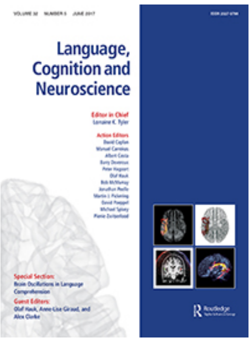

Language, Cognition and Neuroscience

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

Rule-based and word-level statistics-based processing of language: insights from neuroscience

Nai Ding, Lucia Melloni, Xing Tian & David Poeppel

To cite this article: Nai Ding, Lucia Melloni, Xing Tian & David Poeppel (2017) Rule-based and word-level statistics-based processing of language: insights from neuroscience, Language, Cognition and Neuroscience, 32:5, 570-575, DOI: 10.1080/23273798.2016.1215477

To link to this article: https://doi.org/10.1080/23273798.2016.1215477

Published online: 06 Aug 2016.

LANGUAGE, COGNITION AND NEUROSCIENCE, 2017
VOL. 32, NO. 5, 570–575
http://dx.doi.org/10.1080/23273798.2016.1215477

REGULAR ARTICLE

Rule-based and word-level statistics-based processing of language: insights from neuroscience

Nai Ding[a,b,c], Lucia Melloni[d,e], Xing Tian[f,g] and David Poeppel[h,i]

[a]College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, People's Republic of China; [b]Interdisciplinary Center for Social Sciences, Zhejiang University, Hangzhou, People's Republic of China; [c]Neuro and Behavior EconLab, Zhejiang University of Finance and Economics, Hangzhou, People's Republic of China; [d]Department of Neurology, New York University Langone Medical Center, New York, NY, USA; [e]Department of Neurophysiology, Max-Planck Institute for Brain Research, Frankfurt, Germany; [f]Neural and Cognitive Sciences, New York University Shanghai, Shanghai, People's Republic of China; [g]NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, Shanghai, People's Republic of China; [h]Department of Psychology, New York University, New York, NY, USA; [i]Neuroscience Department, Max-Planck Institute for Empirical Aesthetics, Frankfurt, Germany

ABSTRACT
To flexibly convey meaning, the human language faculty iteratively combines smaller units such as words into larger structures such as phrases based on grammatical principles. During comprehension, however, it remains unclear how the brain encodes the relationship between words and combines them into phrases. One hypothesis is that internal grammatical principles governing language generation are also used to parse the hierarchical syntactic structure of spoken language. An alternative hypothesis suggests, in contrast, that decoding language during comprehension solely relies on statistical relationships between words or strings of words, that is, the *N*-gram statistics, and no hierarchical linguistic structures are constructed. Here, we briefly review distinctions between rule-based hierarchical models and statistics-based linear string models for comprehension. Recent neurolinguistic studies show that tracking of probabilistic relationships between words is not sufficient to explain cortical encoding of linguistic constituent structure and support the involvement of rule-based processing during language comprehension.

ARTICLE HISTORY
Received 15 January 2016
Accepted 5 July 2016

KEYWORDS
Speech; language; neural oscillations; grammar; statistics

Introduction

It is vigorously debated whether language comprehension is driven by rule-based decomposition of hierarchical syntactic structures (Berwick & Weinberg, 1986; Everaert, Huybregts, Chomsky, Berwick, & Bolhuis, 2015; Phillips, 2003) or reflects the online analysis of the statistical relationship between adjacent words which obviates the need for abstract structure building (Elman, 1990; Frank, Bod, & Christiansen, 2012). For rule-based models, the hierarchical structure of linguistic input sequence must be revealed via syntactic analysis in order to comprehend spoken language. *N*-gram statistics-based models in contrast propose that the probabilistic relationships between (typically adjacent) words are sufficient for comprehension. Here we briefly discuss the distinctions and relationships between these two hypotheses and argue that recent neuroscientific data suggest that the brain can and does indeed represent hierarchical linguistic structures, even in the absence of relevant probabilistic information.

The predictive nature of language processing

It is well established that the brain actively makes predictions which allow quick processing of incoming words (Dikker, Rabagliati, Farmer, & Pylkkänen, 2010; Marslen-Wilson & Tyler, 1980; Poeppel, Idsardi, & von Wassenhove, 2008; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) and which can aid in enriching underspecified sensory information in challenging listening environments (Miller, Heise, & Lichten, 1951). For example, in noisy environments, sentences with higher transitional probability between words are better recognised than sentences with lower transitional probability (Miller et al., 1951). Furthermore, when a word in a highly constrained context is replaced by, say, a cough, listeners usually feel that they heard the full word on top of the cough sound (Warren, 1970).

A major motivation of statistical language models is to characterise how the brain generates predictions of future words. For an *N*-gram statistical model (Martin & Jurafsky, 2008), a future word, *W*, is predicted based on the $N-1$ words preceding that word, for example,

CONTACT Nai Ding ding_nai@zju.edu.cn

$W_{-(N-1)} \ldots W_{-2}W_{-1}$. Such predictions are based on the transition probability between the previous $N-1$ words and the future word, that is, $P(W|W_{-(N-1)} \ldots W_{-2}W_{-1})$, which is estimated based on previous language experience. Such models have offered successful applications in engineering contexts.

It has been controversial, however, whether *N*-gram models are sufficient to describe the human comprehension system. First, some sentences, although grammatical, have extremely low transition probabilities between words, as in the famous example coined by Chjomsky: "colorless green idea sleep furiously". Such syntactically correct but very-low predictability (and usually meaningless) sentences are processed differently from syntactically incorrect random word lists, as shown by abundant psycholinguistic and neurolinguistic studies (Friederici, Meyer, & von Cramon, 2000; Marslen-Wilson & Tyler, 1980; Pallier, Devauchelle, & Dehaene, 2011). For example, in a noisy environment, syntactically correct but semantically anomalous sentences are easier to recognise than ungrammatical sentences (Miller & Isard, 1963). Correct syntactic (or phonological) structures may also facilitate language processing by generating predictions (DeLong, Urbach, & Kutas, 2005). Such predictions, however, are based on tacit syntactic (or phonological) knowledge rather than *N*-gram transitional probability. For example, an adjective, for example, "green", predicts the forthcoming category noun, even if a low-probability one, such as "ideas".

Second, the grammars of human languages allow, in numerous linguistic contexts, long-distance dependencies between words. For example, "you can **either** read the first sentence of the first paragraph of the first book **or** not read it". The word "either" predicts the word "or" but the distance between them could be of any arbitrary length depending on the number of embedded clauses, and such long-distance dependencies pose a challenge for *N*-gram models. What underlies this problem is that human language is more complicated that can be described by an *N*-gram model (Berwick, Friederici, Chomsky, & Bolhuis, 2013; Chomsky, 1957; Fitch & Friederici, 2012). Such long-distance dependencies are very frequent and can also occur in other forms. Consider the following example: "These insects can digest wood because … in the morning they really like to eat pine." In this case, "pine" is predictable given the context of the discourse even though the local transitional probability between "eat" and "pine" is very low. Moreover, the long-distance dependency between the antecedent "insects" and the pronoun "they" is regulated by structural factors (i.e. specific structural, grammatical constraints exist that licence the interpretation) and not simple linear word distance.

In summary, an important asset of the *N*-gram statistical model is that it can easily generate predictions about future words and is mathematically approachable. However, not all aspects of human language processing can be characterised based on that model. In particular, the *N*-gram statistic is not the only source of information used to make predictions about incoming linguistic information (Jurafsky, 2003). For example, it cannot characterise predictions made based on syntactic information or discourse-level context. Therefore, the difference between rule-based hierarchical models and *N*-gram models is not whether the brain makes predictions or whether the brain is sensitive to statistical regularities. These points are uncontroversial. The crucial difference is over what kind of linguistic units – hierarchical constituent structures or linear *N*-word strings – the brain tracks statistical regularities and generates predictions (Townsend & Bever, 2001).

Relationship between statistical models and rule-based hierarchical models

Although it has been debated whether the brain processes language based on statistics or rules, statistics-based processing and rule-based processing are related and not mutually exclusive. First, syntactic rules give rise to statistical cues. On the view that language is generated based on a set of rules, only some sequences of words are allowed (and therefore typical and frequent), that is, the grammatical ones. In daily life, the probability of being exposed to an ungrammatical sentence is fairly low and therefore the brain mainly accumulates statistics based on grammatical sentences. In this set of grammatical sequences, the transitional probability is not equal between pairs of words and can be learned to facilitate language processing.

Second, it has been proposed that rules can be learned based on statistical cues. For example, it has been shown that 8-month-old infants are sensitive to the transitional probability between syllables, which may serve as cues to segment a continuous speech stream into words (Peña, Bonatti, Nespor, & Mehler, 2002; Saffran, Aslin, & Newport, 1996). Such statistical learning paradigms can also underpin the learning of phrasal structures (Thompson & Newport, 2007) and rules (Marcus, Vijayan, Rao, & Vishton, 1999). The difference between rules and statistics, however, concerns the levels of abstraction. For example, after being exposed to a large number of noun phrases, one may simply learn the frequency of one word appearing after another $N-1$ words, but one may also abstract a set of rules, for example, a class of words that can be used to modify another class of words (Saffran et al., 1996;

Seidenberg, MacDonald, & Saffran, 2002). Such abstraction could be an implicit and subconscious process during language acquisition, while it could also be an explicit process when learning grammar in school.

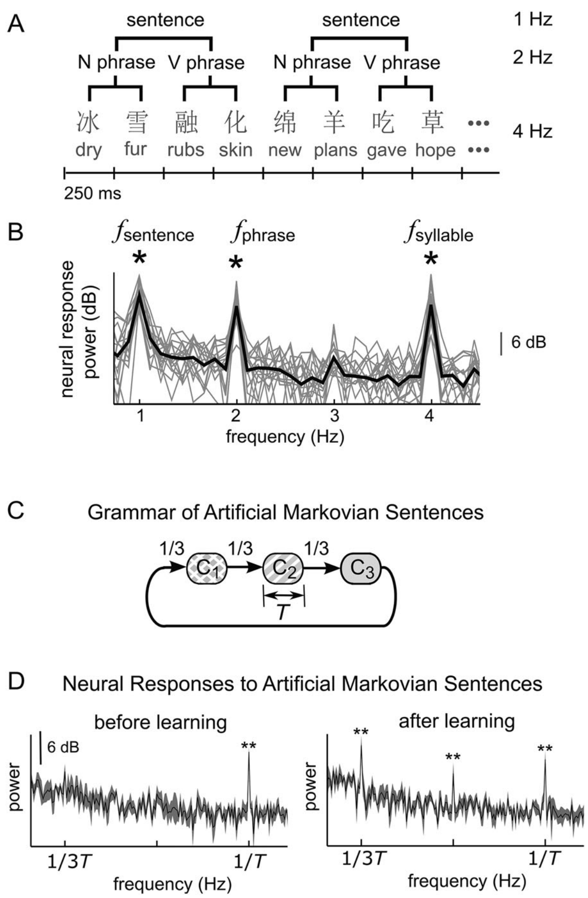

Figure 1. Cortical tracking of the linguistic structure of speech. (a) The grammar of a set of short Chinese sentences in which the syllables are presented at a constant rate of 4 Hz. The phrases and sentences are presented at 2 and 1 Hz, respectively, because of binary embedding of linguistic structures. (b) Neural response spectrum (global field power) shows peaks at the syllabic rate, phrasal rate, and the sentential rate, demonstrating concurrent neural tracking of three linguistic levels. (c) The grammar of a set of Artificial Markovian Sentences (AMS). Each sentence consists of three components, C_1, C_2, and C_3. Each component is independently chosen from three candidate syllables with equal probability. The stimulus-onset asynchrony (SOA) between syllables is a constant, $T = 0.3$ s. In each trial, 33 sentences are played in a sequence without any gap in between. (d) Neural response spectrum (global field power) before and after learning the AMS grammar. Before learning, cortical activity only tracks the syllabic rhythm of speech. After learning, however, cortical activity concurrently follows the syllabic rhythm at $1/T$ and the sentential rhythm at $1/3\ T$. Frequency bins showing power stronger than the mean power of a neighboring 1 Hz region (i.e. 0.5 Hz on each side) are shown by stars ($N = 5$, $P < .001$, paired t-test, false discovery rate corrected) (adapted from Figure 1 and Supplementary Figure 4 of Ding et al., 2016).

Therefore, abstraction/generalisation might be a potential link between rule-based processing and statistics-based processing (Marcus, 1999; Seidenberg et al., 2002). Even if considering only statistics-based processing, generalisation is critical due, for example, to poverty of stimulus considerations (Chomsky, 1957). For example, the sentence "university professors never assign homework" is not a strange sentence – but most people have never been exposed to this exact sentence. It is not sensible to assume the probability of such a sentence to be zero just because it has never been heard/seen. In fact, modern statistical models do not simply count how many times a sequence of words appear but instead build models that can generalise (Pereira, 2000). If the brain is not simply counting word frequencies but instead makes generalisations, it must have an internal model about *how* to make generalisations. Such internal models may not be critically different from syntactic or semantic knowledge. One important question, however, is what kind of generalisation is made by the brain and how abstract, that is, rule-like, such generalisations are.

One fundamental distinction between rule-based models and *N*-gram models is that rule-based linguistic theories describe the relationship between words using hierarchical syntactic "chunks", while *N*-gram models by-and-large assume a linear relationship between words. An *N*-gram model is not the only model to describe the statistical regularities in language. More sophisticated statistical models, such as probabilistic context-free grammars, assume a hierarchically embedded phrasal structure and are compatible with symbolic rules and representations (Chater & Manning, 2006; Hale, 2001). These rule-level or phrasal-level probabilistic models, in contrast with the word-level *N*-gram models, are consistent with the particular rule-based hierarchical structure models we discuss here.

Hierarchical structure building and its neural correlates

Research on the neural encoding of speech and sound-sequence processing can shed light on the debate between rule-based models and *N*-gram models (Bahlmann, Gunter, & Friederici, 2006; Brennan et al., 2012; Dikker et al., 2010; Fitch & Friederici, 2012; Friederici, Bahlmann, Friedrich, & Makuuchi, 2011; Pallier et al., 2011). For example, functional magnetic resonance imaging studies have shown that rule-based construction of hierarchical linguistic structures mainly occurs in the left inferior frontal gyrus, for example, Brodmann area 44 and temporal areas (Fitch & Friederici, 2012).

In terms of neurophysiological studies, on the one hand, a large body of literature has demonstrated that the brain is sensitive to various types of statistical cues, even without any rule-based structure (Kutas & Federmeier, 2000; Näätänen, Paavilainen, Rinne, & Alho, 2007). Furthermore, statistical learning can also lead to neural tracking of statistically defined linguistic structures (Buiatti, Peña, & Dehaene-Lambertz, 2009; Kabdebon, Pena, Buiatti, & Dehaene-Lambertz, 2015). On the other hand, there is also neurophysiological evidence for purely rule-based hierarchical linguistic processing. Here we briefly review neurophysiological evidence supporting the following two claims.

First, the brain can simultaneously represent hierarchical phrasal structures, that is, syntactic chunks of different sizes, resulting in a multi-resolution representation of the input sequence. For example, it has been shown that violating discourse-level context evokes the classic N400 response, similar to what is observed when the local sentential context is violated (Van Berkum, Zwitserlood, Hagoort, & Brown, 2003). This result demonstrates that the brain can detect a violation of local and global context within a similar time window, that is, within half a second after the word onset. This suggests that brain maintains a representation of both local and global context which can be promptly retrieved. It is difficult to explain such a phenomenon using nothing more than an *N*-gram model, since modeling the global context requires integration of tens of words, which is beyond the limit of human working memory. To consider another example, it has been shown that, during listening to spoken language, cortical activity concurrently follows the rhythms of linguistic structures of different sizes, for example, words, phrases, and sentences (Figure 1(a,b)), providing direct evidence for simultaneous neural representations of hierarchical linguistic structures (Ding, Melloni, Zhang, Tian, & Poeppel, 2016).

Second, neural representations of phrasal structures (i.e. syntactic chunks) can be formed without statistical cues. Evidence supporting this claim mostly comes from studies using artificial sequences, which are parsed based on explicitly instructed rules. The logic is to show cortical encoding of phrasal chunks in the absence of any relevant statistical cue, and this is achieved by explicitly learning the phrasal construction rules. In one example, to dissociate linguistic structures from statistical cues, a special Markovian sequence is constructed in which the transitional probability between adjacent syllables is constantly 1/3. The Markovian sequence alternates among three states, C_1, C_2, and C_3 (Figure 1(c)), and the states are independent from one another. In each state, a syllable will be drawn from three candidate syllables with equal probability. Each state is associated with a distinct set of candidate syllables, and a sequence of three consecutive states, that is, $C_1C_2C_3$, is viewed as a sentence. The transitional probability between syllables, however, is constant within a sentence or across sentence boundaries.

When listening to such a sequence without any instruction about the stimulus structure, cortical activity recorded by magnetoencephalography only follows the syllabic rhythm (Figure 1(d)). The listeners were then instructed about the sentential structure and exposed to stimuli that contained a short gap after C_3, which facilitates the learning of sentential structures. They were instructed to memorise the set of syllables belonging to each state. After this learning phase, when exposed to the Markovian sequence again, cortical activity tracking the sentential structure emerges (Figure 1(d)). This result demonstrates that the brain can parse learned linguistic structures even in the absence of transitional probability cues.

Further evidence comes from studies on artificial musical sequences (Nozaradan, Peretz, Missal, & Mouraux, 2011). When listening to an isochronous tone sequence, the listeners were either instructed to listen to an isochronous sequence (i.e. x x x x x x) or to imagine a binary (X x X x X x) or ternary metre structure (X x x X x x). It is observed that cortical activity measured by electroencephalography only follows the repetition rate of tones when the listeners were asked to listen to an isochronous sequence. When asked to imagine a metre structure, however, additional neural tracking of the metre structure emerged. Since the metre structure is *imagined*, not associated with acoustic or statistical cues, neural tracking of the metric structure can only be explained by rule-based processing rather than statistics-based processing. Of course, binary/ternary grouping can be described by a Markov model, and so is the sentential grouping in Figure 1(c,d). Nevertheless, what is important is that even if such grouping is achieved by Markovian processes, the processes are based on rules, not input statistics. The above examples show that cortical activity can track phrasal structures in the absence of any statistical cues, providing compelling evidence that the brain can form phrasal-level representations based on rules.

In summary, word-level input statistics alone are not sufficient and, in many cases, not necessary to explain human language processing performance or the neural responses to language or sound sequences. Input word-level statistics, however, can trigger the learning of syntactic rules or other more abstract processing models. Future research needs to establish what kind of knowledge is gained during statistical learning and how abstract it is. If the brain uses input statistics to fit an internal language model, it has to be investigated what kind of model it is. Furthermore, rule-based processing does not deny that language processing is highly predictive but assumes that predictions will be made, among other factors, based on a hierarchically nested syntactic structure rather than a linear string structure. Using the paradigm in Figure 1(a,b), future neurophysiological studies can shed light on whether hierarchically nested structures are constructed online during speech perception and how deeply embedded the phrasal structure could be.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China [31500873 (ND) and 31500914 (XT)]; Fundamental Research Funds for the Central Universities (ND); Zhejiang Provincial Natural Science Foundation of China [LR16C090002 (ND)]; the Program of Introducing Talents of Discipline to Universities Base [B16018 (XT)]; Major Projects Program of the Shanghai Municipal Science and Technology Commission [15JC1400104 (XT)]; and the US National Institutes of Health grant [NIH R01 DC 05660 (DP)].

References

Bahlmann, J., Gunter, T. C., & Friederici, A. D. (2006). Hierarchical and linear sequence processing: An electrophysiological exploration of two different grammar types. *Journal of Cognitive Neuroscience*, *18*(11), 1829–1842. doi:10.1162/jocn.2006.18.11.1829

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in Cognitive Sciences*, *17*(2), 89–98. doi:10.1016/j.tics.2012.12.002

Berwick, R. C., & Weinberg, A. S. (1986). *The grammatical basis of linguistic performance: Language use and acquisition*. Cambridge, MA: MIT press.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*(2), 163–173. doi:10.1016/j.neuroimage.2012.01.030

Buiatti, M., Peña, M., & Dehaene-Lambertz, G. (2009). Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *Neuroimage*, *44*(2), 509–551. doi:10.1016/j.neuroimage.2008.09.015

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344. doi:10.1016/j.tics.2006.05.006

Chomsky, N. (1957). *Syntactic structures*. Berlin: Mouton de Gruyter.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8* (8), 1117–1121. doi:10.1038/nn1504

Dikker, S., Rabagliati, H., Farmer, T. A., & Pylkkänen, L. (2010). Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science*, *21*(5), 629–634. doi:10.1177/0956797610367751

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*. doi:10.1038/nn.4186

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. doi:10.1016/0364-0213(90)90002-E

Everaert, M. B., Huybregts, M. A., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, *19*(12), 729–743. doi:10.1016/j.tics.2015.09.008

Fitch, W. T., & Friederici, A. D. (2012). Artificial grammar learning meets formal language theory: An overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367* (1598), 1933–1955. doi:10.1098/rstb.2012.0103

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, *279*(1747), 4522–4531. doi:10.1098/rspb.2012.1741

Friederici, A. D., Bahlmann, J., Friedrich, R., & Makuuchi, M. (2011). The neural basis of recursion and complex syntactic hierarchy. *Biolinguistics*, *5*(1–2), 87–104.

Friederici, A. D., Meyer, M., & von Cramon, D. Y. (2000). Auditory language comprehension: An event-related fMRI study on the processing of syntactic and lexical information. *Brain and Language*, *74*, 289–300. doi:10.1006/brln.2000.2313

Hale, J. (2001). *A probabilistic Earley parser as a psycholinguistic model*. Paper presented at the Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics, Pittsburgh, PA. doi:10.3115/1073336.1073357

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod & J. Hay (Eds.), *Probabilistic linguistics*. Cambridge, MA: Mit Press.

Kabdebon, C., Pena, M., Buiatti, M., & Dehaene-Lambertz, G. (2015). Electrophysiological evidence of statistical learning of long-distance dependencies in 8-month-old preterm and full-term infants. *Brain and Language*, *148*, 25–36. doi:10.1016/j.bandl.2015.03.005

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. doi:10.1016/S1364-6613(00)01560-6

Marcus, G. F. (1999). Do infants learn grammar with algebra or statistics? Response to Seidenberg & Elman, Negishi, and Eimas. *Science*, *284*(5413), 436–437. doi:10.1126/science.284.5413.433f

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80. doi:10.1126/science.283.5398.77

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1–71. doi:10.1016/0010-0277(80)90015-3

Martin, J. H., & Jurafsky, D. (2008). *Speech and language processing*. Upper Saddle River, NJ: Prentice Hall.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials.

Journal of Experimental Psychology, *41*, 329–335. doi:10.1037/h0062491
Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, *2*(3), 217–228. doi:10.1016/S0022-5371(63)80087-0
Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, *118*(12), 2544–2590. doi:10.1016/j.clinph.2007.04.026
Nozaradan, S., Peretz, I., Missal, M., & Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *Journal of Neuroscience*, *31*, 10234–10240. doi:10.1523/JNEUROSCI.0411-11.2011
Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527. doi:10.1073/pnas.1018711108
Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604–607. doi:10.1126/science.1072901
Pereira, F. (2000). Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society of London, Series A, Mathematical, Physical and Engineering Sciences*, *358*(1769), 1239–1253. doi:10.1098/rsta.2000.0583
Phillips, C. (2003). Linear order and constituency. *Linguistic Inquiry*, *34*(1), 37–90. doi:10.1162/002438903763255922
Poeppel, D., Idsardi, W. J., & von Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1071–1086. doi:10.1098/rstb.2007.2160
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. doi:10.1126/science.274.5294.1926
Seidenberg, M. S., MacDonald, M. C., & Saffran, J. R. (2002). Does grammar start where statistics stop. *Science*, *298*(5593), 553–554. doi:10.1126/science.1078094
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634. doi:10.1126/science.7777863
Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1–42. doi:10.1080/15475440709336999
Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. Cambridge, MA: MIT Press.
Van Berkum, J. J., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research*, *17*(3), 701–718. doi:10.1016/S0926-6410(03)00196-4
Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393. doi:10.1126/science.167.3917.392