Draft of October 4, 2018, comments welcome

Sentence stress in presidential speeches

Arto Anttila† Timothy Dozat† Daniel Galbraith† Naomi Shapiro‡

†Stanford University ‡University of Washington

Abstract

Sentential prominence is not represented in writing, it is hard to measure phonetically, and it is highly variable, yet it undoubtedly exists. Here we report preliminary findings from our study of sentential prominence in the inaugural addresses of six U.S. presidents. We confirm the familiar hypothesis that sentential prominence has two sources (Jespersen 1920): it is partly MECHANICAL and depends on syntax (Chomsky and Halle 1968, Liberman and Prince 1977, Cinque 1993) and partly MEANINGFUL in that it highlights informative material (Bolinger 1972). Both contribute independently to perceived prominence. Pursuing the view that sentential prominence is a matter of STRESS, we provide evidence for the linguistic reality of the Nuclear Stress Rule (Chomsky and Halle 1968) as well as the view that information coincides with stress peaks in good prose (Bolinger 1957). We also observe that part of speech matters to sentence stress: noun and adjective stresses are loud and mechanical; verb and function word stresses are soft and meaningful. We suggest that this may explain why parts of speech differ in word phonology as well.

1. Introduction

Sentential prominence is a complex and fascinating topic: it is not represented in writing, it is hard to measure by phonometric methods, and it is highly variable, yet it undoubtedly exists.[1] Consider the following excerpt from Jimmy Carter's inaugural address (January 20, 1977, sentence 3), transcribed by a native speaker of English:

(1) We mùst adjùst to chánging TIMES and stìll hóld to unchánging PRINCIPLES.

The transcriber heard four degrees of prominence: primary (capitalized), secondary (acute accent), tertiary (grave accent), and none (unmarked). The same utterance can be heard slightly differently by different speakers. Here's another transcriber:

(2) We mùst adjúst to chànging TIMES and stìll hóld to unchànging príncíples.

[1] This research was partially funded by the Office of the Vice-Provost for Undergraduate Education (VPUE) at Stanford University and by the Roberta Bowman Denning Initiative Committee, H&S Dean's Office, as part of the project *Prose Rhythm and Linguistic Theory* (Arto Anttila, PI). Our work was inspired by the project *The sound of text*, partially funded by the VPUE and carried out in collaboration with Joshua Falk and Ryan Heuser in the summers of 2010-2011. We are deeply indebted to Alex Wade for his help with the dataframe construction. This paper has benefited from presentations at the Stanford Phonology Workshop (March 3, 2017), the 39th Annual Meeting of the Deutsche Gesellschaft für Sprachwissenschaft (DGfS), Workshop on *Prosody in Syntactic Encoding*, Saarland University, Saarbrücken (March 10, 2017), and the Archbishop Mitty High School Linguistics Club (April 6, 2017). We thank the audiences for helpful feedback. In particular, we thank Jared Bernstein, Canaan Breiss, Joan Bresnan, Uriel Cohen Priva, Vivienne Fong, Penny Eckert, Isabelle Franz, Bruce Hayes, Ryan Heuser, Larry Hyman, Gerrit Kentner, Paul Kiparsky, Mark Liberman, Geoffrey Pullum, Nathan Sanders, Madeline Snigaroff, Benjamin Storme, Meghan Sumner, and Simon Todd. Finally, we thank the Department of Statistics Consulting Services at Stanford University for their advice. We are responsible for any errors.

The two transcribers differ by one step on the words *adjust*, *changing*, *unchanging*, and *principles*. However, the essential shape of the perceived prominence contour remains the same. This becomes evident if we convert the accents to numbers, marking higher level of prominence with a higher number (primary = 3, secondary = 2, tertiary = 1, none = 0). Anticipating our conclusion, we will call such prominences STRESSES.

(3) (a) 0 1 1 0 2 3 0 1 2 0 2 3 transcriber (1)
(b) 0 1 2 0 1 3 0 1 2 0 1 2 transcriber (2)

Stress is a matter of relative, not absolute prominence, and numerical differences can be irrelevant. Consider the opening phrase *My fellow citizens* from Barack Obama's first inaugural address (January 20, 2009, sentence 1) transcribed 1 2 3 by the first transcriber and 0 1 2 by the second. The numbers differ for each word, but both transcribers heard the same steadily rising contour.

Why are sentences stressed the way they are? Linguistic tradition holds that there are two kinds of sentence stress (Jespersen 1920: 212-222). First, there is MEANINGFUL stress, sometimes called "semantic" or "psychological" stress (Jespersen's *Wertdruck*, *Neuheitsdruck*, *Gegensatzdruck*). Meaningful stress is illustrated in (4):

(4) In the Gilmore Girls universe, Luke and Lorelai seemed inevitable.
He served the coffee; she needed the coffee. (Correction: NEEDED the coffee.) [2]

Under one natural reading, *needed* is contrastively stressed (as opposed to *served*) and *NEEDED* receives an additional dose of emphatic stress as indicated by the capitalization in the original. Both stresses are individually meaningful.

There is also MECHANICAL STRESS, sometimes called "physiological" stress (Jespersen's *rhythmischer Druck*, *Einheitsdruck*). Mechanical stress is illustrated in (5) with an example from Ladd 1996: 166:

(5) Q: How much did they pay you for participating in the experiment?
A: Five francs.

In the answer, "*francs* is almost entirely predictable if the conversation takes place in a country where the unit of currency is the franc; *five* is the information of interest. Yet the accent is on *francs*." (Ladd 1996: 166). The stress on *francs* is thus not meaningful, but assigned mechanically to the rightmost content word of the sentence.

We first show that both kinds of stresses are real. The evidence is consistent with the view that sentence stress is assigned based on syntax (Chomsky and Halle 1968, Liberman and Prince 1977, Cinque 1993) and that information seeks out stress peaks, especially in good prose (Bolinger 1957). We then present evidence that sentence stress impacts different parts of speech differently: noun and adjective stresses are loud and mechanical; verb and function word stresses are soft and meaningful. We suggest that this may explain why parts of speech differ in word phonology as well (Smith 2011).

[2] *Entertainment Weekly*, November 17, 2016, http://www.ew.com/article/2016/11/17/gilmore-girls-luke-originally-woman

Our assumption that sentential prominence is a matter of stress is controversial and requires a comment. Theories of sentential prominence differ on two key questions:

(6) (a) Are sentential prominences metrical (stresses) or tonal (pitch accents)?
(b) Do sentential prominences reflect syntax directly or indirectly?

In this paper, we take the view that sentential prominence is a metrical phenomenon: the prominences are genuine stresses, parallel to word stresses (see, e.g., Hayes 1995, Ch. 2). We further assume that sentence stress is assigned directly to words based on the surface syntactic structure. This view may strike some readers as quaint and others as wrong. It most closely resembles the SPE theory of phrasal stress (Chomsky and Halle 1968) and its immediate descendants, such as Liberman and Prince 1977 and Cinque 1993. It is therefore good to briefly motivate our theoretical choices before turning to the empirical part of the study.

Our first assumption, that sentential prominence is a matter of stress, implies that degrees of prominence involve degrees of stress: some words are stronger than others and some words are weaker than others. This is the Infinite Stress View (ISV), so named by Gussenhoven (2011: 2779), that is rejected by many of today's researchers, including Gussenhoven (1991, 2011, 2015), who instead advocates the Pitch Accent View (PAV), under which "nuclear stress" is a pitch accent, i.e., a tone or a tone complex, left behind when its neighbors are deaccented. Crucially, pitch accent removal is categorical and there is no notion of gradient prominence, implying that apparent primary, secondary, tertiary, etc., stresses are not representable at the sentence level, and any impressions to the contrary must have some other explanation. Gussenhoven further notes that "[t]oday, there are probably no linguists who adhere to the ISV in its original form" (Gussenhoven 2011: 2779). Our analysis builds on the ISV and provides new evidence for it.

Our second assumption, that sentence stress is assigned to words based on surface syntax, as in, e.g., Kaisse 1985, glosses over three decades of work on the prosodic hierarchy; for overviews, see, e.g., Inkelas and Zec 1990, 1995; Truckenbrodt 2007; Selkirk 1995, 2011; and Nespor and Vogel 2007; among others. These theories assume that there exists a hierarchical prosodic structure – essentially an imperfect phonological reflection of syntax – that plays a role in phrasing and prominence. In our analysis, the prosodic hierarchy plays no role.

Why go back to these earlier theories of sentence stress? Our reasons were first and foremost practical. Over the past decades the computational infrastructure for linguistics has taken enormous strides forward. It is now easy to take a large amount of text, parse it syntactically, apply sentence stress rules to the resulting trees, and compare the predicted stress contours to human judgments. This seemed like an interesting project, especially as it had to the best of our knowledge never been attempted before. The Nuclear Stress Rule (NSR) and Compound Stress Rule (CSR) of Chomsky and Halle (1968) are simple rules: given the syntax, implementing them is a programming task. Pitch accent theories (e.g., Gussenhoven 1983; Ladd 1996) are less straightforward to test, mainly because they often rely on notions like accentuation domains that are harder to operationalize. Similar reasons kept us from adopting the prosodic hierarchy. We also note that the prosodic hierarchy is most commonly employed to provide domains for external sandhi, not for assigning stress, and for that reason it seemed less relevant for the task at hand.[3]

Practical reasons aside, systematically testing a theoretical claim, even an incorrect one, is a useful mode of investigation. Applying a theory to a significant amount of data often brings up unexpected evidence that speaks to important theoretical questions. Indeed, we believe that our results provide a new argument for the gradience of sentence stress, and hence for the Infinite Stress View, against the Pitch Accent View.

The paper is structured as follows. Section 2 lays out the empirical procedure and explains our stress model. Section 3 shows that both meaningful and mechanical stress matter to perceived stress. Section 4 discusses stress differences among parts of speech. Section 5 concludes the paper.

2. Procedure

Why are sentences stressed the way they are? We approached this question by taking the following steps:

(7) (a) Find a scripted speech, with a transcription, audio, and video.
(b) Assign mechanical stress to the text by a computer.
(c) Annotate the text for informativity by a computer.
(d) Collect perceived stress judgments from native speakers.
(e) Try to predict perceived stress from mechanical stress and informativity.

Our data consist of the first inaugural addresses of six presidents: Carter (1977), Reagan (1981), Bush Sr., (1989), Clinton (1993), Bush Jr. (2001), and Obama (2009), available from the American Presidency Project (Peters and Woolley 1999-2018) in script, audio, and video. The work involved annotating the speeches for mechanical stress, informativity, and perceived stress. For mechanical stress, we used the METRICALTREE software written by Timothy Dozat that implements a version of Liberman and Prince's (1977) stress algorithm in conjunction with a syntactic analysis by the Stanford Parser (Klein and Manning 2003, Chen and Manning 2014, Manning et al. 2014).[4] For informativity, we used bigram informativity (e.g., Pan and Hirschberg 2000, Piantadosi et al. 2011, Cohen Priva 2012, 2015). Both aspects of the analysis will be explained below. Finally, the perceived stress judgments were collected with the web application METRICGOLD developed by Naomi Shapiro concurrently with the annotation process.[5] The result was a spreadsheet of about 11,500 words annotated for several syntactic and phonological variables, crucially mechanical stress, informativity, and perceived stress. We then used statistical tools to evaluate to what extent perceived stress is determined by phonology, syntax, and informativity.

Inaugural addresses are a very specific genre: the delivery is slow and fluent, most likely well rehearsed, and the text has benefited from the skills of professional

[3] Bellik, Bellik, and Kalivoda (2017) develop software for generating and evaluating prosodic hierarchy candidate sets based on syntactic trees using violable constraints. Their experiments illustrate that small differences in constraint definitions can have significant consequences for the analysis.
[4] The source code for MetricalTree is freely available at https://github.com/tdozat/Metrics.
[5] The source code for MetricGold is freely available at https://github.com/tsnaomi/metric-gold.

speechwriters. However, we have no particular reason to believe that the underlying stress contours would be any different from those of ordinary English. Rather, it seems that oratorical prose maximizes the use of natural prosodic resources in ordinary speech, highlighting properties that are hard to detect in rapid conversation. Like the speech of radio announcers, which is characterized by "natural but controlled style, combining the advantages of both read speech and spontaneous speech" (Hasegawa-Johnson et al. 2005), oratorical prose turns out to be a rich source of evidence for the study of sentence prosody.

2.1 The mechanical stress model

In classical generative phonology (Chomsky and Halle 1968, henceforth *SPE*), stress is a feature that takes numerical values: [1 stress], [2 stress],…, [*n* stress]. Content words start out with [1 stress] assigned by word stress rules, while function words have no word stress. Sentence stress rules apply cyclically to syntactic constituents, starting from the innermost constituent, assigning [1 stress] to a designated word and reducing stress elsewhere by one (stress subordination). The sentence stress rules are stated below:

(8) **The Nuclear Stress Rule (NSR):** Assign [1 stress] to the rightmost vowel bearing the feature [1 stress]. Applies to phrases (NP, VP, AP, S).

(9) **The Compound Stress Rule (CSR):** Skip over the rightmost word and assign [1 stress] to the rightmost remaining [1 stress] vowel; if there is no [1 stress] to the left of the rightmost word, then try again without skipping the word. Applies to words (N, A, V).

The NSR and CSR are illustrated below by the famous sentence *John's blackboard eraser was stolen* (Chomsky and Halle 1968: 15-24). The parentheses indicate syntactic constituents. Each cycle is represented as a row. The outcome is the stress contour 3 2 5 4 1. Compared to the perceived stress numbers in our earlier examples, the predicted stress numbers are inverted, with the consequence that they are readily translatable into ordinary English: 1 = primary stress, 2 = secondary stress, 3 = tertiary stress, etc.

(10) [[[John's] [[[black] [board]] [eraser]]] [was stolen]]

1	1	1	1	1	
	[1	2]			Cycle 1
	[1	3	2]		Cycle 2
[2	1	4	3]		Cycle 3
[3	2	5	4	1]	Cycle 4

In Liberman and Prince's (1977) revision of the SPE stress theory, the two stress rules are defined on local syntactic trees as follows: in a configuration [A B], if the constituent is a phrase, B is strong (= NSR); if the constituent is a word, B is strong iff it branches (= CSR).[6] Our program METRICALTREE essentially implements Liberman and Prince's

[6] Under Cinque's (1993) reformulation, stress falls on the most deeply embedded phrase, i.e., to the right of V in VO languages and to the left of V in OV languages. More generally, complements win over heads and specifiers, and in the absence of complements, heads wins over specifiers. Cinque's proposal is an

sentence stress rules building on the phrase structures provided by the Stanford Parser (Klein and Manning 2003, Chen and Manning 2014, Manning et al. 2014), with a number of modifications to be discussed shortly.

In the SPE theory, words come labeled as either stressed (content words) or unstressed (function words). Our mechanical stress model adopts a more fine-grained three-way taxonomy: words can be stressed, unstressed, or stress-ambiguous. For earlier proposals along the same lines, see, e.g., Hirschberg 1993 and Shih 2014. We considered all words lexically stressed except for those listed in (11). These special words were identified as unstressed or stress-ambiguous by three sometimes overlapping criteria: (a) word form, (b) part of speech, and (c) syntactic dependency. This taxonomy is a working hypothesis that we expect to revisit in future work (Snigaroff 2017). For the meaning of the part of speech and dependency labels, see Appendix A.

(11) Lexically unstressed and stress-ambiguous words (first approximation)

	UNSTRESSED	STRESS-AMBIGUOUS
(a)	it	this, that, these, those
(b)	CC, PRP$, TO, UH, DT	MD, IN, PRP, WP$, PDT, WDT, WP, WRB
(c)	det, expl, cc, mark	cop, neg, aux, auxpass

Positing a stress-ambiguous category reflects our uncertainty about the presence vs. absence of lexical stress on some function words. However, it also serves a deeper purpose in making the realistic prediction that sentence stress is variable. Consider the following sentence from Ronald Reagan's first inaugural address (January 20, 1981, sentence 63):

(12) I do not believe in a fate that will fall on us no matter what we do.

By our taxonomy, this sentence has nine stress-ambiguous words. Since the sentence stress rules presuppose that words are either lexically stressed or unstressed, such words must be first disambiguated before the sentence stress rules can apply. Assuming that the actual stress values of ambiguous words are independent of each other, the sentence has $2^9 = 512$ stress paths, each one potentially resulting in a distinct sentence stress contour.[7] Instead of examining the 512 readings individually we opted for three basic stress models described in (13).

(13) Model 1: Ambiguous words are stressed.
Model 2: Ambiguous monosyllables are unstressed, polysyllables stressed.
Model 3: Ambiguous words are unstressed.

In addition, we constructed an "ensemble model" that takes the mean of the three basic models. Our model thus diverges from SPE in permitting lexical stress ambiguity. As a result, we have the beginnings of an explanation for why sentence stress is variable. The

improvement in terms of cross-linguistic coverage, but does not affect our discussion as we are only concerned with English at the moment.

[7] The unambiguously stressed words are *believe, fate, fall, on, matter, do*; the unambiguously unstressed words are *a, no*. All other words are stress-ambiguous.

stress rules (NSR, CSR) are invariant; variation comes from lexical stress ambiguity in certain common function words.

2.2 Information-theoretic variables

According to the information-theoretic view, sentence accent is a matter of information, not of structure: what is informative is accented, what is uninformative is unaccented.[8] This is the gist of the Focus-to-Accent theory of sentence accent (Gussenhoven 1983) famously championed by Bolinger and illustrated in the following quote:

> "In phrases like *bóoks to write*, *wórk to do*, *clóthes to wear*, *fóod to eat*, *léssons to learn*, *gróceries to get* - as they occur in most contexts - the verb is highly predictable: food is to eat, clothes are to wear, work is to do, lessons are to learn. Less predictable verbs are less likely to be de-accented - where one has *léssons to learn*, one will probably have *pàssages to mémorize*." (Bolinger 1972: 634)

Information content has been found to be a good predictor of perceived prominence in earlier work, see, e.g., Pan and McKeown 1999. Cole et al. (2010: 435) call this EXPECTATION-DRIVEN PROMINENCE: "[T]he listener may judge prominence based on information status alone, rather than judging the acoustic form directly. In this sense, the listener's judgment of word prominence is driven by their expectation based on prior experience of the word." According to Pan and Hirschberg (2000: 239-240), "of all the collocation measures we investigated, bigram word predictability has the strongest correlation with pitch accent assignment."

We operationalized a word's information content using its BIGRAM INFORMATIVITY, the word's average predictability across the entire corpus of inaugural addresses beginning with Roosevelt 1933. This informativity measure is an extension of conditional probability. Conditional probability is a measure of local predictability, in that it estimates the probability of an event based on its immediate context. In particular, the conditional probability of a word *w* is the probability of seeing *w* in a specific context or corpus *c*:

(14) $P(w|c)$

For instance, if we take Barack Obama's 2013 address as our scope, we might be interested in the probability of seeing the word *Americans* in his address:

(15) $P(Americans|\text{Barack Obama's 2013 address})$

We approximate (15) by limiting the scope of the context to the tokens directly preceding *w*, such as estimating the probability of seeing *Americans* following the word *fellow*. This probability is calculated by taking the number of times we see the phrase *fellow Americans* in Barack Obama's address and dividing it by the number of times we see the word *fellow*.

[8] The terminological shift from "sentence stress" to "sentence accent" reflects research tradition. In this paper, we assume that sentence-level prominences are genuine stresses, analogous to word stresses, accompanied by an overlay of pitch accents (see, e.g., Hayes 1995, Ch. 2).

Specifically, this produces a bigram probability, since *fellow Americans* is two words in length.

(16) $P(Americans|fellow) = \frac{\text{count}(\textit{fellow Americans})}{\text{count}(fellow)}$

The informativity of a word *w* is the average predictability of *w* across contexts (cf. Piantadosi et al. 2011, Cohen Priva 2012, 2015). It is measured as the weighted average of the negative log probability of seeing *w* given every context *c* that *w* follows in a corpus. These probabilites are weighted by the conditional probabilities of seeing *c* given *w*. Let *C* be the set of contexts in which *w* appears:

(17) $\text{Informativity}(w) = -\sum_{c \in C} P(c|w)\ \log_2 P(w|c)$

Tying this to our previous example, the word *Americans* appears 4 times in Barack Obama's 2013 address: twice after the word *fellow*, once after *of*, and once after *as*. In addition, *fellow* appears 3 times in total, *of* 69 times, and *as* 12 times. Therefore, the bigram informativity of the word *Americans* in the 2013 speech is 2.7159, as shown in (18). Likewise, the bigram informativity of *Americans* across the entire inaugural corpus is 5.4126.

(18) $\text{Informativity}(Americans) = -\frac{2}{4}\log_2\frac{2}{3} - \frac{1}{4}\log_2\frac{1}{69} - \frac{1}{4}\log_2\frac{1}{12} = 2.7159$

Intuitively, words vary in their informativity. For instance, we might expect prepositions to be generally less informative than nouns and verbs. Consider the informativity of the words *Americans*, *confronting*, and *of*, given the entire inaugural corpus: [9]

(19) $\text{Informativity}(of) = 1.7024$
$\text{Informativity}(Americans) = 5.4126$
$\text{Informativity}(confronting) = 6.0549$

In Bolinger's view, meaningful stress is about informativity. How are mechanical stress and meaningful stress related? Bolinger proposes that there is a natural alignment between the two:

> "The recipe for reconciling the two functions is simple: the writer should make them coincide as nearly as he can by maneuvering the semantic heavy stress into the position of the mechanical loud stress; that is, toward the end." (Bolinger 1957: 235)

Bolinger then makes the interesting suggestion that choices among syntactic variants, such as actives and passives, may be driven by sentence stress:

[9] An anonymous reviewer asks why we calculated bigram informativity based on the inaugural corpus instead of larger corpora. We suspect that the vocabulary of inaugural addresses may be rather specific to this genre, so using the same genre for calculating frequency-based measures seems appropriate.

> "To circumvent the arbitrariness of grammar, the writer may now choose between parallel structures differing only in sentence order. He picks the one that allows him to get his stress at the end. If he defines Canada as *the place where Canadian bacon was invented*, he miscues his reader, for he wants the stress on *bacon* (and can easily put it there if he says it aloud), but his order suggests that it falls on *invented*. A parallel structure, *the place where they invented Canadian bacon*, avoids the trap." (Bolinger 1957: 236)

In other words, meaningful stress and mechanical stress preferably coincide: information seeks out stress peaks, especially in good prose. We call this hypothesis STRESS-INFORMATION ALIGNMENT; for relevant discussion, see, e.g., Calhoun 2010 and Cohen Priva 2012. Note that Bolinger says "writer", not "speaker", so one would expect his theory to hold particularly well in a scripted genre like the inaugural speech corpus. A plausible alternative to Stress-Information Alignment is COMMUNICATIVE DYNAMISM, a hypothesis that goes back to the Prague School (see, e.g., Firbas 1971), according to which information typically piles up towards the end of the sentence. We will return to both theories shortly.[10]

2.3 Collecting perceived stress judgments

We define perceived stress as syllable prominence intuitively felt by a native speaker. This view is succinctly stated in the following quote:

> "what makes a syllable accented is for the large part the work of the perceiver, generating his internal accent pattern on the basis of a strategy by which he assigns structures to the utterances. These structures, however, are not fabrications of the mind only, for they can be related to sound cues." (van Katwijk 1974: 5, cited in Baart 1987: 4)

Native speakers usually find it easy to tap, hum, exaggerate, or otherwise highlight the "rhythm" of a sentence they hear. Our goal was to make maximal use of this native speaker ability to interpret prosody. In interpreting prominence contours, a native speaker is in the position to draw upon rich resources, not only the familiar objective phonetic cues, such as pitch, amplitude, and duration that are in principle available to anyone, but also the speaker's overall knowledge of the language. One way to make such intuitions explicit is through gestures, for example, by tapping out the rhythm of the sentence. In fact, the presidents themselves often involuntarily annotate the text they are delivering by head nods and hand gestures. In our annotation practice, we considered the cues in roughly the following order. For a more detailed explanation, see Appendix B.

(20) (a) native speaker intuitions
(b) embodied cues, e.g., tapping (annotator), nodding (president)
(c) explicit phonetic cues

[10] Yet another hypothesis is UNIFORM INFORMATION DENSITY (Levy and Jaeger 2007; Jaeger 2010), which proposes, roughly, that information strives to be evenly spread across the time it takes to utter a sentence. At this level of generality, this theory flatly contradicts the other two.

A key question is the choice of a transcription system. The alternatives include the popular ToBI (Silverman et al. 1992, Veilleux et al. 2006), RaP (Rhytm and Pitch, Dilley and Brown 2005), and RPT (Rapid Prosody Transcription, Cole et al. 2017). The matter is effectively decided by our hypothesis: since our goal is to verify the NSR/CSR, which predicts gradient stress contours, the optimal annotation system must allow for similar gradience. ToBI only allows for the binary prominence distinction pitch-accented vs. unaccented and does not describe metrical prominence; RaP allows for the three-way distinction non-prominent, prominent but not pitch-accented, and prominent and pitch-accented (Breen et al. 2012: 284); finally, RPT uses the binary distinction prominent vs. not-prominent. The binarity assumption seems normal in computational work (see, e.g., Pan et al. 2002, Nenkova et al. 2007).

To achieve the level of granularity appropriate for verifying the NSR/CSR we chose to transcribe the data directly in terms of metrical grids (Prince 1983). This involves taking an utterance of a suitable length and marking the words with grid columns of different heights that translate into whole numbers ranging from 0 (non-prominent) to 8 (highly prominent). In practice, the annotators limited themselves to a scale from 0 to 6. The annotation method is described in more detail in Appendix B. Stress judgments were collected from two native speaker annotators using the web application METRICGOLD developed concurrently with the annotation process with the help of the annotators. The annotators were native speakers of two varieties of English: Irish and American (West). Both had completed coursework in phonology and took part in weekly project meetings, but did the actual annotation work independently of each other. The meetings took place over the summer of 2016 and resulted in the informal protocol summarized in Appendix B. The resulting data frame contains approximately 11,500 words coded for perceived stress as well as for various phonological, syntactic, and frequency variables.

Any transcription is to some extent subjective and there is always variation across transcribers. However, in our view, such variation is not noise, but data. This is not simply making a virtue out of necessity, but has a serious scientific rationale. Assuming that prominence is largely the work of the perceiver, it is possible that two native speakers with different linguistic experience genuinely differ in their interpretation of the same prominence contour.[11] Variation in the stress annotation thus does not necessarily mean that at least one of the annotators must be wrong. It may mean that the annotators have subtly different grammars and consequently interpret the same signal differently. Moreover, eliminating variation across annotators by "harmonizing" the discrepancies is a procedure fraught with difficulties (see, e.g., Ernestus and Baayen 2011). For these reasons, we took no explicit steps to eliminate variation from our stress transcriptions, beyond the loose guidelines outlined above. Even with this freedom, the interannotator reliability turned out good (Cronbach's alpha = 0.85).

3. Both mechanical and meaningful stress matter

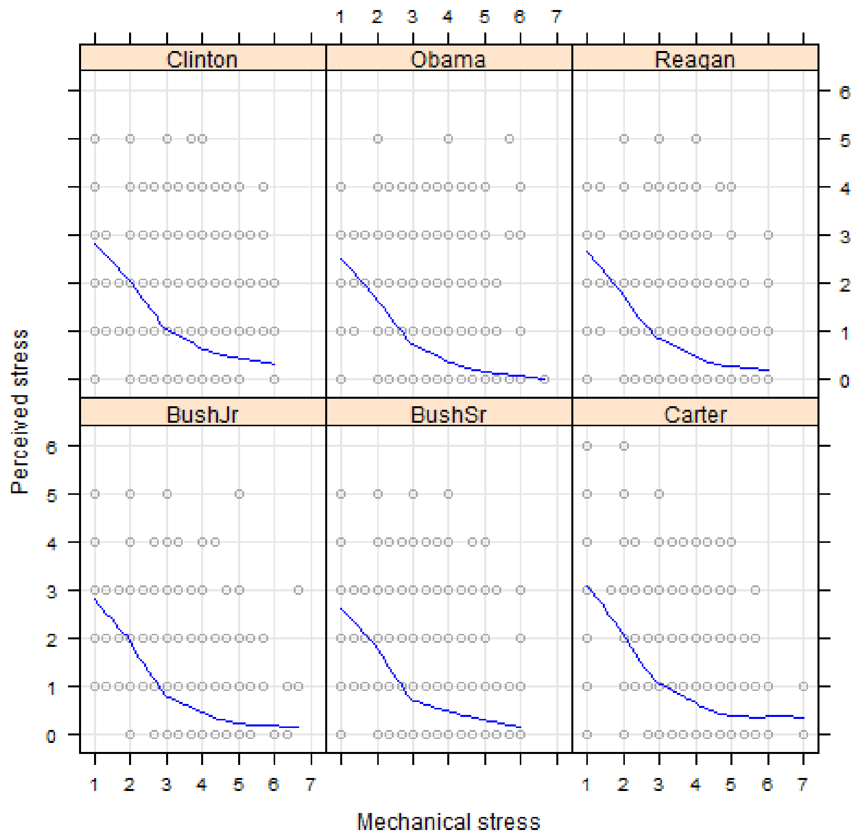
The goal of this section is twofold. First, we visualize the key relationships between perceived stress and other variables to help the reader intuitively appreciate the systematic patterns in the data. Based on earlier linguistic literature we have a fairly good sense of

[11] This opens up the interesting possibility of using the stress annotation to study, not just the grammar of the president being transcribed, but also the grammar of the annotator doing the transcription.

what to expect and the first order of business is to confirm these expectations. Second, we present regression models that support the composite nature of perceived stress: both meaningful stress and mechanical stress are real.

We start by focusing on four major syntactic categories, namely nouns, adjectives, verbs, and function words, trimming the data down from 11,641 to 10,982 words. The obvious first thing to look at is the relationship between mechanical and perceived stress. The scatterplots in (21) visualize the relationship between the "ensemble model", i.e., the mean of the three basic mechanical stress models, and perceived stress, with both annotators included.[12]

(21) Mechanical stress vs. preceived stress by president



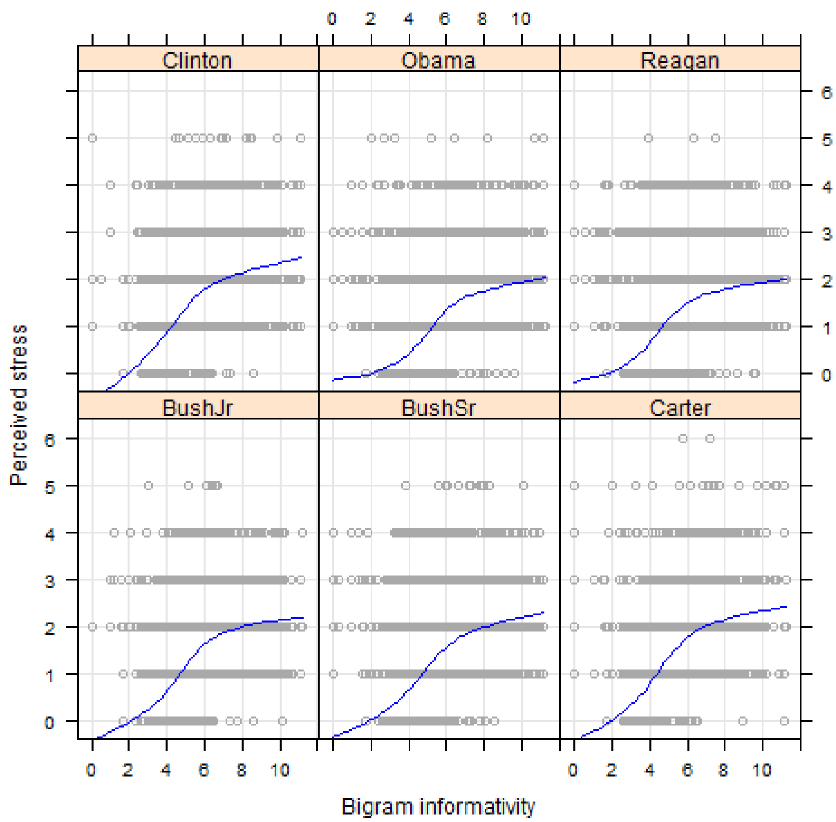
Recall that mechanical stress follows the SPE convention: 1 = primary stress, 2 = secondary stress, etc., where a lower number stands for more stress. Perceived stress inverts this order: 0 = no stress, 1 = one degree of stress, 2 = two degrees of stress, etc., where a lower number stands for less stress. We would thus expect a negative correlation between mechanical and perceived stress and this is indeed what we see (Spearman's $\rho = -0.4957$, $p < 2.2\text{e-}16$). One is immediately struck by the shape of the curve: the relationship is nearly one-to-one at the loud end of the spectrum where primary stress is heard as 3, secondary stress as 2, and tertiary stress as 1, but as soon as we go below tertiary stress the graph starts to flatten out. This is what we might expect: the theory predicts quaternary, quinary, senary, and

[12] These scatterplots were drawn using the xylowess.fnc convenience function from the languageR package with the default smoother added (Baayen 2008, 2013).

septenary stresses with no problem, but to the human ear such distinctions become increasingly hard to hear.

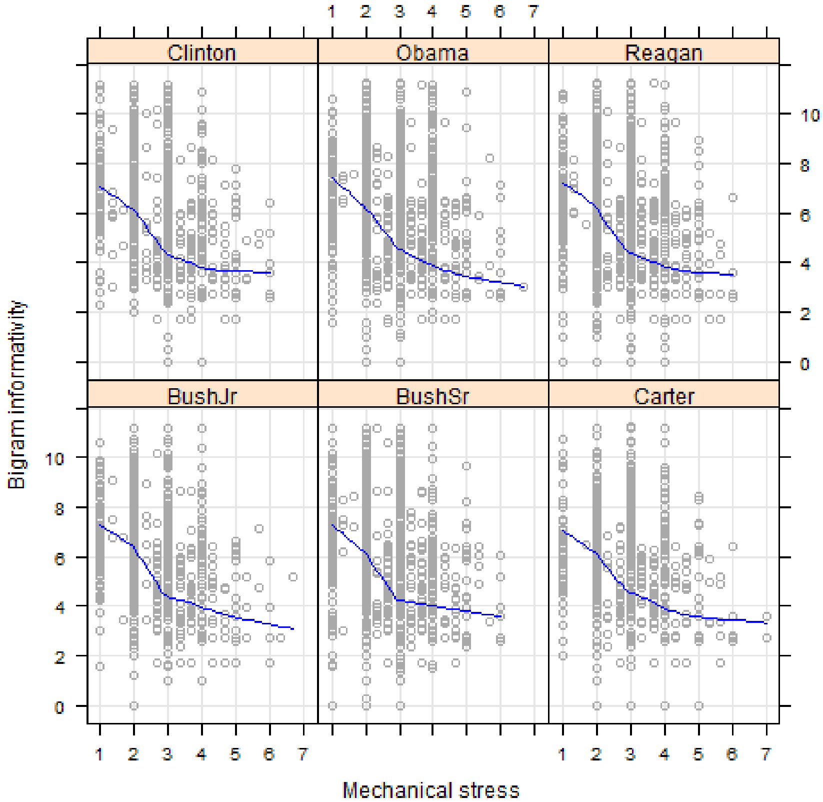
The obvious next thing is to examine the relationship between informativity and perceived stress. One would expect more informative words to have more stress, i.e., the correlation should be positive. Again, this turns out to be so ($\rho = 0.6034$, $p < 2.2e\text{-}16$). The scatterplots in (22) visualize the relationship between bigram informativity and perceived stress, with both annotators included.

(22) Bigram informativity vs. perceived stress by president



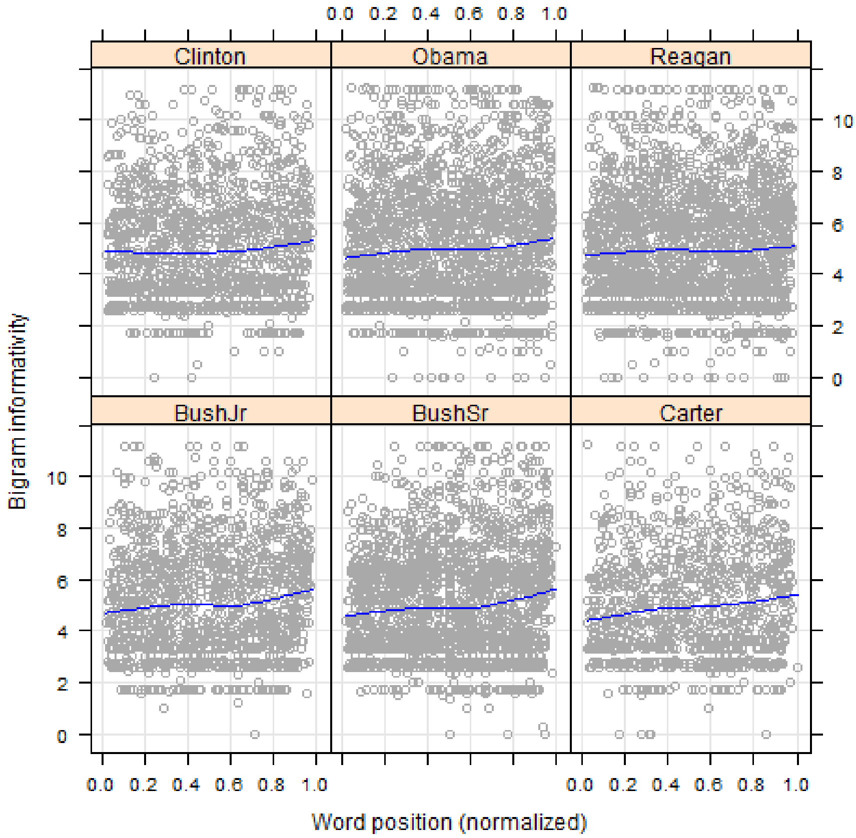
We now turn to Stress-Information Alignment. Recall Bolinger's (1957) suggestion that meaningfulness and mechanical stress ideally coincide in good prose. Let us start by checking whether bigram informativity and mechanical stress are correlated. One would expect a negative correlation because our mechanical stress model follows the SPE numerology: the larger the stress number, the weaker the stress. Indeed, we find a strong negative correlation ($\rho = -0.4341$, $p < 2.2e\text{-}16$), which is consistent with Bolinger's (1957) hypothesis.

(23) Mechanical stress vs. bigram informativity by president



An alternative to Bolinger's Stress-Information Alignment hypothesis is Communicative Dynamism which predicts that information should pile up towards the end of the sentence. The scatterplots in (24) visualize the relationship between a word's linear position in the sentence and bigram informativity. Since sentences vary greatly in length we normalized word position by sentence length, so that every word falls on a scale between 0 (beginning) and 1 (end). Communicative Dynamism predicts that we should find a positive correlation and that is indeed what we do find ($\rho = 0.0621$, $p < 2.2e\text{-}16$). However, this correlation is much weaker than the correlation between mechanical stress and bigram informativity pictured above.

(24) Word position (normalized) vs. bigram informativity by president



Do all three predictors (bigram informativity, mechanical stress, word position) independently contribute to perceived stress? In order to find out, we fitted a linear regression model to the data. The model is summarized in (25). The response variable is perceived stress on a scale from 0 to 6, log transformed.[13]

(25) A linear regression model of perceived stress (logged)

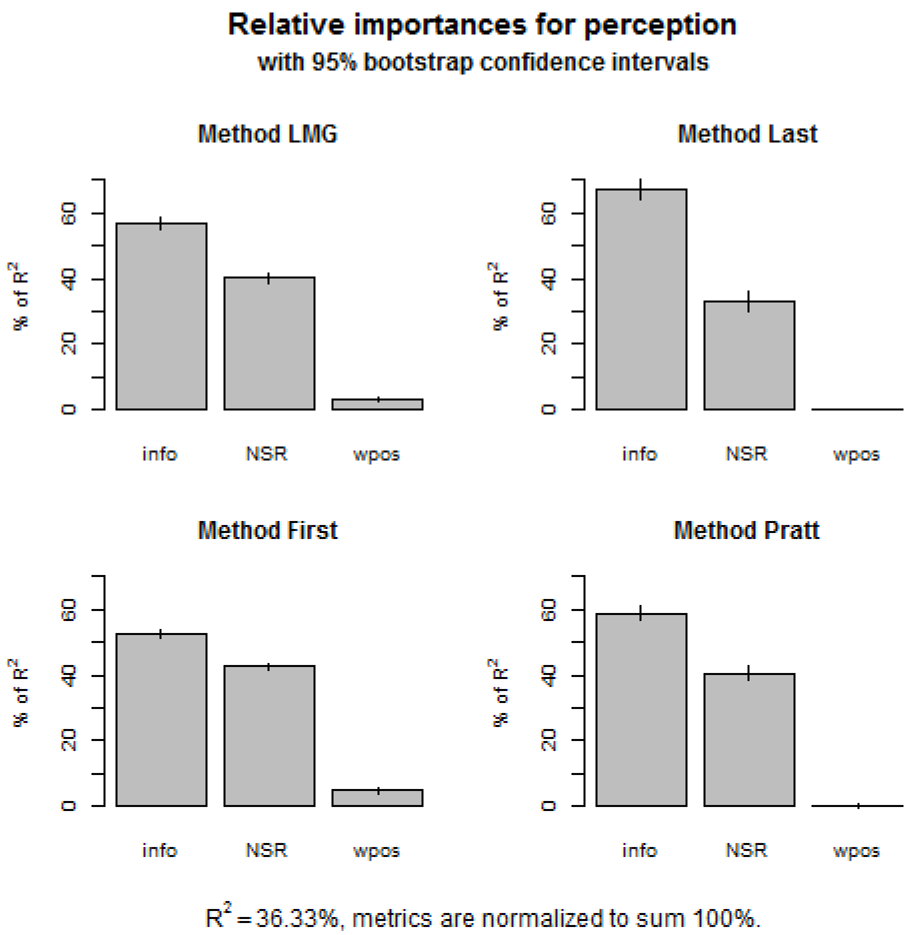
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.163686	0.039481	29.47	<2e-16 ***
bigram informativity	0.215886	0.003171	68.07	<2e-16 ***
mechanical stress	-0.366282	0.007644	-47.92	<2e-16 ***
word position	0.032694	0.025341	1.29	0.197

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.9417 on 21901 degrees of freedom
(59 observations deleted due to missingness)
Multiple R-squared: 0.3633, Adjusted R-squared: 0.3632
F-statistic: 4165 on 3 and 21901 DF, p-value: < 2.2e-16

[13] The 59 observations deleted due to missingness are of two kinds: (i) the contractions *‘ll, ‘m, ‘re, ‘s, ‘ve* were not assigned a mechanical stress value by MetricalTree; (ii) one of the annotators declined to give a perceived stress value for the word *4-year*.

Only bigram informativity and mechanical stress come out significant, both in the expected direction. In order to get a sense of each predictor's relative importance we calculated the contribution of each predictor using the relaimpo package (Grömping 2006). All four methods point at bigram informativity (info) as being the most important predictor, followed by mechanical stress (NSR), followed by the word's normalized position in the sentence (wpos).

(26) The relative importance of predictors for perceived stress (logged)



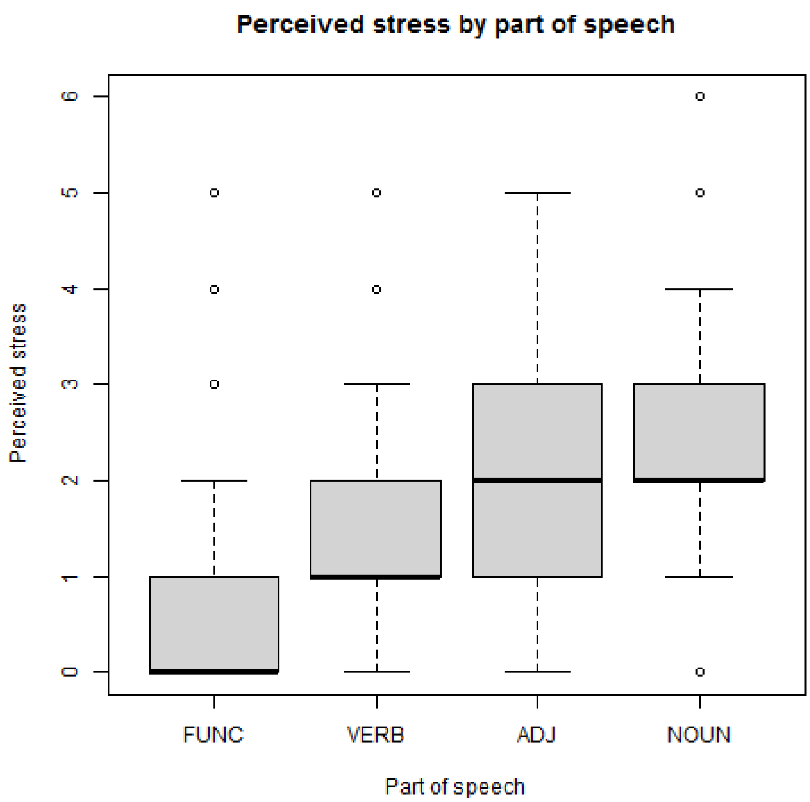
We conclude that both meaningful stress and mechanical stress are real and both contribute to perceived stress. In particular, the evidence supports the reality of the Nuclear Stress Rule (Chomsky, Halle, and Lukoff 1956; Chomsky and Halle 1968; Liberman and Prince 1977; Cinque 1993) as an independent source of stress.

There remain a number of additional structural predictors that one can expect to influence stress perception. In the next section we will consider one such predictor: PART OF SPEECH. While the general picture will not change with the addition of the part-of-speech variable, the linguistic plot will thicken.

4. Part of speech effects

It is well known that parts of speech differ in degree of stress (see, e.g., Altenberg 1987, Hirschberg 1993, German et al. 2006). The boxplot in (27) shows the level of perceived stress by part of speech, with both annotators included. The bottom and top of the box are the 25th and the 75th percentiles; the black band is the median.

(27) Perceived stress vs. part of speech



What we see here is a stress hierarchy FUNC < VERB < ADJ < NOUN. Function words are soft, nouns are loud. Adding part of speech into our model shows that NOUN is a good predictor of perceived stress just as FUNC and VERB are good predictors of its absence.

(28) A linear regression model of perceived stress (logged)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.335652	0.046119	28.961	< 2e-16 ***
bigram informativity	0.123341	0.003421	36.052	< 2e-16 ***
mechanical stress	-0.125354	0.008629	-14.527	< 2e-16 ***
category FUNC	-0.976762	0.026459	-36.916	< 2e-16 ***
category NOUN	0.198743	0.024628	8.070	7.41e-16 ***
category VERB	-0.403530	0.025801	-15.640	< 2e-16 ***
word position	0.196662	0.024064	8.173	3.18e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8734 on 21898 degrees of freedom
(59 observations deleted due to missingness)
Multiple R-squared: 0.4523, Adjusted R-squared: 0.4521
F-statistic: 3014 on 6 and 21898 DF, p-value: < 2.2e-16

All the predictors (bigram informativity, mechanical stress, part of speech, word position) are independently significant in the expected direction. The dependence of stress on syntactic category has been stated in various ways by earlier researchers: arguments are more "accentable" than predicates (Schmerling 1976; Ladd 1996: 187-193) and fine-grained language-specific "accentability hierarchies" have been proposed; for a summary, see Baart 1987: 56-57. Pan and Hirschberg (2000: 237) note that "[i]n general, nouns, especially head nouns, are very likely to be accented", while Pan et al. (2002) report that "[v]erbs, which are content words, are not accented, according to the POS-pitch accent model."

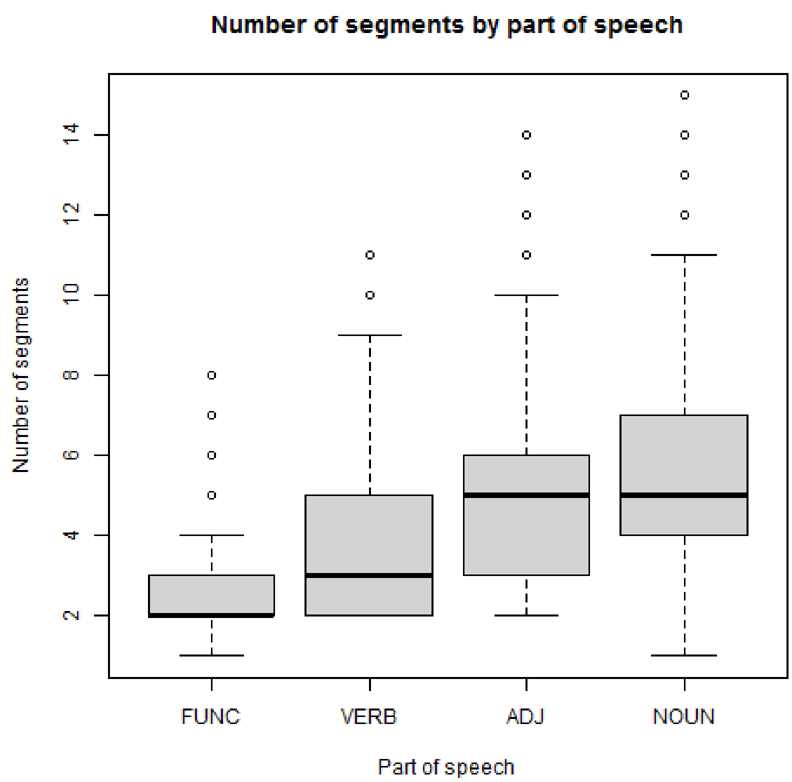
The stress hierarchy FUNC < VERB < ADJ < NOUN bears an uncanny resemblance to hierarchies independently discovered in word phonology. In many languages, verbs, adjectives, and nouns exhibit different phonological behavior. In an overview article, Smith (2011: 2439) employs the term "*phonological privilege* […] understood to mean the ability to support a greater array of phonological contrasts, whether this is manifested as a larger number of underlying distinctions, more variety in surface patterns, or a greater resistance to assimilation or other phonological processes." Intuitively, nouns are phonologically more resilient than adjectives which are more resilient than verbs. Smith concludes with the following statement:

> "Many, although perhaps not all, cases are consistent with a universal scale of phonological privilege, N > A > V. Furthermore, the overwhelming majority of cases involve prosodic and suprasegmental phenomena […]" (Smith 2011: 2459).

Could the similarity between "accentability" in sentence phonology and "privilege" in word phonology be an accident? That seems unlikely. We would like to put forward the speculative hypothesis that word-level privilege scales reflect sentence-level accentability scales, synchronically or diachronically. This is based on the following reasoning. For syntactic reasons, nuclear stress typically falls on nouns because they are arguments and avoids verbs because they are predicates. We also know that stress inhibits phonetic reduction. Therefore, words that are typically stressed, i.e., nouns, become resistant to reduction and are better able to support contrasts, whereas words that are typically unstressed, i.e., verbs, are more susceptible to reduction and tend to lose contrasts. In this way, word phonology ultimately depends on sentence phonology. Note that this explanation makes sense only if sentential prominence is stress: stress-based segmental effects are common, whereas tone-based segmental effects are virtually unknown (see, e.g., Hyman and Schuh 1974: 108). This view is also in general agreement with the position of Kelly and Bock (1988) and Kentner (2012) who argue based on experimental evidence that word stress can be sensitive to sentence stress.

Smith's privilege scale is strikingly manifested in word length. The boxplot in (29) visualizes word length by part of speech in the inaugural address corpus. We measured word length in terms of the number of phonological segments based on the CMU pronouncing dictionary (Weide 1998).

(29) Number of segments vs. part of speech



What we see here is a length hierarchy FUNC < VERB < ADJ < NOUN. Function words are short, nouns are long. The obvious parallel between the stress hierarchy in (27) and the length hierarchy in (29) strongly suggests that the two are related. An anonymous reviewer points out that the NSR was developed for English and might apply fairly straightforwardly to other SVO languages with similar prosodic systems. One such language for which we have suitable corpora available is Finnish, an SVO language with a phrasal stress system much like in English, although different in terms of intonation (Sadeniemi 1949: 78-100, Iivonen et al. 1987: 219-250, Vilkuna 1998). In the *Aamulehti 1999* newspaper corpus

(Aamulehti 1999) the mean lengths of verbs, adjectives, and nouns are neatly arranged along Smith's privilege hierarchy: verbs are shortest, nouns are longest. It remains to be seen to what extent the parallel holds up cross-linguistically. An anonymous reviewer asks what the predictions are for languages without lexical stress, tone languages, verb-final languages, and languages without clear sentence stress. Languages without lexical stress cannot have an NSR that refers to word stress, but they might well have other ways to classify words for the purposes of phrasal stress. Tone languages could have an NSR to the extent they also have lexical stress (Hyman 2006). Verb-final languages with lexical stress, such as German, seem to behave largely as predicted by the NSR suitably generalized, see, e.g., Cinque 1993, Wiese 1996, Sect. 8.5.1, and Wagner 2005, and one would thus expect a similar word length effect. As for languages without clear sentence stress, one would have to know the details.

The evidence suggests that a word's typical sentence prosody is somehow crystallized in its lexical representation.[14] This might explain the success of ACCENT RATIO as a measure of accentability (Nenkova et al. 2007). The intuition is that people simply memorize how likely a word is to be accented and do this separately for each word. Nenkova et al. (2007) define accent ratio as equal to the estimated probability of a word being accented if this probability is significantly different from 0.5 and equal to 0.5 otherwise. This predictor alone turns out to perform about as well as five other predictors together (unigram and bigram probability, part of speech, word length, and word position in the utterance) and generalizes well across genres. One predictor not included in the Nenkova et al. (2007) models is the NSR/CSR. It would be interesting to see to what extent the NSR/CSR can explain accent ratio, but we will leave that for future work.

Our final model of perceived stress that includes word length as a predictor is summarized in (30). This is a linear mixed model that uses the R packages lme4 (Bates et al. 2014) and lmerTest (Kuznetsova et al. 2016). The dependent variable is perceived stress (logged); the fixed effects are bigram informativity, mechanical stress, word length (logged), part of speech, and word position. President and annotator were included in the model as random effects. All the fixed effects are independently highly significant in the expected direction.

[14] There is segmental evidence that nouns are more exhaustively footed than verbs in Finnish (Anttila 2006). Adams (2014) argues that low-frequency adjectives are more completely prosodified than high-frequency adjectives in English.

(30) A mixed effects linear regression model of perceived stress (logged)

Formula: perception ~ info + NSR + nseg.log + category + wpos +
(1 | president) + (1 | annotator)

Data: presidents.data.words.navf.final

Random effects:

Groups	Name	Variance	Std.Dev.
president	(Intercept)	0.01199	0.1095
annotator	(Intercept)	0.04476	0.2116
Residual		0.71199	0.8438

Number of obs: 21830, groups: president, 6; annotator, 2

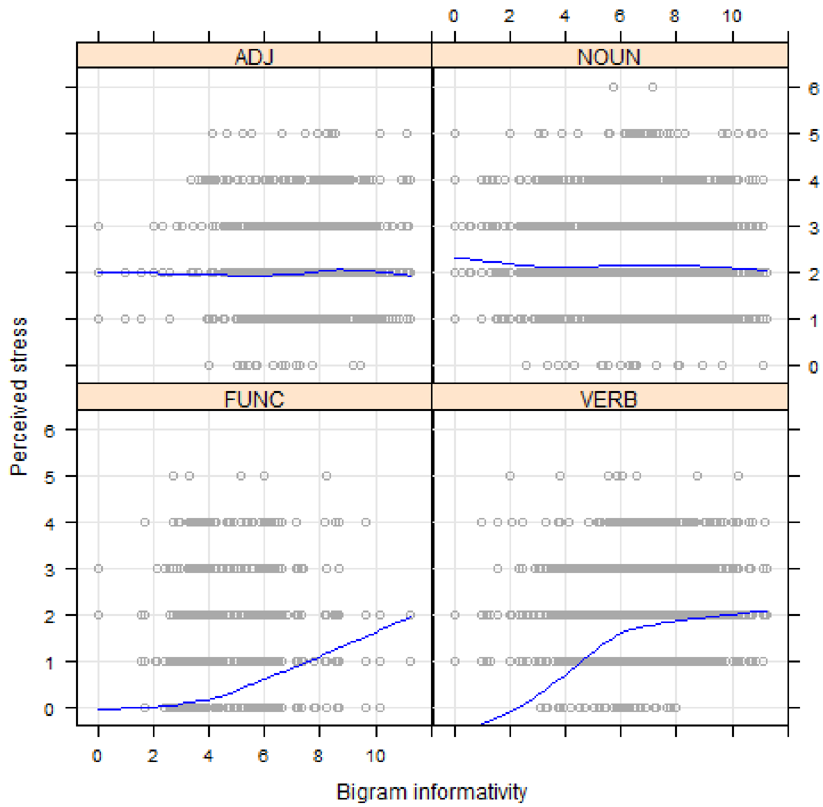
Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	7.926e-01	1.640e-01	1.000e+00	4.833	0.0737	.
bigram informativity (info)	1.056e-01	3.451e-03	2.182e+04	30.594	< 2e-16	***
mechanical stress (NSR)	-8.785e-02	8.502e-03	2.182e+04	-10.333	< 2e-16	***
word length (nseg.log)	3.676e-01	1.578e-02	2.182e+04	23.304	< 2e-16	***
category FUNC	-8.012e-01	2.664e-02	2.182e+04	-30.069	< 2e-16	***
category NOUN	1.705e-01	2.403e-02	2.182e+04	7.096	1.33e-12	***
category VERB	-3.003e-01	2.541e-02	2.182e+04	-11.817	< 2e-16	***
word position (wpos)	2.246e-01	2.332e-02	2.182e+04	9.629	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

We conclude with a final remark on parts of speech. Our final regression model shows that bigram informativity matters to perceived stress, but interestingly, this effect seems to be entirely driven by verbs and function words. The scatterplots in (31) show the level of perceived stress as a function of bigram informativity, with the data grouped by part of speech.

(31) Perceived stress vs. bigram informativity by part of speech



A clear two-way split among parts of speech is immediately apparent. There is no correlation between perceived stress and bigram informativity in nouns (ρ = −0.0116, *p* = 0.3851) or adjectives (ρ = 0.0415, *p* = 0.08522), but a strong positive correlation emerges in verbs (ρ = 0.4810, *p* < 2.2e-16) and function words (ρ = 0.3810, *p* = < 2.2e-16). We can summarize this finding as follows: noun and adjective stresses are loud and mechanical; verb and function word stresses are soft and meaningful.[15]

These observations are consistent with Bolinger's remark about verbs. Recall that Bolinger argues that stress differences are a matter of informativity, in particular "[l]ess predictable verbs are less likely to be de-accented - where one has *léssons to learn*, one will probably have *pàssages to mémorize*" (Bolinger 1972: 634). Some of Bolinger's key examples are repeated in (32).

(32) (a) I have léssons to learn.
I have pàssages to mémorize.
(b) I have a póint to make.
I have a pòint to émphasize.

Although the facts are beyond dispute, our results suggest an alternative explanation. A fact not commented on by Bolinger is that *memorize* and *emphasize* are simply longer

[15] A related observation is made by Cole et al. (2010: 438) who note that frequency matters for prominence in function words, but less in content words: "for most content words factors other than word frequency play a larger role in prominence perception than they do for function words."

words than *learn* and *make*. Our final regression model shows that word length matters to perceived stress. Could it be that the stress Bolinger is hearing depends, not on informativity, but on word length? Since the two are correlated it may be hard to tell, but the explanations are entirely different and the possibility of a length effect cannot be ruled out. We also note that the correlation between perceived stress and word length is strongest precisely in verbs ($\rho = 0.4957$, $p < 2.2\text{e-}16$), where length is measured in terms of the number of segments (logged). The upshot is that a simple phonological variable such as word length may partially undermine explanations based on informativity.

5. Summary

The results presented here are preliminary. One area where improvement is clearly needed is our mechanical stress model, in particular our lexical stress taxonomy which is a first approximation. An initial study has been conducted to evaluate the quality of the model (Snigaroff 2017), but the results were not used in the present paper.

We offer three main results. First, we have seen new evidence that syntax plays an important role in sentence accentuation. While earlier studies have approximated mechanical stress in various ways (see, e.g., Calhoun 2010), we explicitly combined stress rules with syntax and compared the predicted stress contours to the perceived stress contours. Second, we observed that noun and adjective stresses are loud and mechanical whereas verb and function word stresses are soft and meaningful. The gradience that emerges from the transcriptions suggests that binary classification schemes are insufficient, posing a problem for a pure pitch accent view, which assumes that the grid, i.e., a gradient representation of sentential prominence, does not exist (see, e.g., Gussenhoven 2015). Third, we speculated that sentence stress may explain part-of-speech effects in word phonology. For example, nouns tend to be resistant to reduction and better able to support lexical contrasts than verbs because nouns are typically sustained by higher levels of sentence stress. If that is correct, word phonology is not a self-contained system and cannot be completely understood without first understanding sentence phonology.

Appendix A: Lexical stress annotation

Our mechanical stress model relies on the output of the Stanford Parser (Klein and Manning 2003, Chen and Manning 2014, Manning et al. 2014). The Stanford Dependency Parser (Chen and Manning 2014) follows the Stanford Dependency labeling scheme, where the label a word is assigned is dictated both by the lexical properties of the word and by the role that word plays in the sentence. In this Appendix we describe the labels used by the metrical parser in addition to some prototypical examples of their intended usage. While there is some overlap between the part-of-speech and dependency labels, we found both types of annotations were needed for best performance.

We marked words as unstressed and stress-ambiguous based on the part-of-speech and dependency labels listed below. The following words override part-of-speech tags and dependency labels: *it* (unstressed), *this, that, these, those* (stress-ambiguous). In addition to labels and informal descriptions, we include representative examples. For the part-of-speech tags we also give the number of matching words in our corpus of 11,641 words. All other words were labeled as lexically stressed.

(33) Unstressed part-of-speech labels

CC	coordinating conjunction (*and*, *but*, *nor*, *or*, *so*, *yet*)	639
PRP$	possessive pronoun (*my, your, his, her, its, their*)	445
TO	*to*	352
UH	interjection (*yes*, *well*, *amen*)	7
DT	determiner (*a, all, an, another, any, both, each, every, many, neither, no, some, that, the, these, this, those*)	1,174

(34) Stress-ambiguous part-of-speech labels

MD	modal (*can, could, may, might, must, shall, will, would, need, should, 'll , can't*)	276
IN	preposition, subordinating conjunction (*about, above, across, after, against, alongside, among, around, as, at, because, before, beneath, between, beyond, but, by, during, except, for, from, if, in, into, lest, like, of, off, on, out, outside, over, since, so, than, that, though, through, throughout, 'til, toward, under, unless, until, up, upon, whether, while, with, within, without*)	1,309
PRP	personal pronoun (*he, himself, I, it, itself, me, myself, one, ours, ourselves, she, thee, them, themselves, they, us, we, you, your*)	762
PDT	predeterminer (*all*, *such*)	10
WDT	wh-determiner (*that*, *which*, *whatever*)	78
WP	wh-pronoun (*what, who, whom, whoever*)	94
WRB	wh-adverb (*how, when, where, why, whenever, wherever*)	60

(35) Unstressed dependency labels

det	determiner (*a, an, the, this, that, these, those, every, each*)
expl	expletive (*there*)
cc	coordinating conjunction (*and*, *or*, *but*, *yet*, *as*, *nor*)
mark	complementizer or similar introducing a subordinate clause or phrase (*that, if, whether, to, because, in, of, than, as, for, upon, though, while, since*)

(36) Stress-ambiguous dependency labels

cop	copula (*be, am, is, are, were, was, been, being*)
neg	negation (*no, not, never, n't*)
aux	auxiliary (*have, has, must, will, can, might, may, do, does, did, should, is, are, were*)
auxpass	passive auxiliary (*been, is, be, was, were, been, are, being*)

Appendix B

The annotators were instructed as follows: "If intuitions are crystal clear, no further confirmation is necessary and the annotator should feel free to move on to the next sentence. If the intuitions are not so clear, there are various ways to make them more explicit: (i) Replay the utterance multiple times in quick succession to hear and feel its rhythm; (ii) While listening, tap out the utterance rhythm. Most people find it easier to tap on stressed syllables than on unstressed syllables; (iii) Humming along, i.e., substituting a syllable like *ma* for each syllable of the utterance, may also help one. This is called "reiterant speech" (Liberman and Prince 1977: 250); (iv) Take note of visual cues in the video, such as the speaker's head nods and hand gestures. If none of the above methods helps, the annotator should pay explicit attention to the linguistic cues in the signal. Cues for presence of stress: an abrupt pitch movement on a syllable (rise, fall); longer vowel duration (*this is soooo difficult*); strong aspiration (*This is p[hhh]retty difficult*); selection of the stressed allomorph (*a*/*ey*), *the*/*thee*); etc. Cues for absence of stress: vowel reduction (schwa); auxiliary contraction (e.g., *have* → *'ve*); etc."

Data sources

Aamulehti 1999. An electronic document collection of the Finnish language containing 16.6 million words. Gatherers: The Department of General Linguistics, University of Helsinki; The University of Eastern Finland; CSC - IT Center for Science Ltd. Available through the Language Bank of Finland, http://www.kielipankki.fi/. URN: http://urn.fi/urn:nbn:fi:lb-201403268.

Peters, Gerhard and John T. Woolley. 1999-2018. The American Presidency Project, http://www.presidency.ucsb.edu/index.php.

References

Adams, Matthew E. 2014. *The Comparative Grammaticality of the English Comparative*. Ph.D. dissertation, Stanford University.

Altenberg, Bengt. 1987. *Prosodic Patterns in Spoken English: Studies in the Correlation Between Prosody and Grammar for Text-to-Speech Conversion*. Vol. 76. Studentlitteratur, Lund.

Anttila, Arto. 2006. Variation and opacity. *Natural Language & Linguistic Theory* 24(4), 893–944.

Baart, Joan. 1987. *Focus, Syntax, and Accent Placement*. Ph.D. dissertation, University of Leiden.

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge.

Baayen, R. H. 2013. languageR: Data sets and functions with *Analyzing Linguistic Data: A practical introduction to statistics*, R package version 1.4.1. http://CRAN.R-project.org/package=languageR.

Bates, Douglas, Martin Maechler, Ben Bolker and Steven Walker. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-6. http://CRAN.R-project.org/package=lme4

Bellik, Jenny, Ozan Bellik, and Nick Kalivoda. 2017. SPOT. JavaScript application. https://github.com/syntax-prosody-ot.

Bolinger, Dwight L. 1957. Maneuvering for Stress and Intonation. *College Composition and Communication*, 8(4), 234-238.

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

Breen, Mara, Laura C. Dilley, John Kraemer, and Edward Gibson. 2012. Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory* 8(2), 277 – 312.

Calhoun, Sasha. 2010. How does informativeness affect prosodic prominence? *Language and Cognitive Processes* 25(7-9), 1099-1140.

Chen, Danqi and Christopher Manning. 2014. A Fast and Accurate Dependency Parser Using Neural Networks. In *Proceedings of EMNLP* 2014.

Chomsky, Noam, Morris Halle and Fred Lukoff. 1956. On accent and juncture in English. In Morris Halle et al. (eds.), *For Roman Jakobson: Essays on the occasion of his sixtieth birthday*, Mouton & Co., The Hague, pp. 65-80.

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. Harper and 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.

Cole, Jennifer, Timothy Mahrt, and Joseph Roy. 2017. Crowd-sourcing prosodic annotation. *Computer Speech & Language* 45, 300-325.

Cole, Jennifer, Yoonsook Mo, and Mark Hasegawa-Johnson. 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology* 1(2), 425-452.

Dilley, Laura C., and Meredith Brown. 2005. The RaP (Rhythm and Pitch) Labeling System. Unpublished manuscript, MIT, available at http://web.mit.edu/tedlab/tedlab_website/RaP%20System/RaP_Labeling_Guide_v1.0.pdf

Ernestus, Miriam and R. Harald Baayen. 2011. Corpora and exemplars in phonology. In John A. Goldsmith, Jason Riggle, and Alan Yu (Eds.), *The Handbook of Phonological Theory*, Second Edition, Wiley Blackwell, pp. 374-400.

Firbas, Jan. 1971. On the concept of communicative dynamism in the theory of functional sentence perspective. *Sbornik Prací Filosofické Fakulty Brnénské University* (Studia Minora Facultatis Philosophicae Universitatis Brunensis) A-19. 135-144.

German, James, Janet Pierrehumbert, and Stefan Kaufmann. 2006. Evidence for phonological constraints on nuclear accent placement. *Language* 82(1), 151-168.

Grömping, Ulrike. 2006. Relative importance for linear regression in R: the package relaimpo. *Journal of statistical software* 17(1), 1-27.

Gussenhoven, Carlos. 1983. Focus, mode and the nucleus. *Journal of linguistics* 19(2), 377-417.

Gussenhoven, Carlos. 1991. The English Rhythm Rule as an accent assignment rule. *Phonology* 8, 1-35.

Gussenhoven, Carlos. 2011. Sentential prominence in English. *The Blackwell Companion to Phonology*, Vol. 2, 2778-2806.

Gussenhoven, Carlos. 2015. Does phonological prominence exist? *Lingue e linguaggio* 14(1), 7-24.

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.

Hayes, Bruce. 1995. *Metrical Stress Theory: Principles and Case Studies*. Chicago University Press, Chicago.

Hirschberg, Julia. 1993. Pitch accent in context: predicting intonational prominence from text. *Artificial Intelligence* 63, 305-340.

Hyman, Larry M. 2006. Word-prosodic typology. *Phonology* 23(2), 225-257.

Hyman, Larry M., and Russell G. Schuh. 1974. Universals of tone rules: evidence from West Africa. *Linguistic Inquiry* 5(1), 81-115.

Iivonen, Antti, Terttu Nevalainen, Reijo Aulanko, Hannu Kaskinen. 1987. *Puheen intonaatio* [*Speech Intonation*]. Gaudeamus, Helsinki.

Inkelas, Sharon and Draga Zec, ed. 1990. *The Phonology–Syntax Connection*. University of Chicago Press, Chicago/London.

Inkelas, Sharon and Draga Zec. 1995. Syntax–phonology interface. In John A. Goldsmith (ed.), *The Handbook of Phonological Theory*, Blackwell, Cambridge, MA, pp. 535-49.

Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology* 61(1), 23-62.

Jespersen, Otto. 1920. *Lehrbuch der Phonetik: Mit 2 Tafeln*. BG Teubner.

Kaisse Ellen M. 1985. *Connected Speech: The Interaction of Syntax and Phonology*. Academic Press, Orlando, FL.

Kelly, Michael H., and J. Kathryn Bock. 1988. Stress in time. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3), 389-403.

Kentner, Gerrit. 2012. Linguistic rhythm guides parsing decisions in written sentence comprehension. *Cognition* 123(1), 1-20.

Klein, Dan and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Kuznetsova, Alexandra, Per Bruun Brockhoff and Rune Haubo Bojesen Christensen. 2016. lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-30. http://CRAN.R-project.org/package=lmerTest

Ladd, D. Robert. 1996. *Intonational Phonology*. Cambridge University Press.

Levy Roger P. and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Proceedings of the 20th Conference on Advances in Neural Information Processing Systems* (*NIPS 2007*), ed. J. C. Platt, D. Koller, Y. Singer, S. T. Roweis, Curran Assoc., Red Hook, N.Y., pp. 849-56.

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

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Nenkova, Ani, Jason Brenier, Anubha Kothari, Sasha Calhoun, Laura Whitton, David Beaver, and 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*, http://www.aclweb.org/anthology/N07-1002. Also posted at ScholarlyCommons http://repository.upenn.edu/cis_papers/732.

Nespor, Marina, and Irene Vogel. 2007. *Prosodic Phonology: With a new foreword*. Studies in Generative Grammar, Vol. 28. Mouton de Gruyter, Berlin/New York.

Pan, Shimei, and Julia Hirschberg. 2000. Modeling local context for pitch accent prediction. In *Proceedings of the 38th Annual Meeting on Association for Computational linguistics*, Association for Computational Linguistics, pp. 233-240.

Pan, Shimei, and Kathleen R. McKeown. 1999. Word informativeness and automatic pitch accent modeling. In 1999 *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, available at http://www.aclweb.org/anthology/W99-0619.

Pan, Shimei, Kathleen McKeown, and Julia Hirschberg. 2002. Exploring features from natural language generation for prosody modeling. *Computer Speech & Language* 16(3-4), 457-490.

Piantadosi, Steven T., Harry Tily, and 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.

Sadeniemi, Matti. 1949. *Metriikkamme perusteet* [*Fundamentals of Finnish Metrics*]. Suomalaisen Kirjallisuuden Seura, Helsinki.

Schmerling, Susan F. 1976. *Aspects of English Sentence Stress*. Texas University Press, Austin.

Selkirk Elisabeth O. 1995. Sentence prosody: intonation, stress, and phrasing. In John A. Goldsmith (ed.), *The Handbook of Phonological Theory*, Blackwell, Cambridge, MA, pp. 550-69.

Selkirk Elisabeth O. 2011. The syntax–phonology interface. In In John A. Goldsmith, Jason Riggle, and Alan Yu (eds.), *The Handbook of Phonological Theory, 2nd Ed.* Blackwell, Malden, MA, pp. 435-83.

Shih, Stephanie. 2014. *Towards Optimal Rhythm*. Ph.D. dissertation, Stanford University.

Silverman, Kim, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. ToBI: A standard for labeling English prosody. In *Second international conference on spoken language processing*, ICSLP-1992, 867-870.

Smith, Jennifer L. 2011. Category-specific effects. *The Blackwell Companion to Phonology* 4, 2439-2463.

Snigaroff, Madeline K. 2017. Evaluating MetricalTree, Ms., Stanford University.

Truckenbrodt, Hubert. 2007. The syntax–phonology interface. In Paul de Lacy (ed.), *The Cambridge Handbook of Phonology*, Cambridge University Press, Cambridge, UK, pp. 435-56.

Veilleux, Nanette, Stefanie Shattuck-Hufnagel, Alejna Brugos. 2006. 6.911 Transcribing Prosodic Structure of Spoken Utterances with ToBI. January IAP 2006. Massachusetts Institute of Technology: MIT Open Course Ware, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

Vilkuna, Maria. 1998. Word order in European Uralic. In Anna Siewierska (ed.), *Constituent Order in the Languages of Europe*, Empirical approaches to language typology 20, No. 1, Mouton de Gruyter, Berlin and New York, pp. 173-233.

Wagner, Michael. 2005. Asymmetries in Prosodic Domain Formation, in Norvin Richards and Martha McGinnis (eds.), *Perspectives on Phases*, MITWPL 49, 329-367.

Weide, Robert. 1998. The Carnegie Mellon Pronouncing Dictionary [cmudict.0.6], ftp://ftp.cs.cmu.edu/project/speech/dict/, Carnegie Mellon University.

Wiese, Richard. 1996. *The Phonology of German*. Oxford University Press, Oxford.