Syntactic surprisal affects spoken word duration in conversational contexts

Vera Demberg, Asad B. Sayeed, Philip J. Gorinski, and Nikolaos Engonopoulos
M2CI Cluster of Excellence and
Department of Computational Linguistics and Phonetics
Saarland University
66143 Saarbrücken, Germany
{vera,asayeed,philipg,nikolaos}@coli.uni-saarland.de

Abstract

We present results of a novel experiment to investigate speech production in conversational data that links speech rate to information density. We provide the first evidence for an association between syntactic surprisal and word duration in recorded speech. Using the AMI corpus which contains transcriptions of focus group meetings with precise word durations, we show that word durations correlate with syntactic surprisal estimated from the incremental Roark parser over and above simpler measures, such as word duration estimated from a state-of-the-art text-to-speech system and word frequencies, and that the syntactic surprisal estimates are better predictors of word durations than a simpler version of surprisal based on trigram probabilities. This result supports the uniform information density (UID) hypothesis and points a way to more realistic artificial speech generation.

1 Introduction

The uniform information density (UID) hypothesis suggests that speakers try to distribute information uniformly across their utterances (Frank and Jaeger, 2008). Information density can be measured in terms of the surprisal incurred at each word, where surprisal is defined as the negative log-probability of an event. This paper sets out to test whether UID holds across different linguistic levels, i.e. whether speakers adapt word duration during production to syntactic surprisal, such that words with higher surprisal have longer durations than words with lower surprisal. We investigate this question in a corpus of transcribed speech from a mix of native and non-native English speakers, a population that is a non-trivial component of the user base for language technologies developed for English. This data reflects a casual, uncontrolled conversational environment.

Using linear mixed-effects modeling, we found that syntactic surprisal as calculated from a top-down incremental PCFG parser accounts for a significant amount of variation in spoken word duration, using an HMM-trained text-to-speech system as a baseline. The findings of this paper provide additional support the uniform information density hypothesis and furthermore have implications for the design of text-to-speech systems, which currently do not take into account higher-level linguistic information such as syntactic surprisal (or even word frequencies) for their word duration models.

1.1 Related work

The use of word-level surprisal as a predictor of processing difficulty is based on the notion that processing difficulty results when a word is encountered that is unexpected given its preceding context. The amount of surprisal on a word w_i can be formalized as the log of the inverse conditional probability of w_i given the preceding words in the sentence $w_1 \dots w_{i-1}$, or $-\log P(w_i|w_{1\dots i-1})$. If this probability is low, then the word is unexpected, and surprisal is high. Surprisal can be estimated in different ways, e.g. from word sequences (n-grams) or with respect to the possible syntactic structures covering a sentence prefix (see Section 4).

Hale (2001) showed that surprisal calculated from a probabilistic Earley parser correctly predicts well-

known processing phenomena that were believed to emerge from structural ambiguities (e.g., garden paths) and Levy (2008) further demonstrated the relevance of surprisal to human sentence processing difficulty on a range of syntactic processing difficulty phenomena.

There is existing work in correlating information-theoretic measures of linguistic redundancy to the observed duration of speech units. Aylett and Turk (2006) demonstrate that the contextual predictability of a syllable (n-gram log probability) has an inverse relationship to syllable duration in speech. Their experiments were performed using a carefully articulated speech synthesis training corpus.

This type of work fits into a larger programme of understanding how speakers schedule utterances to avoid high variation in the transmission of linguistic information over time, also known as the Uniform Information Density (UID) hypothesis (Florian Jaeger, 2010). Levy and Jaeger (2007) show that the reduction of optional *that*-complementizers in English is related to trigram surprisal; low surprisal predicts a high likelihood of reduction. Florian Jaeger (2010) shows the same result of increased reduction when the complementizer is more predictable according to information density calculated in terms of the main verb's subcategorization frequency.

Frank and Jaeger (2008) provide evidence that a UID account can predict the use of reduced forms of "be", "have", and "not" in English. They use the surprisal of the candidate word itself as well as surprisals of the word before and after, computing bigram and trigram estimates directly from the corpus without smoothing or backoff.

Jurafsky et al. (2001) report a corpus study similar to ours, showing that words that are more predictable from context are reduced. As measures of word predictability, they use bigram and trigram models, as well as joint probabilities, but not syntactic surprisal.

Within the same theme of utterance duration vs. information content, Piantadosi et al. (2011) performed a study using Google-derived n-gram datasets on the lexica of multiple languages, including English, Portuguese, and Czech. For every word in a given language's lexicon, they calculated 2-, 3-, and 4-gram surprisal values using the Google dataset for every occurrence of the word, and then they took the mean surprisal for that word over all occurrences. The 3-gram surprisal values in particular were a better predictor of orthographic length than unigram frequency, providing evidence for the use of information content and contextual predictability as improvement over a Zipf's Law view of communicative efficiency. This is an n-gram approach to supporting the UID hypothesis.

However, there is some counter-evidence for the UID-based view. Kuperman et al. (2007) analyzed the relationship between linguistic unit predictability and syllable duration in read-aloud speech in Dutch. Dutch makes use of interfix morphemes *-s-* and *-e(n)-* in certain contexts to make compound nouns, preferring a null interfix in most cases. For example, the Dutch noun *kandidaatsexamen* ("Bachelor's examination") is composed of *kandidaat-*, *-s-*, and *-examen*.

Kuperman et al. find that the greater the predictability of the interfix from the morphological context (i.e., the surrounding members of the compound), the *longer* the duration of the pronunciation of the interfix. To illustrate, if *-s-* is more expected after *kandidaat* or if *kandidaatsexamen* is a frequent compound, we would therefore expect the *-s-* to be pronounced longer, given the correlations they found. Their finding runs counter to a strong view of UID's fine-grained control over speech rate, but it is focused on the morphological level. They hypothesize that this counter-intuitive result may be driven by complex paradigmatic constraints in the choice of morpheme.

Our work, however, focuses on the syntactic level rather than the paradigmatic. What we seek to answer in our work is the extent to which an information density-based analysis can not only be applied to real speech data in context but also be derived from higher-level syntactic analyses, a combination hitherto little explored. Existing broad-coverage work on syntactic surprisal has largely focused on comprehension phenomena, such as Demberg and Keller (2008), Roark et al. (2009), and Frank (2010). We provide a production study in a vein similar to that of Kuperman et al., but show that frequency effects work in the expected direction at the syntactic level. This in turn expands upon the view supported by n-gram-based work such as that

of Piantadosi et al. (2011); Levy and Jaeger (2007); Jurafsky et al. (2001), showing that information content above the n-gram level is important in guiding spoken language production in humans.

1.2 Implications for Potential Applications

Spoken dialogue systems are of increasing economic and technological importance in recent times, particularly as it is now feasible to include this technology in everything from small consumer devices to industrial equipment. With this increase in importance, there is also unsurprisingly growing scientific emphasis in understanding its usability and safety characteristics. Recent work (Fang et al., 2009; Taube-Schiff and Segalowitz, 2005) has shown that linguistic information presentation has an effect on user behaviour, but the overall granularity of this behaviour is still not well-understood.

Other potential applications exist in any place where text-to-speech technologies can be applied, such as in real-time spoken machine translation and communications systems for the disabled.

In demonstrating that we can observe speakers behaving in the manner predicted by the UID hypothesis in conversational contexts, we provide evidence for a finer-level of granularity necessary for controlling the rate of information presentation in artificial systems.

1.3 AMI corpus

The Augmented Multi-Party Interaction (AMI) corpus is a collection of recorded, transcribed conversations spanning 100 hours of simulated meetings. The corpus contains a number of data streams including speech, video, and whiteboard writing. Transcription of the meetings was performed manually, and the transcripts contain word-level time bounds that were produced by an automatic speech recognition system.

The freely-available AMI corpus is one of a very small number of efforts that contain orthographic transcriptions that are time-aligned at a word level. We chose it for the realism of the setting in which it was recorded; the physical presence of multiple speakers in an unstructured discussion reflects a potentially high level of noise in which we would be looking for surprisal correspondences, potentially increasing the application value of the correspondences we find.

1.4 Organization

The remainder of this paper proceeds as follows. In section 2, we describe at a high level the procedure we used to test our hypothesis that parser-derived surprisal values can partly account for utterance-duration variation. Then (section 3.2) we discuss the MARY text-to-speech system, from which we derive "canonical" word utterance durations. We describe the way we process and filter the AMI meeting corpus in section 3.1. In section 4, we describe in detail our predictors, frequency counts, trigram surprisal, and Roark parser surprisal. Sections 5 and 6 describe how we use linear mixed effects modeling to find significant correlations between our predictors and the response variable, and we finally make some concluding remarks in section 7.

2 Design

The overall design of our experiment is schematically depicted in Figure 1. We extract the words and the word-by-word timings from the AMI corpus, keeping track of each word's position in the corpus by conversation ID, speaker turn, and chronological order. As we describe in the next section, we filter the words for anomalies.

After pre-processing, for each word in the corpus, we extract the following predictors: canonical speech durations from the MARY text-to-speech system, logarithmic word frequencies, n-gram surprisal, and surprisal values produced by the Roark (2001a); Roark et al. (2009) parser (see Section 4). The next sections describe how and from where these values are obtained[1].

Finally, we run mixed effects regression model analyses (Baayen et al., 2008) with the observed durations as a response variable and the predictors mentioned above in order to detect whether syntactic surprisal is a significant positive predictor of spoken word durations above and beyond the more basic effects of canonical word duration and word frequency.

[1] We will make this data widely available upon publication.

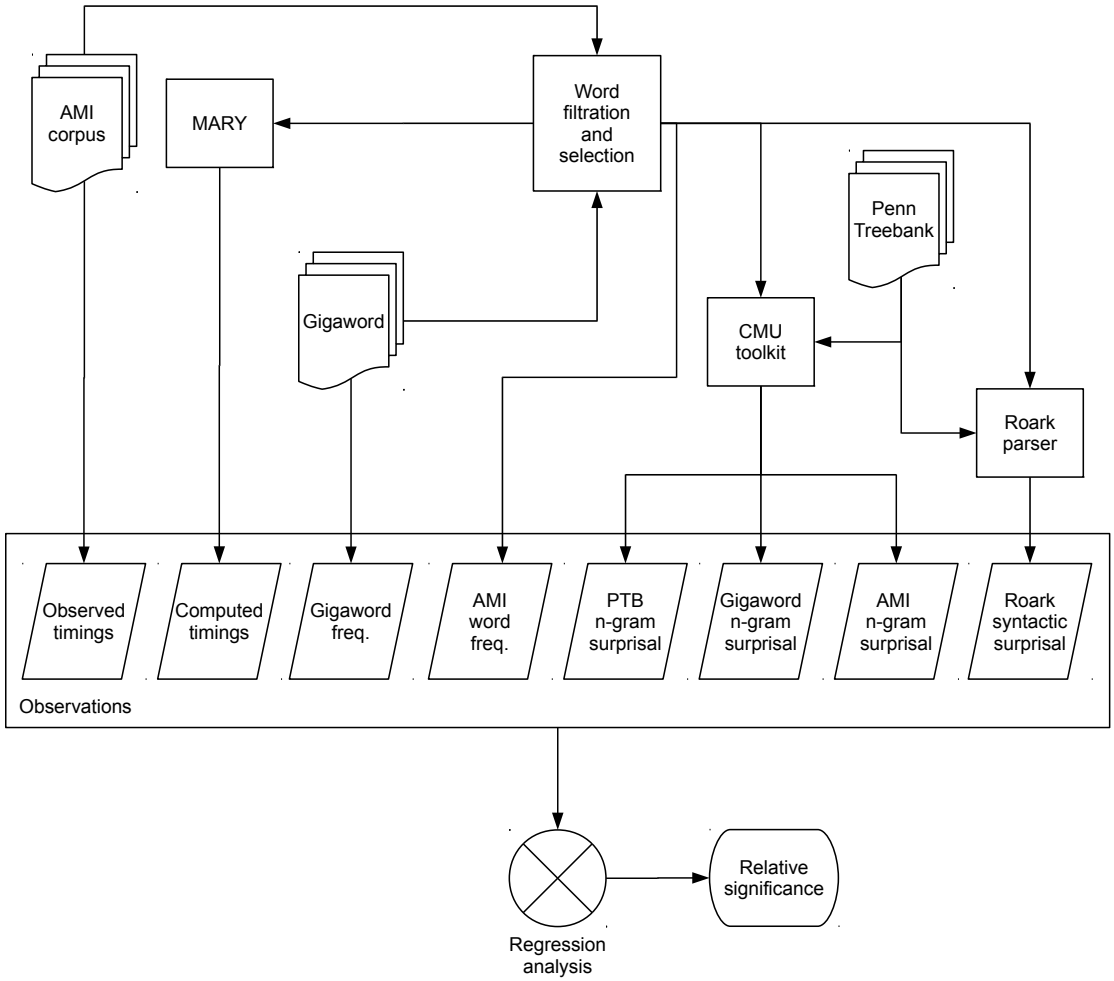

Figure 1: Schematic overview of experiment.

3 Experimental materials

3.1 Corpus preparation

The AMI corpus is provided in the NITE XML Toolkit (NXT) format. We developed a custom interpreter to assemble the relevant data streams: words, meeting IDs, speaker IDs, speaker turns, and observed word durations.

In addition to grouping and re-ordering the information found in the original XML corpus, two more steps were taken to eliminate confounding noise from the data. Non-words (e.g. "uhm", "uh-hmm", etc.) were filtered out, as were incomplete words or incorrectly transcribed words (e.g. "recogn", "somethi", etc); the criterion for rejection was presence in the English Gigaword corpus with subsequent minor corrections by hand, e.g., mapping unseen verbs back into the corpus and correcting obvious common misspellings.[2]

Finally, turns that did not make for complete sentences, e.g., utterances that were interrupted in mid-

[2] A reviewer asks about the extent to which our Gigaword filtering process may remove words we might want to keep but admit words we want to reject. As Gigaword is mostly newswire text, we do not expect the latter case to hold often. AMI is hand-transcribed and uses consistent spellings for non-word interjections (easy to remove), and any spelling mistakes would have to coincide exactly with a Gigaword mistake.

The other way around (rejecting what should be allowed) is easier to check, and we find that of 13K word types in AMI, about 7.2% are rejected for non-appearance in Gigaword, after filtering for interjections like "mm-hmm". However, we manually checked them and returned all but 2.9% of word types to the corpus. These tend to be very low-frequency types. The manual check suggests that ultimately there would be few false rejections.

sentence, were filtered out in order to maximize the proportion of complete parses in surprisal calculation.

3.2 Word duration model

In order to investigate whether there is an association between high/low surprisal and increased/decreased word duration, one needs to have a baseline measure of what constitutes the "canonical" duration of each word—in other words, to account for the fact that some words have longer pronunciations than others. As one reviewer notes, one way of estimating word durations would be to calculate the average duration of each word in the corpus. However, this approach would be insensitive to the phonological, syllabic and phrasal context that a word occurs in, which can have a large effect on word duration. Therefore, we use word duration estimates from the state-of-the-art open-source text-to-speech system MARY (Schröder et al., 2008, version 4.3.1), with the default voice package included in this version (`cmu-slt-hsmm`).

The `cmu-slt-hsmm` voice package uses a Hidden Markov model, trained on the female US English section of the CMU ARCTIC database (Kominek and Black, 2003), to predict prosodic attributes of each individual synthesized phone, including duration. Training was carried out using a version of the HTS system (Zen et al., 2007), modified for using the MARY context features (Schröder et al., 2008) for estimating the parameters of the model and for decoding. Those features include[3]:

- phonological features of the current and neighboring phonemes
- syllabic and lexical features (e.g. syllable stress, (estimated) part-of-speech, position of syllable in word)
- phrasal / sentential features (e.g. sentence/phrase boundaries, neighboring pauses and punctuation)

For each word in the AMI corpus, we obtained two alternative estimates of word duration: one version which is independent of a word's sentential context, and a second version which does take into account the sentential context (such as phrasal/sentential and across-word-boundaries phonological features) the word occurs in. In other words, we obtain MARY word duration estimates in the second version by running individual whole sentences through MARY, segmented by standard punctuation marks used in the AMI corpus transcriptions. For each version, we obtained phone durations using MARY and calculate the total duration of a word as the sum of the estimated phone durations for that word. These durations serve as the "canonical" baselines to which the observed durations of the words in the AMI corpus are compared.

3.3 Word frequency baselines

In order to account for the effects of simple word frequency on utterance duration, we extracted two types of frequency counts. One was taken directly from the AMI corpus alone. The other was taken from a 151 million-word (4.3 million full-paragraph) sample of the English Gigaword corpus. These came from the following newswire sources: Agence France Press, Associated Press Worldstream, New York Times Newswire, and the Xinhua News Agency English Service. These sources are organized by month-of-year. We selected the subset of Gigaword by randomly selecting month-of-year files from those sources with uniform probability. Punctuation was stripped from the beginnings and ends of words before taking the frequency counts.

4 Surprisal models

For predicting the surprisal of utterances in context, two different types of models were used— n-gram probabilities models, as well as Roark's 2001 incremental top-down parser capable of calculating prefix probabilities. We also estimated word frequencies to account for words being spoken more quickly due to their higher frequency which is independent of structural surprisal.

The n-gram probabilities models, while being fast in both training and application, inherently capture very limited contextual influences on surprisal. The full-fledged parser, on the other hand, quantifies sur-

[3]For further information about how HMM-based voices for MARY TTS are trained, see `http://mary.opendfki.de/wiki/HMMVoiceCreation`

prisal based in the prefix probability of the complete sentence prefix and captures long-distance effects by conditioning on c-commanding lexical items as well as non-local node labels such as parents, grandparents and siblings from the left context.

CMU n-grams We used the CMU Statistical Natural Language Modeling Toolkit to provide a convenient way to calculate n-grams probabilities. For the prediction of surprisal, we calculated 3-gram models, 4-gram models and 5-gram models with Witten-Bell smoothing. Different n-gram models were trained on the full Gigaword corpus, as well as the AMI corpus.

To avoid overfitting, the AMI text corpus was split into 10 sub-corpora of equal word counts, preserving coherence of meetings. N-gram probabilities were then calculated for each of the sub-corpora using models trained on the 9 others.

We also produced a trigram model using the text of chapter 2–21 of the Penn Treebank's (PTB) underlying Wall Street Journal corpus. This consists of approximately one million tokens. We generated this model because it is the underlying training data for the Roark parser, described below.

Syntactic Surprisal from Roark parser In order to capture the effect of syntactically expected vs. unexpected events, we can calculate the syntactic surprisal of each word in a sentence. The syntactic surprisal at word S_{w_i} is defined as the difference between the prefix probability at word w_i and the prefix probability at word w_{i-1}. The prefix probability at word w_i is the sum of the probabilities of all trees T spanning words $w_1 \ldots w_i$; see also (Levy, 2008; Demberg and Keller, 2008).

$$S_{w_i} = \log \sum_T P(T, w_1..w_{i-1}) - \log \sum_T P(T, w_1..w_i)$$

The top-down incremental Roark parser (Roark, 2001a) has the characteristic that all partial left-to-right parses are *rooted*: they form a single tree with one root. A set of heuristics ensures that rule application occurs only through node expansion within the connected structure.[4] The grammar-derived prefix probabilities of a given sentence prefix can therefore be calculated directly by multiplying the probabilities of all rules used to generate the prefix tree. The Roark parser shares this characteristic of generating fully connected structures with Earley parsers (Earley, 1970) and left corner parsers (Rosenkrantz and II, 1970).

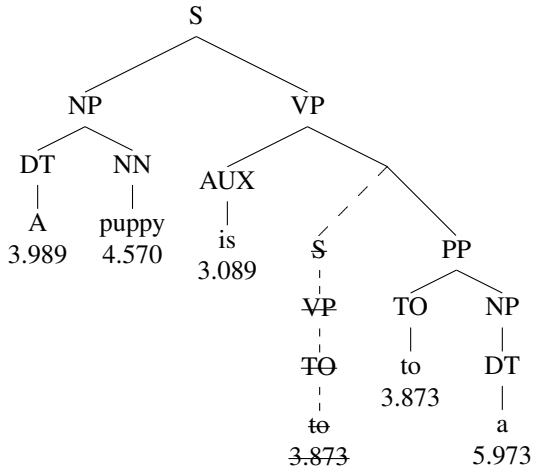

Figure 2: Top-ranked partial parse of *A puppy is to a dog what a kitten is to a cat.*, stopping at the second *a* and providing the Roark parser surprisal values by word. The branch with dashed lines and struck-out symbols represents an analysis abandoned at the appearance of the *a*.

The Roark parser uses a beam search. As the amount of probability mass lost has been shown to be small (Roark, 2001b), the surprisal estimates can be assumed to be a good approximation. The beam width of the parser search is controlled by a "base parsing threshold", which defines the distance in terms of natural log-probability between the most probable parse and the least probable parse within the beam. For the experiments reported here, the parsing beam was set to 21 (default setting is 12). A wider beam also reduces the effects of pruning.

The parser was trained on Wall Street Journal sections 2–21 and applied to parse the full sentences of the AMI corpus, collecting predicted surprisal at each word (see Figure 2 for an example).

The syntactic surprisal can be furthermore be decomposed into a structural and a lexical part: sometimes, high surprisal might be due to a word being incompatible with the high-probability syntactic structures, other times high surprisal might just be due to a lexical item being unexpected. It is inter-

[4]The formulae for the calculation of the prefix probabilities from the PCFG rules can be found in Roark et al. (2009).

esting to evaluate these two aspects of syntactic surprisal separately, and the Roark parser conveniently outputs both surprisal estimates. Structural surprisal is estimated from the occurrence counts of the application of syntactic rules during the parse discounting the effect of lexical probabilities, while lexical surprisal is calculated from the probabilities of the derivational step from the POS-tag to lexical item.

5 Linear mixed effects modelling

In order to test whether surprisal estimates correlate with speech durations, we use linear mixed effects models (LME, Pinheiro and Bates (2000)). This type of model can be thought of as a generalization of linear regression that allows the inclusion of random factors as well as fixed factors.We treat speakers as a random factor, which means that our models contain an intercept term for each speaker, representing the individual differences in speech rates. Furthermore, we include a random slope for the predictors (e.g. frequency, canonical duration, surprisal), essentially accounting for idiosyncrasies of a participant with respect to the predictor, such that only the part of the variance that is common to all participants and is attributed to that predictor.

In a first step, we fit a baseline model with all predictors related to a word's canonical duration and its frequency as well as their random slopes to the observed word durations. Models with more than two random slopes generally did not converge. We therefore included in the baseline model only the two best random slopes (in terms of model fit). We then calculated the residuals of that model, the part of the observed word durations that cannot be accounted for through canonical word durations or word frequency.

For each of our predictors of interest (n-gram surprisal, syntactic surprisal), we then fit another linear mixed-effects model with random slopes to the residuals of the baseline model. This two-step procedure allows us to make sure to avoid problems of collinearity between e.g. surprisal and word frequency or canonical duration. A simpler (but less conservative) method is to directly add the predictors of interest to the baseline model. Results for both modelling variants lead to the same conclusions for our model, so we here report the more conservative two-step model. We compare models based on the Akaike Information Criterion (AIC).

6 Results

Our baseline model uses speech durations from the AMI corpus as the response variable and canonical duration estimates from the MARY TTS system and log word frequencies as predictors. We exclude from the analysis all data points with zero duration (effectively, punctuation) or a real duration longer than 2 seconds. Furthermore, we exclude all words which were never seen in Gigaword and any words for which syntactic surprisal couldn't be estimated. This leaves us with 771,234 out of the 799,997 data points with positive duration.

MARY duration models As mentioned in the earlier sections, we have calculated different versions of the MARY estimated word durations: one model without the sentential context and one model with the sentential context. In our regression analyses, we find, as expected, that the model which includes sentential context achieves a much better fit with the actually measured word durations from the AMI corpus (AIC = 32167) than the model without context (AIC = 70917).

Word frequency estimates We estimated word frequencies from several different resources, from the AMI corpus to have a spoken domain frequency and from Gigaword as a very large resource. We find that both frequency estimates significantly improve model fit over a model that does not contain frequency estimates. Including both frequency estimates improves model fit with respect to a model that includes just one of the predictors (all $p < 0.0001$).

Furthermore, including into the regression an interaction of estimated word duration and word frequency also significantly increases model fit ($p < 0.0001$). This means that words which are short and frequent have longer duration than would be estimated by adding up their length and frequency effects.

Baseline model Fixed effects of the fitted model are shown in Table 2. We see a highly significant effect in the expected direction for both the canonical duration estimate and word frequency. The positive

coefficient for MARY_CONTEXT means that TTS duration estimates are positively correlated with the measured word durations. The negative coefficient for WORDFREQUENCY means that more frequent words are spoken faster than less frequent words. Finally, the negative coefficient for the interaction between word durations and frequencies means that the duration estimate for short frequent and long infrequent words is less extreme than otherwise predicted by the main effects of duration and frequency.

	Ami Dur	Mary Word	Mary Cntxt	Giga Freq	PTB Freq	AMI Freq	AMI 3grm	Giga 4grm
Mary_Word	.36	1						
Mary_Cntxt	.42	.72	1					
GigaFreq	-.35	-.52	-.65	1				
PTBFreq	-.33	-.48	-.62	.98	1			
AMIFreq	-.33	-.61	-.57	.65	.62	1		
AMI3gram	.21	.40	.41	-.41	-.39	-.68	1	
Giga4gram	.24	.33	.44	-.59	-.59	-.44	.61	1
Srprsl	.29	.40	.48	-.71	-.73	-.50	.50	.73

Table 1: Correlations (pearson) of model predictors.

Note though that the predictors are also correlated (for correlations of the main predictors used in these analyses, see Table 1), so there is some collinearity in the below model. Since we are less interested in the exact coefficients and significance sizes for these baseline predictors, this does not have to bother us too much. What is more important, is that we remove any collinearity between the baseline predictors and our predictors of interest, i.e. the surprisal estimates from the ngram models and parser. Therefore, we run separate regression models for these predictors on the residuals of the baseline model.

N-gram estimates We estimated 3-gram, 4-gram and 5-gram models on the AMI corpus (9-fold-cross), the Penn Treebank and the Gigaword Corpus. We found that coefficient estimates and significance levels of the resulting models were comparable. This is not surprising, given that 4-gram and 5-gram models were backing of to 3-grams or smaller contexts for more than 95% of cases on the AMI and PTB corpora (both ca. 1m words), and thus were correlated at $p > .98$. On the Gigaword Corpus, the larger contexts were seen more often (5-grams: 11%, 4-grams: 36%), but still correlation with 3-grams were high at ($p > .96$).

Predictor	Coef	t-value	Sig
INTERCEPT	0.3098	212.11	***
MARY_CONTEXT	0.4987	95.48	***
AMIWORDFREQUENCY	-0.0282	-32.28	***
GIGAWORDFREQUENCY	-0.0275	-62.44	***
MARY_CNTXT:GIGAFREQ	-0.0922	-45.41	***

Table 2: Baseline linear mixed effects model of speech durations on the AMI corpus data for MARY_CONTEXT (including the sentential context), WORDFREQUENCY under speaker with random intercept for speaker and random slopes under speaker. Predictors are centered.

Predictor	Coef	t-value	Sig
INTERCEPT	0.3099	212.94	***
MARY_CONTEXT	0.4970	94.60	***
AMIWORDFREQUENCY	-0.0279	-31.98	***
GIGAWORDFREQUENCY	-0.0254	-53.68	***
GIGA4GRAMSURPRISAL	0.0027	11.81	***
MARY_CNTXT:GIGAFREQ	-0.0912	-44.87	***

Table 3: Linear mixed effects model of speech durations including 4-gram surprisal trained on gigaword as a predictor.

N-gram model surprisal estimated on newspaper texts from PTB or Gigaword were statistically significant positive predictors of spoken word durations beyond simple word frequencies (but PTB ngram surprisal did not improve fit over models containing Gigaword frequency estimates). Counter-intuitively however, ngram models estimated based on the AMI corpus have a small *negative* coefficient in models that already include word frequency as a predictor – residuals of an AMI-estimated ngram model with respect to word frequency are very noisy and do not show a clear correlation anymore with word durations.

Surprisal Surprisal effects were found to have a robust significant positive coefficient, meaning that words with higher surprisal are spoken more slowly / clearly than expected when taking into account only canonical word duration and word frequency. Surprisal achieves a better model fit than any of the n-gram models, based on a comparsion of AICs, and Surprisal significantly improved model fit over a model including frequencies and ngram models based on AMI and Gigaword. Table 4 shows the estimate for SURPRISAL on the residuals of the model in Table 2.

Predictor	Coef	t-value	Sig
INTERCEPT	-0.0154	-23.45	***
SURPRISAL	0.0024	26.09	***

Table 4: Linear mixed effects model of surprisal (based on Roark parser) with random intercept for speaker and random slope. The response variable is residual word durations from the model shown in Table 3.

Surprisal estimated from the Roark parser also remains a significant positive predictor when regressed against the residuals of a baseline model including both 3-gram surprisal from the AMI corpus and 4-gram surprisal from the Gigaword corpus. In order to make really sure that the observed surprisal effect has indeed to do with syntax and can not be explained away as a frequency effect, we also calculated frequency estimates for the corpus based on the Penn Treebank. The significant positive surprisal effect remains stable, also when run on the residuals of a model which includes PTB trigrams and PTB frequencies.

It is difficult from these regression models to intuitively grasp the size of the effect of a particular predictor on reading times, since one would have to know the exact range and distribution of each predictor. To provide some intuition, we calculate the estimated effect size of Roark surprisal on speech durations. Per Roark surprisal "unit", the model estimates a 7 msec difference[5]. The range of Roark surprisal in our data set is roughly from 0 to 25, with most values between 2 and 15. For a word like "thing" which in one instance in the AMI corpus was estimated with a surprisal of 2.179 and in another instance as 16.277, the estimated difference in duration between these instances would thus be 104msec, which is certainly an audible difference. (Full range for Roark surprisal: 174msec, whereas full range for gigaword 4gram surprisal is 35 msec.)

When analysing the surprisal effect in more detail, we find that both the syntactic component of surprisal and its lexical component are significant positive predictors of word durations, as well as the interaction between them, which has a negative slope. A model with the separate components and their interaction achieves a better model fit (in AIC and BIC scores) than a model with only the full surprisal effect. The detailed model is shown in Table 5.

[5] 2.4msec for a unit of residualized Roark surprisal, but it is even less intuitive what that means, hence we calculate with non-residualized surprisal here.

Predictor	Coef	t-value	Sig
INTERCEPT	-0.0219	-18.77	***
STRUCTSURPRISAL	0.0009	2.71	**
LEXICALSURPRISAL	0.0044	24.00	***
STRUCT:LEXICAL	-0.0004	- 6.83	***

Table 5: Linear mixed effects model of residual speech durations wrt. baseline model from Table 3, with random intercept for speaker and random slope for structural and lexical component of surprisal, estimated using the Roark parser.

To summarize, the positive coefficient of surprisal means that words which carry a lot of information from a structural point of view are spoken more slowly than words that carry less such information. These results thus provide good evidence for our hypothesis that the predictability of syntactic structure affects phonetic realization and that speakers use speech rate to achieve more uniform information density.

Native vs. non-native speakers Finally, we also compared effects in our native vs. non-native speaker populations, see Table 6. Both populations show the same effects and tell the same story (note that significance values can't be compared as the sample sizes are different). It might be possible to interpret the findings in the sense that native speakers are more proficient at adapting their speech rate to (syntactic) complexity to achieve more uniform information density, given the slightly higher coefficient and significance for Surprisal for native speakers. Since the effects are statistically significant for both groups, we don't want to make too strong claims about differences between the groups.

7 Conclusions and future work

We have shown evidence in this work that syntactic surprisal effects in transcribed speech data can be detected through word utterance duration in both native and non-native speech, and we did so using a meeting corpus not specifically designed to isolate these effects. This result is the potential foundation for futher work in applied, experimental, and

Predictor	Native English Coef	t-value	Sig	Non-native Coef	t-value	Sig
INTERCEPT	0.2947	149.74	***	0.3221	175.38	***
MARY_CONTEXT	0.5304	69.27	***	0.4699	67.77	***
AMIWORDFREQUENCY	-0.0226	-18.10	***	-0.0321	-28.00	***
GIGAWORDFREQUENCY	-0.0264	-41.19	***	-0.0248	-39.58	***
GIGAWORD4-GRAMS	0.0018	5.36	***	0.0033	10.85	***
MARY_CONTEXT:GIGAFREQ	-0.0810	-27.20	***	-0.0993	-35.71	***
SURPRISAL	0.0033	24.21	***	0.0018	15.09	***
no of data points	320,592			391,106		

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 6: Native speakers are possibly slightly better at adapting their speech rate to syntactic surprisal than non-native speakers. Surprisal value is for model with residuals of other predictors as dependent variable.

theoretical psycholinguistics. It provides additional direct support for approaches based on the UID hypothesis.

From an applied perspective, the fact that frequency and syntactic surprisal have a significant effect beyond what a HMM-trained TTS model would predict for individual words is a case for further research into incorporating syntactic models into speech production systems. Our methodology immediately provides a framework for estimating the word-by-word effect on duration for increased naturalness in TTS output. This is relevant to spoken dialogue systems because it appears that synthesized speech requires a greater level of attention from the dialogue system users when compared to the same words delivered in natural speech (Delogu et al., 1998). Some of this effect may be attributable to peaks in information density which are caused by current generation systems not compensating for areas of high information density through speech rate, lexical and structural choice.

Furthermore, syntax and semantics have been observed to interact with the mode of speech delivery. Eye-tracking experiments by Swift et al. (2002) showed that there was a synthetic vs. natural speech difference in the time required to pay attention to an object referred to using definite articles, but not indefinite articles. Our result points a way towards a direction for explaining of this phenomenon by demonstrating that the differences between current-technology artificial speech and natural speech can be partially explained through higher-level syntactic features.

However, further experimentation is required on other measures of syntactic complexity (e.g. DLT, Gibson (2000)) as well as other levels of representation such as the semantic level. From a theoretical and neuroanatomical perspective, the finding that a measure of syntactic ambiguity reduction has an effect on the phonological layer of production has additional implications for the organization of the human language production system.

References

Aylett, M. and Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the acoustical society of America*, 119(5):3048–3059.

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Delogu, C., Conte, S., and Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication*, 24(2):153–168.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109:193–210.

Earley, J. (1970). An efficient context-free parsing algorithm. *Commun. ACM*, 13(2):94–102.

Fang, R., Chai, J. Y., and Ferreira, F. (2009). Be-

tween linguistic attention and gaze fixations inmultimodal conversational interfaces. In *International Conference on Multimodal Interfaces*, pages 143–150.

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

Frank, A. and Jaeger, T. F. (2008). Speaking rationally: uniform information density as an optimal strategy for language production. In *The 30th annual meeting of the Cognitive Science Society*, pages 939—944.

Frank, S. (2010). Uncertainty reduction as a measure of cognitive processing effort. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, pages 81–89, Uppsala, Sweden.

Gibson, E. (2000). Dependency locality theory: A distance-dased theory of linguistic complexity. In Marantz, A., Miyashita, Y., and O'Neil, W., editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT Press, Cambridge, MA.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, volume 2, pages 159–166, Pittsburgh, PA.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. (2001). Evidence from reduction in lexical production. *Frequency and the emergence of linguistic structure*, 45:229.

Kominek, J. and Black, A. (2003). The cmu arctic speech databases for speech synthesis research. *Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMULTI-03-177 http://festvox. org/cmu arctic*.

Kuperman, V., Pluymaekers, M., Ernestus, M., and Baayen, H. (2007). Morphological predictability and acoustic duration of interfixes in dutch compounds. *The Journal of the Acoustical Society of America*, 121(4):2261–2271.

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

Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*.

Piantadosi, S., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9).

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Statistics and computing series. Springer-Verlag.

Roark, B. (2001a). Probabilistic top-down parsing and language modeling. *Computational linguistics*, 27(2):249–276.

Roark, B. (2001b). *Robust probabilistic predictive syntactic processing: motivations, models, and applications*. PhD thesis, Brown University.

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Rosenkrantz, D. J. and II, P. M. L. (1970). Deterministic left corner parsing (extended abstract). In *SWAT (FOCS)*, pages 139—152.

Schröder, M., Charfuelan, M., Pammi, S., and Türk, O. (2008). The MARY TTS entry in the Blizzard Challenge 2008. In *Proc. Blizzard Challenge*. Citeseer.

Swift, M. D., Campana, E., Allen, J. F., and Tanenhaus, M. K. (2002). Monitoring eye movements as an evaluation of synthesized speech. In *Proceedings of the IEEE 2002 Workshop on Speech Synthesis*.

Taube-Schiff, M. and Segalowitz, N. (2005). Linguistic attention control: attention shifting governed by grammaticized elements of language. *Journal of experimental psychology Learning memory and cognition*, 31(3):508–519.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., and Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0.

In *Proc. of Sixth ISCA Workshop on Speech Synthesis*, pages 294–299.