Proceedings of the Society for Computation in Linguistics

Volume 2 Article 39

2019

On the Interaction Between Dependency Frequency and Semantic Fit in Sentence Processing

Soo Hyun Ryu
University at Buffalo, soohyun422@gmail.com

Rui P. Chaves
University at Buffalo, rchaves@buffalo.edu

Recommended Citation
Ryu, Soo Hyun and Chaves, Rui P. (2019) "On the Interaction Between Dependency Frequency and Semantic Fit in Sentence Processing," *Proceedings of the Society for Computation in Linguistics*: Vol. 2 , Article 39.
DOI: https://doi.org/10.7275/neag-b065
Available at: https://scholarworks.umass.edu/scil/vol2/iss1/39

On the interaction between dependency frequency and semantic fit in sentence processing

Soo Hyun Ryu
Department of Linguistics
University at Buffalo
soohyunr@buffalo.edu

Rui P. Chaves
Department of Linguistics
University at Buffalo
rchaves@buffalo.edu

1 Introduction

The frequency of linguistic input is a key factor in predicting sentence processing difficulty, which motivated probabilistic models of sentence processing (Hale, 2001; Levy, 2008). Another important factor is plausibility, *i.e.* semantic fit, which indicates how consistent the linguistic input is with world knowledge. Up until recently, plausibility was regarded as a by-product of frequency. For example, Padó et al. (2009) define plausibility as the joint probability of lexical, semantic and syntactic information, and for Levy (2008) any representation that affects reading difficulty does it in proportion to predictability. However, Kang et al. (2018) recently provide experimental evidence suggesting that predictability and plausibility are distinct types of knowledge; see also DeLong et al. (2014) for ERP data consistent with this conclusion.

Kang et al. (2018) measured verb-object predictability with a 10-rank cloze task and measured plausibility with a verb-object plausibility judgment task. By using a cloze task, Kang et al. effectively avoid sparseness problems caused by estimating predictability from corpora. On the other hand, a 10-rank cloze task is a rather unnatural task for comprehenders, and potentially creates noisy results. Suppose, for example, that participants choose a prototype *'a gift'* as the object of *'The girl sent her boyfriend ___'*. In that case, producing nine additional cloze alternatives is difficult and likely produces artificial results.

In other words, both the corpus-driven approach to dependency frequency estimation and the elicitation approach have shortcomings. In the present work, we investigate whether the sparseness problem can be minimized by using very large syntactic corpora, and provide independent corroborating evidence for Kang et al.'s (2018) claim that predictability alone cannot explain semantic fit.

2 Method

Through Amazon.com's Mechanical Turk (AMT) crowdsourcing marketplace, we recruited 41 self-reported native speakers of English with IP addresses originating from the United States. Participants read sentences on a self-paced moving window display (Just et al., 1982). Programming and the presentation of the experimental stimuli were done using Ibex 0.3.9 (Drummond, 2013).[1] Four lists of sentences were created and pseudo-randomized with 24 distractors, half of which were followed by comprehension questions as illustrated in (1).

(1) Someone $_1$| is $_2$| printing $_3$| many $_4$| copies $_5$| of $_6$| the memo. $_7$|
Q: This person likely knows how to operate a xerox machine. [True/False]

The experimental items were constructed using the verb-object patienthood norms obtained in Experiment 2A of McRae (2010) and the corresponding verb-object dependency bigram frequencies from the 345 billion words Google Syntactic N-gram corpus (Goldberg and Orwant, 2013). Ratings of verb-object patienthood greater than 4 or higher were considered 'High Fit (H^{fit})' ($M = 5.67$, $SD = 0.70$) and ratings less than 4 were considered 'Low Fit (L^{fit})' ($M = 2.35$, $SD = 0.72$). Conversely, verb-object dependency frequencies greater than 10 were considered high (H^{freq}) ($M = 204.90$, $SD = 253.00$) while dependencies with 0 frequency were considered low (L^{freq}). It was not

[1] See Futrell (2012, 25–35) for validation studies showing that Ibex self-paced reading experiments with AMT participants can replicate classic self-paced reading experiments run in the laboratory.

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 330-333.
New York City, New York, January 3-6, 2019

Sentence	Condition	Fit	Freq.
We $_1$| intended $_2$| to $_3$| select $_4$| the contractor $_5$| after $_6$| the workshop. $_7$|	$H^{fit}H^{freq}$	5.0	10
We $_1$| intended $_2$| to $_3$| select $_4$| the contestant $_5$| after $_6$| the workshop. $_7$|	$H^{fit}L^{freq}$	5.2	0
We $_1$| intended $_2$| to $_3$| select $_4$| the secretary $_5$| after $_6$| the workshop. $_7$|	$L^{fit}H^{freq}$	3.9	52
We $_1$| intended $_2$| to $_3$| select $_4$| the gourmet $_5$| after $_6$| the workshop. $_7$|	$L^{fit}L^{freq}$	3.8	0

Table 1: Example *stimuli*

possible to find verb-object combinations in the McRae (2010) that had frequencies of exactly the same magnitude in Syntactic N-gram corpus, and so the H^{freq} condition values varied substantially, which reflects the problem of estimating frequencies from corpora.

12 sets of experimental items in a 2×2 Latin Square design were constructed, pitting (i) high/low semantic fit between the verb and its direct object and (ii) high/low verb-object frequency, as illustrated in Table 1.[2] All item quadruples had similar syntactic structure, but different content words. To minimize the influence of the subject phrase on the semantic fit of the direct object, only pronouns were used as subjects, which reduced the risk of the subject affecting the processing time of the direct object.[3]

The data from three participants were removed from the analysis for having comprehension question accuracy below 75%, and data points were excluded if the reading time for the corresponding region was less than 100 ms or more than 1500 ms. After residual reading times were computed for each participant according to region length, LMER models were fit to compare conditions at each region. In all models, the intercept was allowed to be adjusted by items, subjects, and lists in order to account for random effects.

3 Results

Residual reading times by regions are shown in Figure 1. The $H^{fit}L^{freq}$ items were read slower than $H^{fit}H^{freq}$ items at spill-over Region 6 (β = 41.50, t = 2.72, $p < 0.01$). Similarly, $L^{fit}H^{freq}$ items were read slower than $H^{fit}H^{freq}$ items (β = 29.52, t = 2.00, $p < 0.05$) at Region 6. No difference was found in any other regions, before or after the region of interest.

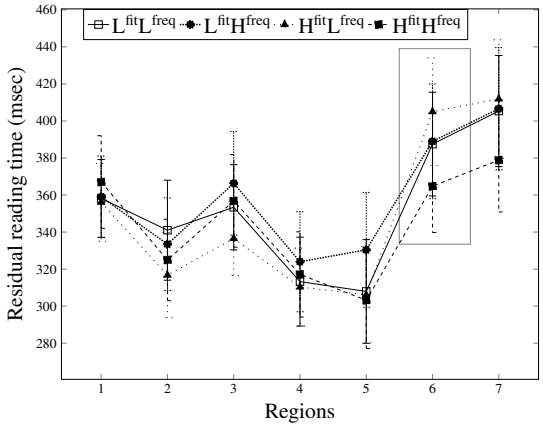

Figure 1: Residual reading time by regions

Subsequent investigation of Region 6 revealed an interaction between semantic fit and frequency (β = -45.59, t = -2.09, $p < 0.05$), as shown in Figure 2. In a deviation coding scheme, both frequency (β = -18.82, t = -1.73, p = 0.84) and semantic fit (β = -5.92, t = -0.54, p = 0.5) were shown to have no main effects. In a treatment coding scheme, a simple effect of frequency (β = 41.61, t = 2.68, $p < 0.01$) and a near-significant simple effect of semantic fit (β = 28.71, t = 1.89, p = 0.06) were found.

4 Discussion

We found no main effects of frequency or semantic fit. Rather, we found an interaction between the two in which the effect of verb-object frequency on reading time is strong only in the presence of high verb-object thematic fit, and the effect of verb-object semantic fit occurs only when the frequency is high. This interaction between frequency and plausibility is problematic for sen-

[2]The full data can be downloaded at github.com/soohyun422/SprExp/blob/master/Datasheet.xlsx.

[3]In Kang et al. (2018), items like *The girl sent her boyfriend flowers* were rated as low verb-object plausibility presumably not because of the verb-object fit, strictly speaking but rather because of the fact that the the agent and patient combination is non-prototypical for such an event. We avoid this potential confound by using pronouns.

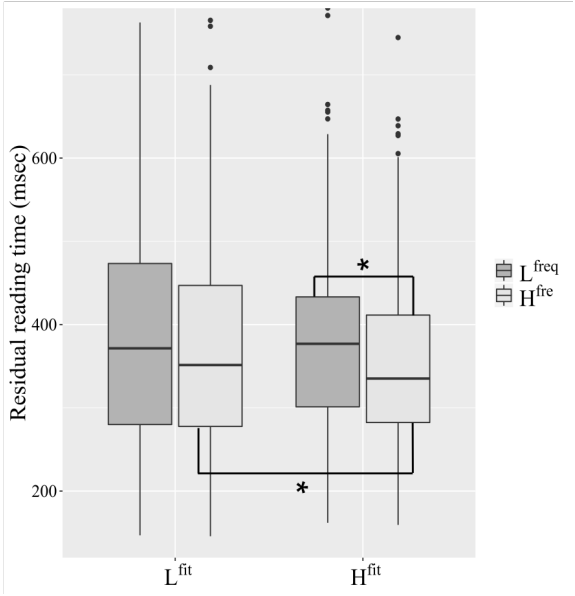

Figure 2: RT by conditions at Region 6

tence processing models that reduce plausibility to frequency (Levy, 2008; Padó et al., 2009).

Assuming that sparseness was not the cause, the null effect of semantic fit, given low frequency, may be a consequence of expectations being primarily driven by high frequency.[4] If the predicted object does not match the object in the sentence, the semantic fit of the latter offers little processing advantage. Particularly so if the observed verb-object combination has a low frequency.

The null effect of frequency in the low semantic fit condition may be due to an inhibition effect. As is well-known, comprehenders produce expectations about the upcoming direct object during on-line sentence processing, based on their world knowledge (Kutas and Hillyard, 1984; Altmann and Kamide, 1999; Arai and Keller, 2013). It is therefore possible that the direct object that comprehenders pick inhibits other highly frequent candidates, and therefore the high frequency of those competitors has little effect on reading times when the object in the sentence does not match the expected candidate.

Before concluding, we should point out the possibility that verb micro-senses may have contributed to the lack of verb-object frequency effect in the low sematic fit condition. The classic example of micro-sense is *cut*, which involves radically different motor programs depending on the object, such as *cut the grass / cake / meat / wood* (Elman, 2009). Even though some of our target sentences involved the same verb sense, they may have invoked different micro-senses. One such example pair is (...) *accused the criminal* (...) which was $H^{fit}H^{freq}$, and (...) *accused the prosecutor* (...) which was $H^{fit}L^{freq}$. Although both sentences have basically the same verb sense, the former is more likely to be interpreted as a technical or specialized legal term and the latter as an accusation with no legal consequences. It is possible that the micro-senses of *accuse* led to different semantic expectations about the object, potentially leading to higher reading times when the expectation was violated. Further research is needed to determine whether fine-grained verb-object semantic fit expectations had an effect on the experiment.

5 Conclusion

The purpose of the present study was to measure the effects of frequency and semantic fit in sentence processing using large syntactic corpora, in spite of the sparseness problem. The results of this study suggest that the effect of semantic fit and frequency in sentence processing are distinct, even using corpora, and interact in complex ways. Our results provide independent support for the cloze-based findings in (Kang et al., 2018), by suggesting that verb-object semantic fit is not a by-product of verb-object probability.

[4] Indeed, Kang et al. (2018) found that the object predictability arises faster than the effect of plausibility (the former appearing at the object region and the latter appearing at the spill-over region.

References

Gerry T. M. Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73:247–264.

Manabu Arai and Frank Keller. 2013. The use of verb-specific information for prediction in sentence processing. *Language and Cognitive Processes*, 4(28):525–560.

Katherine A DeLong, Laura Quante, and Marta Kutas. 2014. Predictability, plausibility, and two late erp positivities during written sentence comprehension. *Neuropsychologia*, 61:150–162.

Alex Drummond. 2013. Ibex 0.3.7 manual. [http://spellout.net/latest_ibex_manual.pdf].

Jeffrey L. Elman. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33:1–36.

Richard Futrell. 2012. *Processing Effects of the Expectation of Informativity*. Ma thesis, Stanford.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 241–247.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Marcel Adam Just, Patricia A. Carpenter, and Jacqueline D. Wooley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2):228–238.

Hong Mo Kang, J. Koenig, and G. Mauner. 2018. Plausibility is not reducible to predictability. 31st Annual CUNY Sentence Processing Conference, Davis, CA.

Marta Kutas and Steven A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 5947(307):161–163.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

K. McRae. 2010. Thematic fit ratings from a number of studies [data file]. Retrieved from World Development Indicators, https://sites.google.com/site/kenmcraelab/norms-data.

Ulrike Padó, Matthew W Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):794–838.