Update

Letters

Weak quantitative standards in linguistics research

Edward Gibson[1] and Evelina Fedorenko[2]

[1] Department of Brain and Cognitive Sciences, Department of Linguistics and Philosophy, Massachusetts Institute of Technology, 46-3035, MIT, Cambridge, MA 02139, USA
[2] Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research, Massachusetts Institute of Technology, 46-4141C, MIT, Cambridge, MA 02139, USA

A serious methodological weakness affecting much research in syntax and semantics within the field of linguistics is that the data presented as evidence are often not quantitative in nature. In particular, the prevalent method in these fields involves evaluating a single sentence/meaning pair, typically an acceptability judgment performed by just the author of the paper, possibly supplemented by an informal poll of colleagues. Although acceptability judgments are a good dependent measure of linguistic complexity (results from acceptability–judgment experiments are highly systematic across speakers and correlate with other dependent measures, but see Ref. [1]), using the researcher's own judgment on a single item/pair of items as data sources does not support effective testing of scientific hypotheses for two critical reasons. First, as several researchers have noted [2–4], a difference observed between two sentences could be as a result of lexical properties of the materials rather than syntactic or semantic properties [5,6]. Multiple instances of the relevant construction are needed to evaluate whether an observed effect generalizes across different sets of lexical items [7]. The focus of this letter, however, is on a second problem with standard linguistic methodology: because of cognitive biases on the part of the researcher, the judgments of the researcher and his/her colleagues cannot be trusted (Box 1) [8,9]. As a consequence of these problems, multiple items and multiple *naïve* experimental participants should be evaluated in testing research questions in syntax/semantics, which therefore require the use of quantitative analysis methods.

The lack of validity of the standard linguistic methodology has led to many cases in the literature where questionable judgments have led to incorrect generalizations and unsound theorizing, especially in examples involving multiple clauses, where the judgments can be more subtle and possibly more susceptible to cognitive biases. As one example, a well-cited observation in the syntax literature is that an object–subject–verb question asking about three elements is more natural than one asking about only two (e.g. *What did who buy where?* is claimed to sound better than *What did who buy?*). Several theories explain and build on this purported phenomenon [10,11]. However, it turns out that the empirical claim is not supported by quantitative measurements [12,13]. There are many other such examples of questionable judgments leading to unsound theorizing {[2–4]; including an example from the first author's PhD thesis (Gibson, E., 1991, PhD Thesis, Carnegie Mellon University), which is discussed along with other examples elsewhere (Gibson, E. and Fedorenko E., *The Need for Quantitative Methods in Syntax*, unpublished)}. Without quantitative data from naïve participants, cognitive biases affect *all* researchers.

Furthermore, even if such cases were rare, the fact that this methodology is not valid has the unwelcome consequence that researchers with higher methodological standards will often ignore the current theories from the field of linguistics. This has the undesired effect that researchers in closely related fields are unaware of interesting hypotheses in syntax and semantics research.

To address this methodological weakness, future syntax/semantics research should apply quantitative standards from cognitive science, whenever feasible. Of course, gathering quantitative data does not guarantee the lack of confounding influences or the correct interpretation of data. However, adopting existing standards for data gathering and analyses will minimize reliance on data that are spurious, driven by cognitive biases on the part of the researchers.

Corpus-based methods provide one way to quantitatively test hypotheses about syntactic and semantic tendencies in language production. A second approach involves controlled experiments. Experimental evaluations of syntactic and semantic hypotheses should be conducted with participants who are naïve to the hypotheses and samples large enough to make use of inferential statistics. A variety of experimental materials should be

Box 1. Cognitive biases and linguistic judgments.

There are at least three types of unconscious cognitive biases [8,9] that can adversely affect the results of intuitive judgments, given the way that they are currently typically gathered in the syntax/semantics literature:

1. Confirmation bias on the part of the researcher: researchers will often have a bias favoring the success of the predicted result, with the consequence that they will tend to treat data that do not support the hypothesis as flawed in some way (e.g. from a not quite native speaker, or from a speaker of a different dialect).
2. Confirmation bias on the part of the participants: individuals that the researcher asks to provide a judgment on a linguistic example – including the researcher him/herself – might be biased because they understand the hypotheses. When faced with complex materials, they could then use these hypotheses to arrive at the judgment.
3. Observer–expectancy effects (the ''clever Hans'' effect): individuals that the researcher asks to provide a judgment could be biased because they subconsciously want to please the researcher and are consequently affected by the researcher's subtle positive/negative reactions.

Corresponding author: Gibson, E. (egibson@mit.edu).

used to rule out effects as a result of irrelevant properties of the experimental items (e.g. particular lexical items). It is our hope that strengthening methodological standards in the fields of syntax and semantics will bring these fields closer to related fields, such as cognitive science, cognitive neuroscience and computational linguistics.

References

1 Edelman, S. and Christianson, M. (2003) How seriously should we take Minimalist syntax? *Trends Cogn. Sci.* 7, 60–61
2 Schütze, C. (1996) *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*, University of Chicago Press
3 Cowart, W. (1997) *Experimental Syntax: Applying Objective Methods to Sentence Judgments*, Sage Publications
4 Wasow, T. and Arnold, J. (2005) Intuitions in linguistic argumentation. *Lingua* 115, 1481–1496
5 MacDonald, M.C. *et al.* (1994) The lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101, 676–703
6 Gibson, E. and Pearlmutter, N. (1998) Constraints on sentence comprehension. *Trends Cogn. Sci.* 2, 262–268
7 Clark, H.H. (1973) The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *J. Verbal Learn. Verbal Behav.* 12, 335–359
8 Evans, J.S. *et al.* (1983) On the conflict between logic and belief in syllogistic reasoning. *Mem. Cogn.* 11, 295–306
9 Nickerson, R.S. (1998) Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175–220
10 Kayne, R. (1983) Connectedness. *Linguist. Inq.* 14, 223–249
11 Pesetsky, D. (2000) *Phrasal Movement and its Kin*, MIT Press
12 Clifton, C., Jr *et al.* (2006) Amnestying superiority violations: processing multiple questions. *Linguist. Inq.* 37, 51–68
13 Fedorenko, E. and Gibson, E. Adding a third *wh*-phrase does not increase the acceptability of object-initial multiple-*wh*-questions. *Syntax* (in press), doi:10.1111/j.1467-9612.2010.00138.x

1364-6613/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tics.2010.03.005 Trends in Cognitive Sciences 14 (2010) 233–234

Letters Response

Quantitative methods alone are not enough: Response to Gibson and Fedorenko

Peter W. Culicover[1] and Ray Jackendoff[2]

[1] Department of Linguistics, The Ohio State University, 222 Oxley Hall, 1712 Neil Ave., Columbus, OH 43210-1298, USA
[2] Center for Cognitive Studies, Tufts University, Medford, MA 02155, USA

Gibson and Fedorenko [1] (see also [2,3]) correctly point out that subjective judgments of grammaticality are vulnerable to investigator bias, and that – where feasible – other types of data should be sought that shed light on a linguistic analysis. Major theoretical points often rest on assertions of delicate judgments that prove not to be uniform among speakers or that are biased by the writer's theoretical predispositions or overexposure to too many examples.

Another problem with grammaticality judgments is that linguists frequently do not construct enough control examples to sort out the factors involved in ambiguity or ungrammaticality. But this problem cannot be ameliorated by quantitative methods: experimental and corpus research can also suffer from lack of appropriate controls (see Box 1).

Nevertheless, theoreticians' subjective judgments are essential in formulating linguistic theories. It would cripple linguistic investigation if it were required that all judgments of ambiguity and grammaticality be subject to statistically rigorous experiments on naive subjects, especially when investigating languages whose speakers are hard to access. And corpus and experimental data are not inherently superior to subjective judgments.

In fact, subjective judgments are often sufficient for theory development. The great psychologist William James offered few experimental results [4]. Well-known visual demonstrations such as the Necker cube, the duck-rabbit, the Kanizsa triangle, Escher's anomalous drawings, and Julesz's random-dot stereograms are quick and dirty experiments that produce robust intuitions [5]. These phenomena do not occur in nature, so corpus searches shed no light on

Box 1. The need for proper controls in Gibson and Fedorenko's experiment

Fedorenko and Gibson's argument turns on the claim that superiority violations with two wh-phrases are supposedly worse than with three. Their experiment [9] disputes this judgment. The relevant sentence types are illustrated in (i).

(i) Peter was trying to remember ...
a. who carried what.
b. who carried what when.
c. what who carried.
d. what who carried when.

They find that, in contrast to longstanding judgments in the literature, (ic) is worse than (id), the two are judged to have about equal (un)acceptability.

They do not control by replacing the third wh-phrase with a full phrase as in (ii).

(ii) Peter was trying to remember ...
a. who carried what last week.
b. what who carried last week.

We find (iia) as good as (ia,b), but (iib) worse than (ic,d). If so, *some* violations with two wh-phrases *are* worse than counterparts with three. The difference calls for a reexamination of the examples in the literature, controlling for this factor. Ratings studies might be helpful in establishing the reliability of these judgments. We doubt relevant examples will be found in corpora of natural speech and writing. And we also doubt that Bolinger's original observation in [10] resulted from investigator bias.

Corresponding author: Jackendoff, R. (ray.jackendoff@tufts.edu).