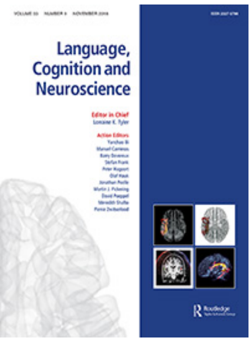

Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

ISSN: 2327-3798 (Print) 2327-3801 (Online) Journal homepage: https://www.tandfonline.com/loi/plcp21

Hierarchical and sequential processing of language

A response to: Ding, Melloni, Tian, and Poeppel (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience. *Language, Cognition and Neuroscience*.

Stefan L. Frank & Morten H. Christiansen

To cite this article: Stefan L. Frank & Morten H. Christiansen (2018) Hierarchical and sequential processing of language, Language, Cognition and Neuroscience, 33:9, 1213-1218, DOI: 10.1080/23273798.2018.1424347

To link to this article: https://doi.org/10.1080/23273798.2018.1424347

Published online: 25 Jan 2018.

LANGUAGE, COGNITION AND NEUROSCIENCE
2018, VOL. 33, NO. 9, 1213–1218
https://doi.org/10.1080/23273798.2018.1424347

RESPONSE

OPEN ACCESS

Hierarchical and sequential processing of language
A response to: Ding, Melloni, Tian, and Poeppel (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience. *Language, Cognition and Neuroscience*.

Stefan L. Frank[a] and Morten H. Christiansen[b]

[a]Centre for Language Studies, Radboud University, Nijmegen, The Netherlands; [b]Department of Psychology, Cornell University, Ithaca, NY, USA

ABSTRACT
Ding et al. (2017) contrast their view that language processing is based on hierarchical syntactic structures, to a view that relies on word-level input statistics. In this response to their paper, we clarify how, exactly, the two views differ (and how they do not), and make a case for the importance of sequential, as opposed to hierarchical, structure for language processing.

ARTICLE HISTORY
Received 28 July 2017
Accepted 2 January 2018

KEYWORDS
Hierarchical structure; sequential structure; language statistics; abstraction and generalisation

Introduction

In their recent paper, Ding, Melloni, Tian, and Poeppel (2017; from here on: DMTP) contrast two opposing views on language processing; one that claims “comprehension is driven by rule based decomposition of hierarchical syntactic structure” (p. 570) and the alternative which argues that comprehension is based on “word-level input statistics” (p. 573). The first view, which DMTP adhere to, relies on linguistic levels of abstraction such as syntactic categories (noun, verb, etc.), phrases (noun phrase, verb phrase, etc.) and hierarchical syntactic structures that follow a rule-based grammar. The opposing view, which they ascribe to us (Frank, Bod, & Christiansen, 2012) is alleged to “reflect[s] the online analysis of the statistical relationship between adjacent words which obviates the need for abstract structure building” (p. 570).

In this response to the DMTP paper, we aim to clear up several (apparent) misunderstandings regarding the controversy sketched above, as well as to explain what, exactly, our claims about language processing are (and are not). Further, we will show that purported evidence for hierarchical processing is not always what it seems. Specifically, DMTP present MEG power spectrum data as evidence for hierarchical linguistic structure but the same data can also be generated by a system that uses nothing but sequences of word representations.

The controversy: what is (not) at stake?

The dichotomy between rule-based hierarchical processing on the one hand and statistical word-level processing on the other presupposes that rules only apply to (or create) hierarchical syntactic structures whereas statistical analysis remains limited to word sequences. However (as DMTP acknowledge) statistics can also apply to parts of a syntactic structure and is therefore not restricted to the level of word sequences. Conversely, non-statistical rule-based processing can be applied to word-level input, as for example in the “Preliminary Phrase Packager” of Fodor and Frazier’s (1978) Sausage Machine model and for certain aspects of f-structure in Lexical Functional Grammar. The fact that the presence of hierarchical structure is independent from the use of statistical information demonstrates that DMTP’s carving up of the views on language processing conflates two orthogonal dimensions: whether processing is *statistical*, and whether it is *hierarchical*.

The role of statistics in language processing

According to DMTP, language statistics at the word level are “in many cases, not necessary to explain human language processing performance” (p. 573). It is of course true that many phenomena cannot be reduced to word-level statistics. However, effects of word-level

CONTACT Stefan L. Frank s.frank@let.ru.nl statistics are (nearly) unavoidable: Word frequencies are for a large part responsible for word recognition times (Gardner, Rothkopf, Lapan, & Lafferty, 1987; among many others) and word reading times closely follow word probability conditional on the sentence context (Smith & Levy, 2013). Similar word-probability effects have been observed in neural activity during sentence or text comprehension (Brennan, Stabler, Van Wagenen, Luh, & Hale, 2016; Frank, Otten, Galli, & Vigliocco, 2015; Nelson et al., 2017; Willems, Frank, Nijhof, Hagoort, & Van den Bosch, 2016 – for review, see Armeni, Willems, & Frank, 2017). Such probabilistic (i.e. statistical) effects form the backdrop against which (perhaps more interesting) processing phenomena take place. And moving beyond the word level, too, there is overwhelming behavioural evidence for effects of the statistics of multiword utterances (Arnon & Christiansen, 2017; Arnon & Snider, 2010), syntactic categories (e.g. Monsalve, Frank, & Vigliocco, 2012), syntactic constructions (e.g. Fine, Jaeger, Farmer, & Qian, 2013; Frank, Trompenaars, & Vasishth, 2016; MacDonald & Christiansen, 2002), and even pragmatic contexts (Goodman & Stuhlmüller, 2013).

This is not to say that people are unable to learn and apply grammar rules in a non-statistical manner, as they often do, for example, when using a second language that has been learned by explicit instruction. Much of (native) language acquisition and processing, however, relies on the language user's knowledge of language statistics. In fact, without assuming a fundamental role for statistics, it is hard to explain the efficiency of comprehension or its robustness against errors and ambiguity. Thus, the real controversy is not about the importance of statistics but about the role of hierarchical structure in language use.

Hierarchical versus sequential processing

Contrary to what DMTP appear to believe, non-hierarchical processing is not restricted to the word level or to *n*-grams. We will therefore refer to the superficial structure of the linguistic signal as *sequential*, highlighting that any spoken or written linguistic utterance takes the form of a sequence of units; and only its *analysis* may (or may not) be hierarchically structured. Sequential structure is, by definition, not hierarchical, but it is present beyond the word level because other linguistic units also display sequential structure. For example, a spoken utterance may be perceived as a sequence of phonemes, and a sentence can also be viewed as a sequence of multiword chunks or even syntactic categories. Moreover, the notion of sequential structure goes beyond *n*-grams in at least two respects. First, there is no fixed maximum string length (*n*) beyond which statistical dependencies are excluded (which does not imply that dependencies can be reliably tracked over an unlimited distance). Second, sequences need not be comprised of words (or other arbitrary symbols) but can instead be modelled using high-dimensional numerical vectors that are learned statistically from unannotated corpora. The use of such analogical representations is well known to improve an *n*-gram model's generalisation ability (Bengio, Ducharme, Vincent, & Jauvin, 2003).

As DMTP point out, hierarchical structures are more abstract than word strings in the sense that they are further removed from the input signal. We agree with DMTP when they argue for the importance of abstraction and its crucial role in generalisation. In fact, we would go even further and point out that abstraction begins to take place long before reaching the word-level "input": phonemes are an abstraction over the raw sound signal and, likewise, graphemes are an abstraction over the visual input in reading. Words themselves form an abstraction too, which is how we can generalise over input modalities and different pronunciation (or spelling) variants. It is therefore uncontroversial that abstraction takes place. The question is: which level of abstraction needs to be assumed to account for particular instances of human language performance? Are syntactic categories (abstracting over words) psychologically real? And does hierarchical syntactic structure building, which requires a very high level of abstraction, always occur during comprehension?

DMTP's position on this issue is quite clear, as the very first sentence of Ding, Melloni, Zhang, Tian, and Poeppel (2016) claims that "To understand connected speech, listeners *must* construct a hierarchy of linguistic structures […] including […] phrases" (p. 158, emphasis added). Hence, these authors take it as a fact that abstraction to the level of phrase structure is required for comprehension. This is precisely where we disagree: According to our view, abstraction from the input can remain limited to what is required for comprehension, which depends on both the nature of the utterance (e.g. frequent multiword strings can be analysed as a whole, so no internal structure is assigned, even if the string also retains its sequential character) and the comprehension setting (e.g. engaging in informal conversation affords more superficial analysis than criticising philosophical discourse). To what extent a particular instance of utterance processing involves hierarchical structure building (and if so: to which level) is an empirical question (Sanford & Sturt, 2002; see Christiansen & Chater, 2016a, 2016b, for discussion) to which we return in the Conclusion.

This principle of "minimal abstraction" extends to accounts of specific language performance phenomena:

One should not assume unnecessary levels of abstraction; and even if higher levels of abstraction are required for comprehension, these may not be responsible for the phenomenon under investigation. For example, nobody believes that *n*-gram statistics suffice for comprehension; nevertheless, particular phenomena may be sufficiently explained by *n*-grams in which case higher levels of abstraction (such as hierarchical structure) should not be posited in models of such phenomena (e.g. Reali & Christiansen, 2005). One specific phenomenon where this may be the case is in the relation between the size of the N400 ERP component in response to reading a word and the word's conditional probability. When these probabilities are estimated by an *n*-gram model they can predict the ERP size better than probabilities that follow from a (particular) hierarchical phrase-structure grammar (Frank et al., 2015), suggesting that the generator of the N400 is not particularly sensitive to highly abstract linguistic knowledge, at least not insofar as it follows word probabilities.

If the presence of hierarchical syntactic structure is not a priori assumed, one must remain sceptical about claims that some phenomenon is caused by hierarchical processing. If a sequential explanation suffices this should be preferred because of the lower level of abstraction involved. As a case in point, we shall revisit one piece of evidence from Ding et al. (2016), reproduced in DMTP. In what follows, we will first briefly discuss the experiment that gave rise to the evidence. Next, we discuss the computational model by Frank and Yang (in press) that assumes no higher level of abstraction than a syntactic/semantic clustering of word representations; in particular, the model has no knowledge about (probabilities of) word sequences or syntactic structures. Nevertheless, we will see that this model precisely predicts the results that DMTP present as evidence in support of hierarchical linguistic processing.

Re-examining the evidence

The Ding et al. study

Participant in the Ding et al. (2016) study listened to a continuous stream of Chinese or English syllables, presented at a fixed rate of 250 ms per syllable (or slightly longer in the English case; a difference we will ignore). In the condition presented by DMTP, each subsequence of four Chinese syllables formed a sentence consisting of a two-syllable noun followed by a two-syllable verb or verb phrase. In the equivalent English condition, each sentence consisted of four monosyllabic words, the first pair forming a noun phrase and the second pair a verb phrase, as in [[dry fur]$_{NP}$ [rubs skin]$_{VP}$]$_S$. The MEG signal recorded during speech perception was subjected to a Fourier analysis to obtain a power spectrum, which showed peaks at exactly the presentation frequencies of the three levels of linguistic structure: words (or Chinese syllables) at 4 Hz, phrases (or Chinese words) at 2 Hz, and sentences at 1 Hz. This was taken as evidence for cortical entrainment to hierarchically related levels of linguistic structure.

The Frank and Yang model

Frank and Yang (in press) developed a model to simulate the Ding *et al*. experiments with the goal to investigate if the results could be explained without recourse to syntactic processing. We will only discuss the English model here.

The model represents each word as a high-dimensional numerical vector, automatically extracted from a large English text corpus by a state-of-the-art neural network model for distributional semantics (Mikolov, Chen, Corrado, & Dean, 2013). Twelve different participants are simulated by obtaining different word vector representations, with different dimensionalities.

The entire stimulus sequence of 60 four-word sentences is represented by simply concatenating the word vectors into a matrix, which then has one row for each vector dimension. Each matrix column corresponds to a simulated MEG sample, with a sampling rate of 200 Hz (i.e. sample duration of 5 ms). The (simulated) word presentation rate is 4 Hz, so there are 50 matrix columns per word token (for details, see Frank & Yang, in press). Next, a Discrete Fourier Transform is applied to each matrix row and the results are averaged over rows to obtain a power spectrum.

Crucially, the only linguistic information available to the model resides at the lexical level. The word representations are not integrated in any way; there is no phrase- or sentence-level processing, only representation of lexical information. Hence, if power spectra predicted by the model are qualitatively similar to those from Ding et al.'s (2016) MEG experiment, these results can be explained from the stimulus' lexical properties alone and may therefore not indicate syntactic processes after all.

Comparing the model to the Ding et al. results

Figure 1 shows the power spectrum on English four-word sentences from the Ding et al. (2016) study and model output on the same sentences. The model predicts the same peaks at the presentation rates of words, phrases, and sentences. In addition, there is a small but significant peak at 3 Hz, which is also visible

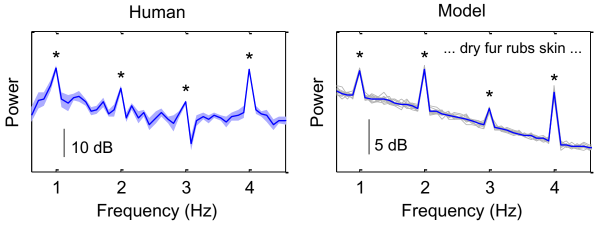

Figure 1. Power spectrum for English four-word sentences from Ding *et al*. (left; with adapted frequency scale) and corresponding model predictions (right). The shaded area (left) is the standard error over participants; and grey lines (right) are results from individual simulated participants. Blue lines represent averages over (simulated) participants. Asterisks indicate frequencies with statistically significant power peaks relative to neighbouring frequency bins, after multiple comparison correction. Adapted from Frank and Yang (in press, Figure 1).

in the human data and reaches significance in Frank and Yang's (in press) reanalysis of this data but not in Ding et al.'s (2016) original analysis. As the stimulus has no property that occurs at this rate, the 3 Hz peak is most likely just the second subharmonic of the 1 Hz peak (Zhou, Melloni, Poeppel, & Ding, 2016). The 2 Hz peak is not merely the first subharmonic, as is clear from the fact that it does not disappear when only the noun phrases or only the verb phrases are presented to the model (Figure 2, top). In those conditions, the stimuli have no sentence structure but only word and phrase structure; and the model predicts no 1 Hz peak in the power spectrum but only the 2 and 4 Hz peaks. Ding *et al*.'s MEG results show the same pattern in the corresponding Chinese conditions.

When word order is scrambled, there is no linguistic structure beyond the word and indeed the resulting power spectrum shows only the 4 Hz peak (Figure 2, bottom left). Finally, for sentences that have the original NP-VP structure but are not semantically interpretable ("Jabberwocky" sentences, taken from Martin & Doumas, 2017), the model predicts the three peaks corresponding to the three levels of linguistic structure (Figure 2, bottom right).

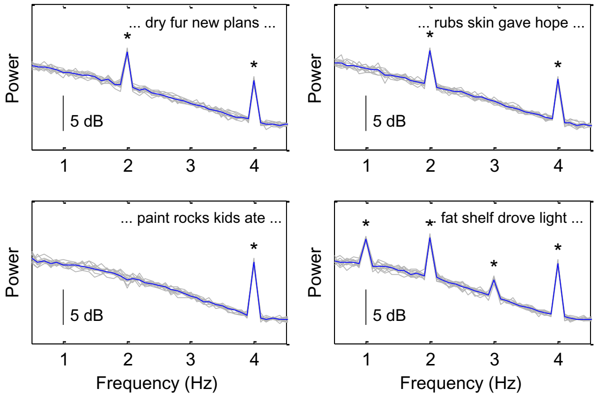

Figure 2. Model predictions for English stimuli not tested by Ding *et al*. Top left: noun phrases only. Top right: verb phrases only. Bottom left: word salad. Bottom right: Martin and Doumas's (2017) "Jabberwocky" sentences.

All these model results correspond to what one would expect if peaks in the power spectrum indicate syntactic processing at different hierarchical levels (see Frank & Yang, in press, for additional results on Chinese stimuli). However, the model has no hierarchical syntax. Consequently, these power spectra can also result from merely representing lexical information in a temporal sequence, which is all the model does. Nevertheless, Ding et al. (2016) are correct when they point out that their results cannot be explained by mere "word-level input statistics". The model abstracts away from the word level, without engaging syntactic structure, by assigning each word a unique vector (itself based on word-level co-occurrence statistics) that incorporates information relevant to the word's syntactic/semantic category.

Conclusion

Language is rife with statistical structure that the cognitive system can use for learning how to comprehend and produce utterances. This applies to all levels of linguistic analysis, from phonetics to syntax to pragmatics. Thus, we do not disagree that there exists a hierarchy in levels of analysis and presumably an (approximately) corresponding hierarchy in the cognitive representation of different aspects of linguistic utterances from sound to discourse (Christiansen & Chater, 2016b). The question we pose is: When, and to what extent, does hierarchical structure play a role in cognitive processing at the level of syntax?

In contrast to what DMTP and others (e.g. Nelson et al., 2017) imply, we never argued that language use can be explained by *n*-grams only, or even that hierarchical structure never plays any role. Our point is that hierarchical processing may be less important than is traditionally assumed because simpler, sequential strategies are often available. Once one takes sequences, instead of syntactic hierarchies, as more fundamental to language use, it becomes pertinent to investigate which phenomena can be explained by sequential processing. We have shown here that the cortical entrainment results DMTP present as evidence for hierarchical processing are also generated by a time-series of word representations. Of course, this merely proves that a non-hierarchical explanation is possible, not that it is correct. In fact, the hierarchical model by Martin and Doumas (2017) can account for the same results.

Occam's Razor, however, would favour the simpler, non-hierarchical account until there is strong-enough evidence in support of the alternative.[1]

The principle of preferring simpler models over (unnecessarily) complex ones implies that hierarchical syntactic processing can only be concluded when the best non-hierarchical alternatives fail. Two recent studies compare specific predictions of a hierarchical processer to those of a sequential model and find that measures of hierarchical structure building do explain unique variance in neural activation (Brennan et al., 2016; Nelson et al., 2017). As such, these results more strongly support the hierarchical view than Ding et al.'s (2016) do. However, in both studies the "competition" is a simple bi- or trigram model, whereas much more sophisticated non-hierarchical language models exist. For example, recurrent neural networks[2] outperform *n*-gram models in predicting human reading times (Frank & Bod, 2011), event-related potential sizes during sentence reading (Frank et al., 2015), and sentence acceptability judgments (Lau, Clark, & Lappin, 2017). Similarly, the chunk-based learner (CBL) model, which uses backward transitional probabilities between words to discover variably-sized multiword chunks in child-directed speech, provides a superior computational account of shallow parsing of parental input and linearisation of child-produced utterances compared to *n*-gram models across several different languages (Chater, McCauley, & Christiansen, 2016; McCauley & Christiansen, 2011, 2017). Moreover, the psycholinguistic validity of the chunks discovered by CBL has been independently verified (Grimm, Cassani, Gillis, & Daelemans, 2017). More generally, the neural activation patterns observed by Brennan et al. (2016) and Nelson et al. (2017) might reflect effects of chunking across different levels of linguistic representation (i.e. Chunk-and-Pass processing[3]; Christiansen & Chater, 2016b), rather than rule-based hierarchical processing at the syntactic level; and may provide yet another explanation for Ding et al.'s (2016) entrainment results. Thus, in the absence of further evidence ruling out these simpler sequential processing accounts, DMTP's conclusion that "Recent neurolinguistic studies […] support the involvement of rule-based processing during language comprehension" (p. 570) seems premature.

Notes

1. For example, Frank and Yang (in press) suggest presenting participants with word-salad stimuli where every fourth word is a verb. Their model predicts a 1 Hz power peak in that case, whereas a syntactic structure building account would predict the absence of a 1 Hz peak.
2. Being universal function approximators, these networks are in principle able to implement a hierarchical system. However, they need to learn this from unannotated (i.e. only linearly structured) texts, and have a learning bias for short-distance dependencies (see, e.g. Christiansen & Chater, 1999). Consequently, they will only come to behave hierarchically when (and to the extent that) this is necessary for the task at hand. That is, any hierarchical processing in a recurrent neural network has emerged from a system that is first and foremost sequential.
3. Because Chunk-and-Pass processing involves both lossy compression, leading to reduction of bottom-up information, and incorporation of top-down information via predictions from semantics, pragmatics and real-world knowledge, chunking across levels of linguistic representations does not involve true part-whole hierarchical relationships (see Christiansen & Chater, 2016b, for further discussion).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Nederlandse Organisatie voor Wetenschappelijk Onderzoek: [grant number 024.001.006].

References

Armeni, K., Willems, R. M., & Frank, S. L. (2017). Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience and Biobehavioral Reviews*, *83*, 579–588. doi:10.1016/j.neubiorev.2017.09.001

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1-L2 differences. *Topics in Cognitive Science*, *9*, 621–636. doi:10.1111/tops.12271

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82. doi:10.1016/j.jml.2009.09.005

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157-158*, 81–94. doi:10.1016/j.bandl.2016.04.008

Chater, N., McCauley, S., & Christiansen, M. H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language*, *89*, 244–254. doi:10.1016/j.jml.2015.11.004

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205. doi:10.1207/s15516709cog2302_2

Christiansen, M. H., & Chater, N. (2016a). *Creating language: Integrating evolution, acquisition, and processing*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (2016b). The Now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, 279. doi:10.1017/S0140525X1500031X

Ding, N., Melloni, L., Tian, X., & Poeppel, D. (2017). Rule-based and word-level statistics-based processing of language: Insights from neuroscience. *Language, Cognition and Neuroscience*, *32*, 570–575. doi:10.1080/23273798.2016.1215477

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*, 158–164.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8*, e77661. doi:10.1371/journal.pone.0077661

Fodor, J. D., & Frazier, L. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*, 229–247. doi:10.1016/0010-0277(78)90002-1

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*, 829–834. doi:10.1177/0956797611409589

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, *279*, 4522–4531. doi:10.1098/rspb.2012.1741

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. doi:10.1016/j.bandl.2014.10.006

Frank, S. L., Trompenaars, T., & Vasishth, S. (2016). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, *40*, 554–578. doi:10.1111/cogs.12247

Frank, S. L., & Yang, J. (in press). Lexical representation explains cortical entrainment during speech perception. *PLoS ONE*.

Gardner, M. K., Rothkopf, E. Z., Lapan, R., & Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory & Cognition*, *15*, 24–28. doi:10.3758/BF03197709

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184. doi:10.1111/tops.12007

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2017). Facilitatory effects of multi-word units in lexical processing and word learning: A computational investigation. *Frontiers in Psychology*, *8*, 814. doi:10.3389/fpsyg.2017.00555

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, *41*, 1202–1241. doi:10.1111/cogs.12414

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review*, *109*, 35–54. doi:10.1037//0033-295X.109.1.35

Martin, A. E., & Doumas, L. A. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS Biology*, *15*, e2000663. doi:10.1371/journal.pbio.2000663

McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 1619–1624). Austin, TX: Cognitive Science Society.

McCauley, S. M., & Christiansen, M. H. (2017). Language learning as language use: A cross- linguistic model of child language development. Manuscript submitted for publication.

Mikolov, T., Chen, K., Corrado, C., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop*.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In W. Daelemans (Ed.), *Proceedings of the 13th conference of the European chapter of the association for computational linguistics* (pp. 398–408). Avignon: Association for Computational Linguistics.

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., … Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*, E3669–E3678. doi:10.1073/pnas.1701590114

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structural dependence and indirect statistical evidence. *Cognitive Science*, *29*, 1007–1028. doi:10.1207/s15516709cog0000_28

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*, 382–386. doi:10.1016/S1364-6613(02)01958-7

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319. doi:10.1016/j.cognition.2013.02.013

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, *26*, 2506–2516. doi:10.1093/cercor/bhv075

Zhou, H., Melloni, L., Poeppel, D., & Ding, N. (2016). Interpretations of frequency domain analyses of neural entrainment: Periodicity, fundamental frequency, and harmonics. *Frontiers in Human Neuroscience*, *10*, 274. doi:10.3389/fnhum.2016.00274