in other domains of the cognitive sciences. While the traditional approach is found throughout, the second ("experimental syntax") is becoming more widespread. As part of this shift, classical theoretical data have been questioned (e.g., [1], [2], [3], [4]) and the first large-scale comparisons of the non-quantitative and experimental data collection methods, a.k.a., data assessment, have been carried out. [5] found a 93%, and [6], 98%, convergence rate between traditional judgments and data gathered experimentally for a 10-year period of research in English published in a top journal and an English syntax textbook, respectively. Questions arise concerning data assessment in other languages. Specifically, languages other than English have not been tested as systematically, which is doubly problematic. First, typological variation might affect the results. This is the case because replication rates varying according to the phenomena under scrutiny. Second, the number of syntacticians working on languages other than English is smaller. Therefore, there could be comparatively more significant noise in the data and the corresponding theoretical developments ([7]). Recent work by [8] focused on Hebrew and Japanese but instead of using a random sample, they focused on what they considered to be problematic judgments found in the literature. Thus, their study, while pertinent, cannot be compared to the work of Sprouse and colleagues. In turn, the relevance of the question in (b.) has been pointed by [8], a.o., e.g., under the assumption that data sets have increasingly become more subtle ([9]) or that quotes can be considered (everything else being equal) a kind of replication ([8]). To my knowledge, this question has not been addressed yet. **Methodology:** To address these issues, a grammaticality judgment task using a 7-point Likert scale including a random sample from Spanish syntax papers published in *Probus* in the period 2006-2017 was developed to establish to what extent recent data gathered informally by syntacticians is representative of Spanish in general. Methodological details were selected so as to be able to compare the current work with [5]'s data assessment project for English, thus addressing the question of whether research on languages other than English faces bigger reliability issues. A second experiment was conducted to study whether number of citations has an effect on the reliability of the data (see [8] and [5] for discussion). To address this second issue, a second experiment was developed using data selected randomly from the most influential work on Spanish syntax, where influential was defined as 100+ quotes. Naturally, these publications are relatively old, the publication year ranging from 1971 to 2009 (mean year: 1990). The dialect was held constant, specifically, Venezuelan Spanish (examples where adapted to control for lexical and/or morphological variation). The analysis of the directionality of the responses based on the difference between means for each minimal pair in paired sample t-test challenge the idea that the data in earlier high impact generative grammar was more reliable or representative than in recent work, at least for random samples. Most importantly for present purposes, the results of both experiments reveal that Spanish data is less reliable than the English data studied by Sprouse et al., irrespective of the number of citations or age of the data, highlighting the need for large-scale data assessment project for Spanish and for other languages. In particular, the Probus sample yielded an 82.352% convergence rate, or 85.294% if marginal results in the predicted direction are counted a convergent, and the high impact sample yielded a 79.069% convergence rate, or 83.720% if, again, marginal results are considered convergent, compared to 93% for English ([5]). A discussion of the potential causes for the Spanish results, as well as the controversial data points is included.

References

[1] Edelman, S., & M. Christiansen, 2003. How seriously should we take Minimalist syntax? *Trends in Cognitive Sciences* 7: 60-61.

[2] Ferreira, F., 2005. Psycholinguistics, formal grammars, and cognitive science. *Linguistic Review* 22 365--380.

[3] Gibson, E., & E. Fedorenko. 2010a. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14.233–234.

[4] Gibson, E., & E. Fedorenko. 2010b. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28:1-2, 88-124.

[6] Sprouse, J., & Almeida, D. 2012. *Assessing the reliability of textbook data in syntax: Adger's Core Syntax. Journal of Linguistics* 48(3), 609-652.

[5] Sprouse, J., C. Schütze & D. Almeida. 2013. Assessing the reliability of journal data in syntax: Linguistic Inquiry 2001-2010. *Lingua* 134, 219-248.

[7] Haider, H. 2016. *Incredible syntax – between Cognitive Science and imposture*. Book ms.

[8] Linzen, T. & Y. Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa* 3(1): 100. 1–25.

[9] Gervain, J. 2003. Syntactic microvariation and methodology: Problems and perspectives. *Acta Linguistica Hungarica*, 50: 405-434.