Corroborating corpus data with elicited introspection data: A case study

Jakob Horsch[1]
[1]Catholic University of Eichstätt-Ingolstadt, Jakob.Horsch@ku.de

Keywords: negative data problem, corroborating corpus data, magnitude estimation

The exponential growth in corpus size over the last three decades has led to a proliferation of corpus studies. However, these come with limitations: Corpora are *finite* samples of language (Hoffmann 2019: 17), whereas language is by definition *infinite* (Chomsky 1965: 6). This leads to the **negative data problem** (“just because a phenomenon cannot be found in a corpus, it cannot be concluded that it is ungrammatical” (Hoffmann 2011: 1)) and the **positive data problem** (“just because a construction appears in a corpus it does not automatically follow that it is grammatical” (Hoffmann 2011: 1)). Therefore, corpora can never be considered fully representative of a language.

One solution is corroborating corpus data with elicited introspection data. I present a case study to show how this can be achieved using the Magnitude Estimation Test (MET) method (Bard et al. 1996; Cowart 1997: 73-84, Hoffmann 2011, Hoffmann 2013), which takes advantage of the fact that humans are better at making relative judgments than absolute judgments (e.g. Likert scales). METs also feature grammatical and ungrammatical fillers, providing ‘baselines’ against which test items can then be compared.

Specifically, I investigate the acceptability of optional *that*-complementizers in the English Comparative Correlative (CC) construction ([*The more we get together*]$_{\text{C1}}$, [*the happier we’ll be*]$_{\text{C2}}$). The CC consists of two subclauses, C1 and C2; *that*-complementizers are optional in C1 (1) and have been claimed to be possible in C2 (2) in “colloquial registers” (den Dikken 2005: 402; see also Hoffmann 2019: 47) (2):

(1) [*The more* [***that***]$_{\textbf{optional THAT-complementizer}}$ *he says*,]$_{\text{C1}}$ [*the less I wanna say*.]$_{\text{C2}}$
(2) [*the larger the settlement becomes*]$_{\text{C1}}$
[*the less* [***that***]$_{\textbf{optional THAT-complementizer (?)}}$ *the reduced number of sites you will have available.*]$_{\text{C2}}$

However, the CC construction is extremely infrequent; Hoffmann et al. indicate a per-million-word frequency of 30-40 (2019: 32). The same applies to *that*-complementizers; in their 2,041-token BNC data set of comparative correlatives, Hoffmann et al. (2020) found just 29 C1 *that*-complementizers and two in C2, concluding that they are “no longer central properties” (2020: 200) of the Present-day English (PdE) CC construction. Any claims based on such sparse evidence are therefore subject to the negative/positive data problems.

To address this issue, I collected grammaticality judgments from 37 L1 American English speakers and normalized them as *z*-scores. The data were tested for significance using mixed-effects models. Fig. 1, which shows the *z*-score means obtained for C1, indicates that the presence of *that*-complementizers in C1 did not change participants’ ratings significantly (this was confirmed by mixed-effects modeling). This suggests that they can indeed be considered grammatical (and optional). Fig. 2, however, shows that *that*-complementizers in C2 (*that*) were rated significantly worse than the alternative, i.e. no *that*-complementizer (Ø). This appears to be in line with claims about them being restricted to colloquial speech (although as pointed out by an anonymous reviewer the underlying reasons – i.e., register – have not been conclusively determined). Thus, it was possible to corroborate findings from corpus studies and address the negative/positive data problems, proving that generally the MET method is indeed viable for complementing corpus data.

Furthermore, the study has led to new findings: Despite being rated worse than the alternative (Ø), *that*-complementizers in C2 must be considered grammatical, since their *z*-score mean is much closer to the grammatical filler mean than the ungrammatical filler mean. This has implications for analyses of the English CC as a hypotactic construction: If *that*-complementizers were indeed markers of subordination (cf. e.g. Borsley (2004) and den Dikken (2005)), they should have been clearly rated as ungrammatical in C2, which under a hypotactic analysis functions as main clause.

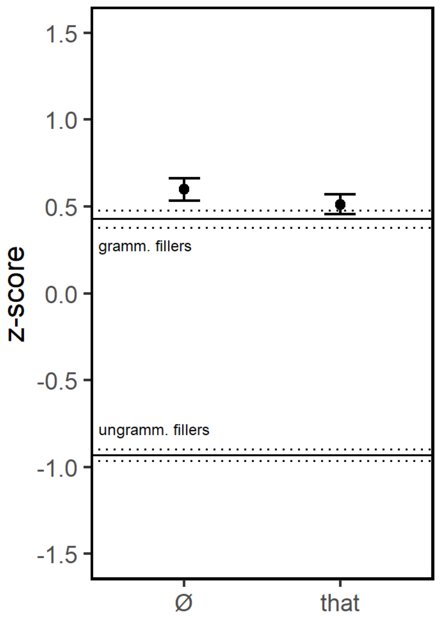

Fig. 1: Z-scores of C1 that-complementizers (n=37)

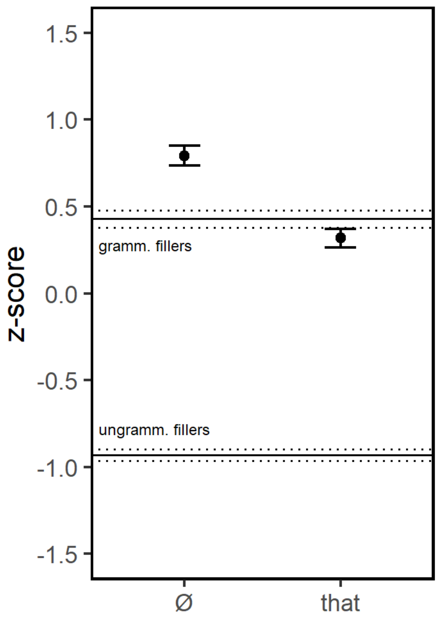

Fig. 2: Z-scores of C2 that-complementizers (n=37)

References

Bard, Ellen G., Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language 72*(1). 32–68.

Borsley, Robert D. 2004. An Approach to English Comparative Correlatives. In Stefan Müller (ed.), *Proceedings of the 11th International Conference on Head-Driven Phrase Structure Grammar*, *Center for Computational Linguistics, Katholieke Universiteit Leuven*, 70–92. Stanford, CA: CSLI Publications.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Cowart, Wayne. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgements*. Thousand Oaks, CA: Sage.

Dikken, Marcel den. 2005. Comparative Correlatives Comparatively. *Linguistic Inquiry 36*(4). 497–532.

Hoffmann, Thomas. 2011. *Preposition Placement in English: A Usage-based Approach* (Studies in English Language). Cambridge: Cambridge UP.

Hoffmann, Thomas. 2013. Obtaining introspective acceptability judgements. In Manfred Krug & Julia Schlüter (eds.), *Research Methods in Language Variation and Change*, 99–118. Cambridge: Cambridge UP.

Hoffmann, Thomas. 2019. *English Comparative Correlatives: Diachronic and Synchronic Variation at the Lexicon-Syntax Interface* (Studies in English Language). Cambridge: Cambridge UP.

Hoffmann, Thomas, Thomas Brunner & Jakob Horsch. 2020. English Comparative Correlative Constructions: A Usage-based account. *Open Linguistics 6*(1). 196–215.

Hoffmann, Thomas, Jakob Horsch & Thomas Brunner. 2019. The More Data, The Better: A Usage-based Account of the English Comparative Correlative Construction. *Cognitive Linguistics 30*(1). 1–36.