To Predict or to Memorize: Prominence in Inaugural Addresses

William Clapp & Arto Anttila
Stanford University

1 Introduction

Words in an English sentence are typically characterized by a range of prominence. Consider the following sentence as delivered by Ronald Reagan on January 20, 1981:

(1) Their patriotism is quiet, but deep.

The prominence contour perceived by one native speaker is shown in Figure 1 (left); see Anttila, Dozat, Galbraith, and Shapiro (2020) for the annotation. Column height corresponds to the degree of perceived prominence. In this six-word sentence the annotator perceived four degrees of stress. The function words *their*, *is*, and *but* have a low degree of stress, whereas the content words *patriotism*, *quiet*, and *deep* show a steadily rising contour, with the primary (nuclear) stress at the right edge on *deep*.

Why are sentences stressed the way they are? Two classical views can be found in the literature (Gussenhoven, 2011). One view considers sentential prominence to be (at least partly) a matter of stress derived mechanically from syntax and phonology (Chomsky & Halle, 1968; Liberman & Prince, 1977; Cinque, 1993). SPE (Chomsky & Halle, 1968) proposed two phrasal stress rules, The Nuclear Stress Rule (NSR) and the Compound Stress Rule (CSR), that assign stress to syntactic constituents cyclically, inside out, yielding a hierarchical stress contour where one word carries primary stress. The theory is simple and explicit and can be practically implemented and tested. Figure 1 (right) shows the stress contour derived by the METRICALTREE implementation of the SPE stress rules (Anttila et al., 2020). The predicted contour is not identical to the perceived contour, but it is reasonably close: the model captures the weakness of the function words as well as the overall rising contour, placing primary stress correctly on *deep*.

An alternative view holds that such mechanical stress rules are illusory: sentential prominences are not a matter of stress, but of pitch accents that are individually meaningful and whose distribution reflects the speaker's intent, with accents falling on information foci in the sentence (Bolinger, 1972). This theory is much harder to test since we do not have direct access to the speaker's (or the listener's) mental states and expressive goals, but we may be able to approximate meaning in terms of informativity: informative words tend to be stressed, uninformative words tend to be unstressed. Starting with the raw lexical frequency of the six words in the inaugural corpus beginning with Roosevelt we get the string 148, 4, 500, 8, 232, 9, which correlates nicely with the stress contour: function words are frequent and weakly stressed, content words are infrequent and strongly stressed. A more sophisticated measure is bigram informativity: the weighted average of the negative log probability of seeing a word w given every context c that w follows in the corpus (Piantadosi et al., 2011; Cohen Priva, 2012, 2015). Using this measure we get 4.89, 9.67, 3.30, 8.36, 4.39, 8.29, showing that weakly stressed function words indeed have low informativity and strongly stressed content words have high informativity. Frequency and informativity seem to capture the function word vs. content word distinction, but the overall rising stress contour and the placement of primary stress on the rightmost content word remain a mystery.

* This work is based on research partially funded by the Office of the Vice-Provost for Undergraduate Education (VPUE) at Stanford University and by the Roberta Bowman Denning Initiative Committee, HS Dean's Office, as part of the project *Prose Rhythm and Linguistic Theory*. We are deeply indebted to Timothy Dozat, Daniel Galbraith, Naomi Shapiro, and Alex Wade for their help in the early stages of this work, some of which is reported in Anttila, Dozat, Galbraith, & Shapiro 2020. We thank Joan Bresnan, Vivienne Fong, Bruce Hayes, Dan Jurafsky, and Michael Wagner for their input. We are responsible for any errors.

Proceedings of *AMP 2020*

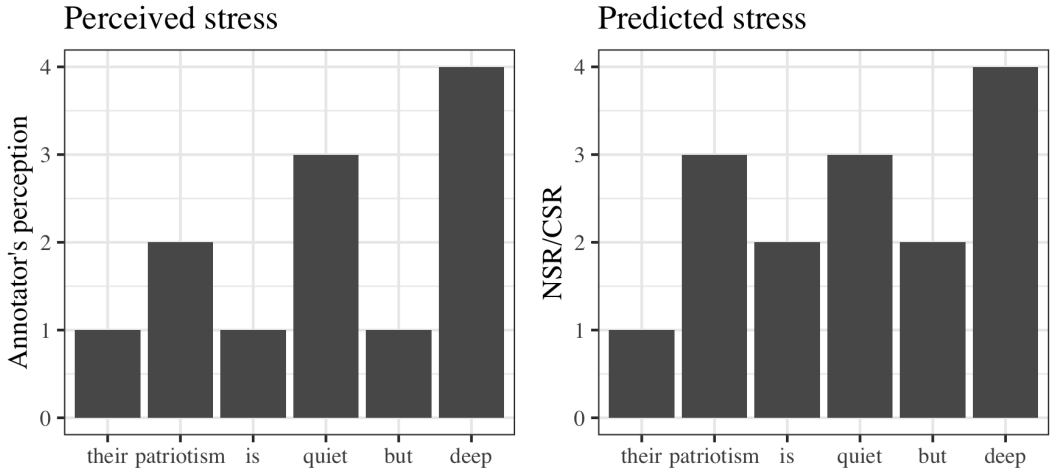

Figure 1: Sentence 75 from Ronald Reagan's first inaugural address as delivered on January 20, 1981. The figure on the left shows one annotator's interpretation of the structure of sentential prominence. The figure on the right shows the predictions made by the METRICALTREE stress algorithm. The scale on the right has been inverted for clarity.

A more recent account consistent with exemplar-theoretic approaches to phonology asserts that prominence is memorized on a by-word basis using ACCENT RATIO (henceforth, AR), which Nenkova et al. (2007) define as the likelihood "that a word belongs to a prominence class or not," assuming a binary of prominence. More specifically, AR for each word is calculated according to the following equation:

(2) $$\textit{AccentRatio(w)} = \begin{cases} \frac{k}{n} & \texttt{if } B(k,n,0.5) \leq 0.05 \\ 0.5 & \texttt{otherwise} \end{cases}$$

In this equation *k* represents the number of times in a given corpus that a word *w* is realized as prominent, and *n* represents the total number of times that word *w* appears. *B*(*k*, *n*, 0.5) is the probability within a binomial distribution of the observed ratio of instances of prominent *k* to total *n*, given that there is an equal probability of *w* being realized as prominent or not. In other words, AR represents the likelihood that a given word will surface as prominent, as long as that likelihood is significantly different than 0.5. Nenkova et al. (2007) showed that AR outperforms many other variables—including part of speech, position in sentence, length of word, and information status, among others—in predicting whether a particular word would be realized as prominent. This was true not only when the AR dictionary was applied to a unique corpus from the same superset as that on which it was trained, namely conversations from the Switchboard Corpus (Godfrey et al., 1992), but also when the AR dictionary was applied to a new corpus across a genre boundary, the Boston University Radio corpus of broadcast news (Ostendorf et al., 2001).

Although the descriptive fit of AR is impressive, the approach raises a number of questions. One is whether sentence accent should be treated as a binary variable. Nenkova et al. (2007) used a corpus where each word was manually labeled for the presence vs. absence of pitch accent. This differs from the metrical view taken by us which considers sentential prominence to be a matter of hierarchical stress coded in terms of degrees of relative prominence with no upper/lower limit, something that is reflected in the traditional descriptive labels "primary stress", "secondary stress", "tertiary stress", etc. (Gussenhoven, 2011). Such labels seem to have some cognitive reality and they are certainly closer to the actual realization of prominence in terms of continuous variables—including f0, duration, intensity, and vowel quality (e.g., van Heuven, 2018, among others).

Another question the AR approach leaves unanswered is how it is determined whether a given token will be realized with or without prominence. For example, the word *again* has an AR value of 0.75, meaning that it ought to surface as prominent approximately 75% of the time, but the AR approach makes no prediction about which variables determine precisely *when* it will be prominent, or *where* on the scale of prominence it will fall. In other words, AR can predict that the word *again* will be prominent 75% of the time, but makes no predictions about in which specific cases that word is likely to be produced with high or low prominence. Along the same lines, research in perceptual phonetics has demonstrated that listeners are often sensitive to variation in prominence, even within a range that a binary assignment system may lump into a single prominence class (Ladd & Morton, 1997; Wagner & Watson, 2010). While Chomsky & Halle's (1968) theory of phrasal stress and Bolinger's (1972) theory of sentence accent are formulated in ways that account for non-polar, non-binary prominence, AR does not have a mechanism to deal with such datapoints without outsourcing them to other theories.

With these successes and shortcomings of AR in mind, we sought to answer the question of whether AR obviates the predictive roles of syntax and informativity in prominence assignment, and if not, whether syntax and phonology shed light on any of the variance in prominence left unexplained by the AR approach.

2 Data & Methods

To answer these questions, we used a corpus of spoken American English consisting of the first inaugural addresses of six recent American presidents, including Jimmy Carter [1977], Ronald Reagan [1981], George H. W. Bush [1989], Bill Clinton [1993], George W. Bush [2001], and Barack Obama [2009]. Each of the approximately 11,500 words in the corpus was annotated for phrasal prominence as perceived by two native English speakers who used the web application METRICGOLD (Shapiro, 2019) to input their responses. Following Prince (1983), stress was transcribed in metrical grids. The definition of "stress" given to the annotators was "syllable prominence intuitively felt by a native speaker." In order to determine the prominence structure of each sentence, annotators were told to use the following cues in the following order:

(3) (i) Their own intuitions as native speakers.
(ii) Embodied cues, including the annotator tapping or humming along, or the speaker's gestures.
(iii) Explicit linguistic (e.g., phonetic) cues.

The two annotators were native speakers of English and had completed coursework in phonology. In general there was a high level of correlation between the responses of the two annotators ($\rho_T = 0.848$).

We have found inaugurals to be a unique source of data for the study of sentential prominence. The speeches are scripted and the performances are virtually disfluency-free. While oratorical prose is a very specific genre we have no reason to believe that its stress rules would be fundamentally different from those of the American English naturally spoken by the president and his audience. Inaugural addresses also provide a context in which the speaker is emotionally involved in the content. Such situations tend to increase the prosodic range of utterances (e.g., Frick, 1985), which makes it easier for the annotators to identify even subtle differences in levels of prominence between words. In this regard, inaugurals are like the speech of radio announcers characterized by "natural but controlled style, combining the advantages of both read speech and spontaneous speech" (Hasegawa-Johnson et al., 2005).

To derive the default stress contour we used a version of the SPE stress rules: the Nuclear Stress Rule (NSR) and the Compound Stress Rule (CSR), with some modifications (Anttila et al., 2020). The NSR/CSR was implemented computationally using the METRICALTREE software (Dozat, 2017) applied to syntactic parse trees generated by the Stanford Parser (Klein & Manning, 2003; Chen & Manning, 2014; Manning et al., 2014). In this study, we used predicted stress values normalized by sentence length. The most prominent word in each sentence was assigned a value of 1, and all other words were assigned a proportional value between 0 and 1. For example, the words of the sentence *Their patriotism is quiet, but deep* are assigned the following values from METRICALTREE: 4, 2, 3, 2, 3, 1, following the SPE numerology where higher numbers are interpreted as less prominent. After normalization, the words of the same sentence are associated with the values 0.25, 0.75, 0.5, 0.75, 0.5, 1.0. Note that the scale has reversed such that higher values are now associated with greater prominence.

Bigram informativity was used as a proxy for Bolinger's (1972) notion of predictability. Bigram informativity is defined as the weighted average of the negative log probability of seeing a word *w* given

every context *c* that *w* follows in the corpus (Piantadosi et al., 2011; Cohen Priva, 2012, 2015). In order to increase the utility of this value for each word, bigram informativity was calculated drawing from the entire corpus beginning with Roosevelt [1933]. Because bigram informativity is an imperfect representation of Bolinger's notion of predictability, additional models were computed using log frequency across the entire corpus instead of informativity.

Rather than relying on syntactic structure or communicative goals, the AR approach posits that prominence is memorized, assigning a value to each lexical item based on the ratio of stressed to total tokens of that lexical item in some corpus, given that the ratio is significantly different from 0.5. This perspective is consistent with exemplar-theoretic approaches to phonology, which assert that speech production and perception are facilitated by the storage of detailed auditory memories. Rather than calculating unique AR values based on prominence in our corpus, we chose to use the preexisting AR dictionaries from Nenkova et al. (2007). This decision was made in part to avoid a circular confound where predictions about sentential prominence would be informed by the same corpus where those predictions would be tested. It has the added advantage of testing Nenkova et al.'s (2007) claim that AR should remain a powerful predictor even across genres of speech. Unfortunately, words not included in the provided AR dictionary had to be excluded from the analysis, which resulted in the exclusion of nearly half the corpus—a decrease from approximately 11,500 words to approximately 6,800.

3 Results & Analysis

All models used in the analysis were mixed effects linear regression models fitted using the *lme4* package in R version 4.0.3 (Bates et al., 2015; R Core Team, 2020). The first model had a response variable of log perceived prominence with predictors of AR, bigram informativity, and the NSR/CSR, in addition to a number of other linguistic variables as controls. These variables included syntactic category, number of segments, and number of lexically stressed syllables. Following Barr et al. (2013), the maximal random effects structure that allowed the model to converge was used. This resulted in random intercepts for annotator and president. The model showed highly significant results for all three of the central predictors. For AR, β = 6.605e-01, SE = 2.130e-02, $p < 0.001$. For bigram informativity, β = 6.713e-02, SE = 2.707e-03, $p < 0.001$. For NSR/CSR, β = 1.231e-01, SE = 2.025e-02, and $p < 0.001$. All linguistic control variables were also highly significant ($p < 0.001$), with the exception of the syntactic category *other* (representing words that cannot be classified as nouns, verbs, adjectives, adverbs, or function words), which was significant ($p < 0.05$) and *noun* which was not significant ($p > 0.05$). The relationship between each of the primary variables and perceived prominence is plotted in Figure 2.

We also fitted an alternative model where all things were held constant except that we used log frequency across the corpus instead of bigram informativity as a predictor. In this model, the three primary predictors remained highly significant. For AR, β = 6.877e-01, SE = 2.100e-02, $p < 0.001$. For log frequency, β = -7.027e-02, SE = 2.901e-03, $p < 0.001$. For NSR/CSR, β = 9.716e-02, SE = 2.029e-02, $p < 0.001$. The pattern of significance among control variables was the same as reported above; namely, all were highly significant ($p < 0.001$) except the *noun* and *other* syntactic categories, which were not significant ($p > 0.05$) with a reference level of *adjective*. It should be noted that the effect direction of log frequency is opposite of that of bigram informativity. This is because prominence is associated with higher informativity values, but lower frequency values (i.e., lower frequency is correlated with higher informativity).

After the initial models had been fitted and the visualizations in Figure 2 had been drawn, we observed that although AR, NSR/CSR, and informativity all have a highly significant degree of predictive power regarding prominence, none of the predictors seems equally predictive across its entire range of values. The curve drawn for AR, for example, is relatively steep for values below 0.5, but nearly flat for values above 0.5. It is possible to interpret this observation such that the predictive success of AR is largely confined to its lower range. In other words, as AR increases from 0 to 0.5, it is very likely that perceived prominence will also increase. However, as AR increases from 0.5 to its ceiling of 1, there may be no observable trend in prominence. The plot associated with mechanical stress shows exactly the opposite trend. In the lower range of NSR/CSR, the smoother is virtually flat, but begins to rise sharply and abruptly in its upper range. One interpretation of these observations is that while a predictor like AR is more influential among low-prominence words, a predictor like NSR/CSR is more influential among high-prominence words.

In order to investigate this further, the data was divided into two subsets: one consisting of all data

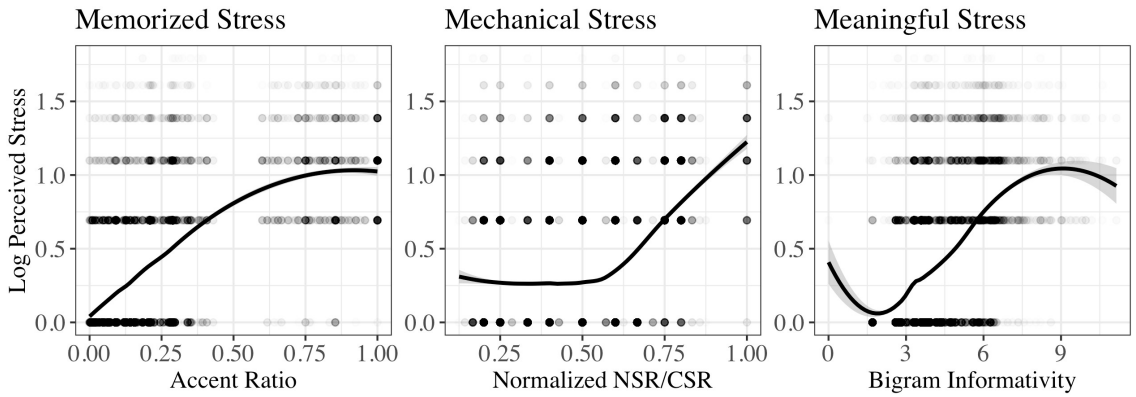

Figure 2: Log perceived stress across the entire corpus, as predicted by AR, NSR/CSR, and bigram informativity. As the loess smoothers show, AR and bigram informativity are most predictive in their lower ranges—below approximately 0.5 and 6, respectively—but that NSR/CSR is most predictive in its upper range, above approximately 0.6.

points where log perceived stress fell above the mean for the entire corpus and the other where all data points fell below the mean for the entire corpus. This schema allows us to address the question: What are the variables that best predict how prominent a word will be at the lower vs. the higher end of the prominence spectrum? Fitting the same model parameters as described above to these more constrained sets of data turned up surprisingly stark results that seem to show that all variables except for NSR/CSR are predictive among low-prominence words, but *only* NSR/CSR is predictive among high prominence words.

Models for the datasets separated based on perceived prominence were again fitted using the *lme4* package in R version 4.0.3 (Bates et al., 2015), and make use of the previously described parameters. The response variable was log perceived prominence, and independent variables included AR for memorized stress, normalized NSR/CSR for mechanical stress, and either bigram informativity or log frequency across the corpus for meaningful stress. The control variables number of stressed syllables, syntactic category, and number of segments were again included. The models were also fitted with random intercepts for annotator and president.

In the model fitted to low-prominence words with bigram informativity standing in for Bolinger's predictability, AR was highly significant (β = 4.775e-01, SE = 1.974e-02, $p < 0.001$), as was bigram informativity (β = 4.551e-02, SE = 2.302e-03, $p < 0.001$). Normalized NSR/CSR was not significant ($p > 0.05$). The control variables were all highly significant except for the categories *noun* and *other*, which were not significant ($p > 0.05$). The results of the model which included log corpus frequency rather than bigram informativity were similar. AR was highly significant (β = 5.168e-01, SE = 1.924e-02, $p < 0.001$), as was log corpus frequency (β = -4.450e-02, SE = 2.389e-03, $p < 0.001$). Normalized NSR/CSR was not significant ($p > 0.05$). The control variables were all highly significant except for the category *noun*, which showed only trend-level significance ($p < 0.1$), and *other*, which was not significant ($p > 0.05$).

The models evaluating the dataset consisting of only highly prominent words turned up quite different results. In the model including bigram informativity for Bolinger's predictability, normalized NSR/CSR was the only highly significant predictor of perceived prominence (β = 7.054e-02, SE = 2.086e-02, $p < 0.001$). Bigram informativity was significant only at the trend level ($p < 0.1$) and all other predictors, including the control variables, were not significant ($p > 0.05$). In the model where log corpus frequency was included rather than bigram informativity, the results were similar. Normalized NSR/CSR was highly significant (β = 7.005e-02, SE = 2.103e-02, $p < 0.001$), but all other variables, including log corpus frequency, were insignificant ($p > 0.05$).

The results of these models seem to point toward a system where the assignment of sentential prominence is influenced by a wide range of factors, including syntactic structure, the speaker's expressive goals, and

memorized prominence patterns associated with individual words. But perhaps more crucially, these factors do not seem to operate symmetrically across situations. Given that a word is relatively unprominent, its specific degree of prominence is much better predicted by informativity and memorized patterns, but for those words that are stressed, syntactic structures have a much greater level of predictive power.

4 Summary & Conclusions

While AR is an excellent predictor of prominence, its explanatory power goes only so far. The significance of the NSR/CSR even in a model controlling for AR and a number of linguistic variables suggests that sentential prominence is assigned at least to some degree according to syntactic structure rather than purely through memorization, informativity, and other linguistic variables. The exclusive significance of the NSR/CSR among words perceived as prominent indicates that syntax guides prominence assignment in its upper register, whereas lexical variables are more influential in the lower register. Although a work in progress, this finding contributes to a body of research seeking to tease out the nuanced ways in which sentential prominence is assigned by a range of variables. In particular, the possible schism in sources of assignment identified in the present work is reminiscent of Anttila et al.'s (2020) observation that "noun and adjective stresses are loud and mechanical whereas verb and function word stresses are soft and meaningful." Indeed, a deeper look into the distribution of syntactic categories that make up the two subsets of our data illuminates a pattern compatible with this notion. In the low-prominence subset, nouns and adjectives make up only 4.2% of the observations, while verbs and function words make up 92.3%. Meanwhile, in the subset of the data consisting of only high-prominence words, nouns and adjectives represent 38.5% of the data while the portion of the data made up of verbs and function words drops to 47.2%. It cannot be an accident that the distributions seen in the two subsets are so obviously different. It follows that nouns and adjectives, which primarily occupy the high-prominence subset, are more responsive to stress assignment through syntax, whereas verbs and function words, which primarily occupy the low-prominence subset, are more responsive to parameters such as AR and informativity along with other linguistic variables.

What all of this together gestures toward is a system where no one predictor can be privileged above any other. All the investigated variables have a substantial role in prominence assignment, but what should not be taken away is that the relationship between the stress-assigning variables is without sense or internal structure. Although each of the factors explored here matters, each is weighted differently depending on the context, suggestive of a complex system of intertwined variables that work in concert with one another to produce the sentential contours we hear as listeners.

References

Anttila, Arto, Timothy Dozat, Daniel Galbraith & Naomi Shapiro (2020). Sentence stress in presidential speeches. Kentner, Gerrit & Joost Kremers (eds.), *Prosody in Syntactic Encoding*, Walter de Gruyter, Berlin/Boston, 17–50.

Barr, Dale J., Roger Levy, Christopher Scheepers & H. J. Tily (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68:3, 255–278.

Bates, Douglas, Martin Mächler, Ben Bolker & S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1, 1–48.

Bolinger, Dwight L. (1972). Accent is predictable (if you are a mind reader). *Language* 48, 633–644.

Chen, Danqi & Christopher Manning (2014). A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, 740–750.

Chomsky, Noam & Morris Halle (1968). *The Sound Pattern of English*. Harper & Row, New York.

Cinque, Guglielmo (1993). A null-theory of phrase and compound stress. *Linguistic Inquiry* 24, 239–298.

Cohen Priva, Uriel (2012). *Sign and Signal: Deriving Linguistic Generalizations from Information Utility*. Ph.D. Dissertation, Stanford University.

Cohen Priva, Uriel (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology* 6:2, 243–278.

Dozat, Timothy (2017). MetricalTree. URL `https://github.com/tdozat/Metrics`.

Frick, Robert W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin* 97:3, 412–429.

Godfrey, John, Edward Holliman & Jane McDaniel (1992). SWITCHBOARD: Telephone speech corpus for research and development. *IEEE ICASSP-92*.

Gussenhoven, Carlos (2011). Sentential Prominence in English. van Oostendorp, Marc, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.), *The Blackwell Companion to Phonology*, John Wiley & Sons, Ltd, Oxford, UK, 2778–2806.

Hasegawa-Johnson, Mark, Ken Chen, Jennifer Cole, Sarah Borys, Sung-Suk Kim, Aaron Cohen, Tong Zhang, Jeung-Yoon Choi, Heejin Kim, Taejin Yoon & Sandra Chavarria (2005). Simultaneous recognition of words and prosody in the Boston University radio speech corpus. *Speech Communication* 46:3-4, 418–439.

van Heuven, Vincent J. (2018). Acoustic Correlates and Perceptual Cues of Word and Sentence Stress: Towards a Cross-Linguistic Perspective. van der Hulst, Harry, Jeffrey Heinz & Rob Goedemans (eds.), *The Study of Word Stress and Accent: Theories, Methods and Data*, Cambridge University Press, Cambridge, 15–59.

Klein, Dan & Christopher D. Manning (2003). Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03*, Association for Computational Linguistics, Sapporo, Japan, vol. 1, 423–430.

Ladd, D. Robert & Rachel Morton (1997). The perception of intonational emphasis: continuous or categorical? *Journal of Phonetics* 25:3, 313–342.

Liberman, Mark & Alan Prince (1977). On stress and linguistic rhythm. *Linguistic Inquiry* 8:2, 249–336.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard & David McClosky (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Baltimore, Maryland, 55–60.

Nenkova, Ani, Jason Brenier, Anubha Kothari, Sasha Calhoun, Laura Whitton, David Beaver & Dan Jurafsky (2007). To memorize or to predict: Prominence labeling in conversational speech. *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*.

Ostendorf, Mari, Izhak Shafran, Stefanie Shattuck-Hufnagel, Lesley Carmichael & William Byrne (2001). A prosodically labeled database of spontaneous speech. *Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding* 119–121.

Piantadosi, Steven T., Harry Tily & Edward Gibson (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108:9, 3526–3529.

Prince, Alan (1983). Relating to the grid. *Linguistic Inquiry* 14:1, 19–100.

R Core Team (2020). R: A language and environment for statistical computing. URL `https://www.R-project.org/`.

Shapiro, Naomi (2019). MetricGold. URL `https://github.com/tsnaomi/metric-gold`.

Wagner, Michael & Duane G. Watson (2010). Experimental and theoretical advances in prosody: A review. *Language & Cognitive Processes* 25:7-9, 905–945.