Generalizing dependency distance: Comment on "Dependency distance: a new perspective on syntactic patterns in natural languages" by Haitao Liu et al.

Richard Futrell, Roger Levy, and Edward Gibson

June 16, 2017

Abstract

With the support of the comprehensive review in Liu et al. (2017), we consider dependency distance minimization to be firmly established as a quantitative property of syntactic trees. In this comment, we consider future empirical and theoretical directions for this concept, including a recent information-theoretic reinterpretation of dependency locality effects as proposed by Futrell and Levy (2017).

Multiple independent lines of work from psycholinguistics, corpus linguistics, linguistic typology, and computational linguistics are converging around a simple quantitative generalization about syntactic trees in natural language: that the linear distance between words linked in dependencies is usually short. Liu et al. (2017, in this issue) provide a thorough and multifaceted review of this concept, which they term dependency distance minimization (DDM). In turn, DDM provides a rich interface between linguistics, graph theory, and complex systems theory.

Based on the work reviewed in Liu et al. (2017), we believe that DDM is now firmly established as a first-order empirical generalization about syntactic trees. This commentary will focus on possible theoretical and empirical directions extending the DDM idea to explain more phenomena in languages and human language processing.

1 Generalizing dependency locality

DDM is motivated by theories of human language processing difficulty based on memory constraints, but as Liu et al. (2017) note, these theories only

account for a subset of observed online processing difficulty. In fact, better predictive accuracy for online reaction times is achieved by an alternative theory, **surprisal theory**, which is based on probabilistic expectations rather than memory constraints. Surprisal theory holds that the online processing effort for a word in context is directly proportional to the information content (log inverse probability) of the word in context (Hale, 2001; Levy, 2008a; Smith and Levy, 2013). There is a long history of evidence for surprisal effects in reaction times for naturalistic text, whereas such evidence for dependency locality effects has only been found recently, in texts that were edited to be difficult to understand (Shain et al., 2016). Nevertheless, while surprisal theory has good coverage, it cannot account for the well-attested dependency locality effects which motivate DDM as a typological principle.

In recent work, Futrell and Levy (2017) introduce a unification of expectation-based and memory-based theories of online processing difficulty, which derives and generalizes dependency locality effects. This theory, called **noisy-context surprisal**, repurposes a noisy-channel model of sentence comprehension (Levy, 2008b; Gibson et al., 2013) to predict online processing difficulty by combining it with surprisal. In this combined theory, the processing effort for current input is proportional to the information content of that input given a *noisy representation* of that context (Levy, 2011), rather than a veridical representation of the preceding context, as was the case in previous surprisal models. Futrell and Levy (2017) introduce the key additional hypothesis that the level of noise affecting the context representation is not uniform: rather, more distal context has higher noise levels than more proximal context. This increasing noise rate is motivated by the Data Processing Inequality: holding an element of context in memory requires continued processing of its representation, and the noise affecting a representation increases monotonically the more data processing is done on it (Cover and Thomas, 2006).

Dependency locality effects emerge from a noisy-context surprisal model in a generalized form: processing difficulty in this model occurs when word pairs with high mutual information (i.e., word pairs which predict each other) are distant in linear order. Under mild assumptions, the predicted processing effort for a word w_i in context $w_1, ..., w_{i-1}$ is approximately:

$$C(w_i|w_1, ..., w_{i-1}) \approx \log \frac{1}{p(w_i)} - \sum_{j=1}^{i-1} f(i-j)\text{pmi}(w_i; w_j), \tag{1}$$

where $\text{pmi}(w_i; w_j)$ is the pointwise mutual information of w_i and w_j and

$f(d)$ is some monotonically decreasing probability mass function indicating the probability that a word of distance d remains unaffected by noise in the context representation. The generalization expressed by Equation 1 is **information locality**: for ease of processing, words that depend on each other statistically should be close to each other.

We can see dependency distance as an approximate metric of information locality if we assume further that syntactic dependencies, as they are annotated in treebanks, identify those word pairs that have high mutual information. It is natural that words in syntactic dependencies would have high mutual information because mutual information is simply a quantification of generic dependence in a statistical sense. If we assume that the correct probability model for sentences has dependents generated conditional on their heads (Eisner, 1996; Klein and Manning, 2004), then heads and dependents are exactly those word pairs with the highest mutual information (Futrell, 2017). The assumption that syntactic dependencies indicate high mutual information is also ubiquitous in NLP, and empirical evidence for this point is given in Futrell and Levy (2017).

While dependency locality denotes processing difficulty when words in syntactic dependencies are distant, information locality describes processing difficulty when a word is distant from *any* relevant contextual information that it must be integrated with. In empirical support of information locality, Futrell and Levy (2017) show that words with high mutual information are usually close, and Gildea and Jaeger (2015) show that word orders are optimized so that mutually predictive words appear within a 3-word window of each other. Li (1989) and Lin and Tegmark (2016) show a power-law decay of mutual information with distance for letters in natural language text. Information locality follows from a generalization of surprisal theory, which is a broad-coverage model of human sentence processing difficulty, so we believe it is promising as a principle of language processing and as a pressure affecting word order.

2 Prospects for information locality

While it remains to be shown that information locality can fully explain dependency locality effects, we believe information locality has the potential to explain detailed word order patterns. The primary prediction beyond DDM is that the strength of the DDM-based attraction between words should be modulated by their mutual information. This modulation might explain why, for instance, adjuncts are typically farther from their heads than ar-

guments, if it is the case that adjuncts have lower mutual information with their heads. Adjective order preferences might also be amenable to this kind of analysis.

Noisy-context surprisal could also permit us to incorporate in a principled way the factors known to modulate dependency locality effects. Noisy-context surprisal holds that the memory representation of context is affected by some noise, but the exact form of the noise function is unspecified: the derivation of information locality assumes only that the noise rate increases the longer a word representation has been in memory. This noise function could be specified to build in effects of primacy, recency, givenness, and intervening material, all of which are noted in Liu et al. (2017) to modulate dependency locality effects. The corpus studies reported can also inform the form of the noise model: the generally power-law decay of dependency distance suggests that the survival probability function f in Equation 1 should have the form of a power law.

Even if dependency locality is a special case of information locality, we expect dependency distance will remain an important metric because of its simplicity. Estimating mutual information from linguistic observations is difficult because of the long tail of possible wordforms, whereas crosslinguistic dependency treebanks are becoming more and more common because of projects such as Universal Dependencies (Nivre, 2015), and one needs only a modest sample size of dependency trees to demonstrate dependency distance minimization. DDM can also be easily formulated as a graph theoretic problem, making it easier to reason about.

The field of quantitative syntax is still just beginning, and DDM is one of its first strong quantitative generalizations, with a solid theoretical motivation coming from psycholinguistic models of incremental language processing. DDM joins the other known factors affecting quantitative word order patterns, such as animacy, definiteness, givenness, and frequency of words. Like DDM, these other factors all appear to be motivated by the principle of minimum effort, in that they involve placing easier and more accessible items earlier in a sequence (an "easy-first" preference in production). Thus, we believe that DDM, joining these other quantitative generalizations, will play a pivotal role in an eventual explanation of the form of human language in terms of optimal communication systems under human-like information processing constraints.

Acknowledgments

This work was supported by NSF grant number 1551543 to R.F. and E.G. and NSF grant number 1534318 to E.G.

References

Cover, T. and Thomas, J. (2006). *Elements of information theory*. John Wiley & Sons, Hoboken, NJ.

Eisner, J. M. (1996). Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 340–345.

Futrell, R. (2017). *Memory and locality in natural language*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Futrell, R. and Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain.

Gibson, E., Bergen, L., and Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

Gildea, D. and Jaeger, T. F. (2015). Human languages order information efficiently. *arXiv*, abs/1510.02823.

Hale, J. (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 and Language Technologies*, pages 1–8.

Klein, D. and Manning, C. D. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, page 478.

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

Levy, R. (2008b). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 234–243.

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *ACL*, pages 1055–1065.

Li, W. (1989). Mutual information functions of natural language texts. Technical report, Santa Fe Institute Working Paper #1989-10-008.

Lin, H. W. and Tegmark, M. (2016). Critical behavior from deep dynamics: A hidden dimension in natural language. *arXiv*, abs/1606.06737.

Liu, H., Xu, C., and Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews [in this issue]*.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., and Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 49–58, Osaka, Japan.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.