# Vector-space calculation of semantic surprisal for predicting word pronunciation duration

**Asad Sayeed, Stefan Fischer, and Vera Demberg**
Computational Linguistics and Phonetics/M$^2$CI Cluster of Excellence
Saarland University
66123 Saarbrücken, Germany
{asayeed,sfischer,vera}@coli.uni-saarland.de

## Abstract

In order to build psycholinguistic models of processing difficulty and evaluate these models against human data, we need highly accurate language models. Here we specifically consider *surprisal*, a word's predictability in context. Existing approaches have mostly used n-gram models or more sophisticated syntax-based parsing models; this largely does not account for effects specific to semantics. We build on the work by Mitchell et al. (2010) and show that the semantic prediction model suggested there can successfully predict spoken word durations in naturalistic conversational data.

An interesting finding is that the training data for the semantic model also plays a strong role: the model trained on in-domain data, even though a better language model for our data, is not able to predict word durations, while the out-of-domain trained language model does predict word durations. We argue that this at first counter-intuitive result is due to the out-of-domain model better matching the "language models" of the speakers in our data.

## 1 Introduction

The Uniform Information Density (UID) hypothesis holds that speakers tend to maintain a relatively constant rate of information transfer during speech production (e.g., Jurafsky et al., 2001; Aylett and Turk, 2006; Frank and Jaeger, 2008). The rate of information transfer is thereby quantified using as each words' *Surprisal* (Hale, 2001), that is, a word's negative log probability in context.

$$Surprisal(w_i) = -\log P(w_i|w_1..w_{i-1})$$

This work makes use of an existing measure of semantic surprisal calculated from a distributional space in order to test whether this measure accounts for an effect of UID on speech production. Our hypothesis is that a word in a semantically surprising context is pronounced with a slightly longer duration than the same word in a semantically less-expected context. In this way, a more uniform rate of information transfer is achieved, because the higher information content of the unexpected word is stretched over a slightly longer time. To our knowledge, the use of this form of surprisal as a pronunciation predictor has never been investigated.

The intuition is thus: in a sentence like *the sheep ate the long grass*, the word *grass* will have relatively high surprisal if the context only consists of *the long*. However, a distributional representation that retains the other content words in the sentence, thus representing the contextual similarity of *grass* to *sheep ate*, would able to capture the relevant context for content word prediction more easily. In the approach taken here, both types of models are combined: a standard language model is reweighted with semantic similarities in order to capture both short- and more long-distance dependency effects within the sentence.

The semantic surprisal model, a reimplementation of Mitchell (2011), uses a word vector $w$ and a history or context vector $h$ to calculate the language model $p(w|h)$, defining this probability in vector space via cosine similarity. Words that have a higher distributional similarity to their context are thus represented as having a higher probability than words that do not. Thus, we calculate probabilities for words in the context of a sentence in a framework of distributional semantics.

Regarding our main hypothesis—that speakers adapt their speech rate as a function of a word's information content—it is particularly important to

*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 763–773,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

us to test this hypothesis on fully "natural" conversational data. Therefore, we use the AMI corpus, which contains transcripts of English-language conversations with orthographically correct transcriptions and precise word pronunciation boundaries in terms of time.

We will explain the calculation of semantic surprisal in section 4 (this is so far only described in Mitchell's 2011 PhD thesis), and then evaluate the effect of an in-domain semantic surprisal model in section 7. Next, we will compare this to the effect of an out-of-domain semantic surprisal model in section 8. The hypothesis is only confirmed for the out-of-domain model, which we argue is due to this model being more similar to the speaker's internal "model" than the in-domain model.

## 2 Background

### 2.1 Surprisal and UID

Surprisal is defined in terms of the negative logarithm of the probability of a word in context: $S(w) = -\log P(w|context)$, where $P(w|context)$ is the probability of a word given its previous (linguistic) context. It is a measure of information content in which a high surprisal implies low predictability. The use of surprisal in psycholinguistic research goes back to Hale (2001), who used a probabilistic Earley Parser to model the difficulty in parsing so-called garden path sentences (e.g. "The horse raced past the barn fell"), wherein the unexpectedness of an upcoming word or structure influences the language processor's difficulty. Recent work in psycholinguistics has provided increasing support (e.g., Levy (2008); Demberg and Keller (2008); Smith and Levy (2013); Frank et al. (2013)) for the hypothesis that the surprisal of a word is proportional to the processing difficulty (measured in terms of reading times and EEG event-related potentials) it causes to a human.

The Uniform Information Density (UID) hypothesis (Frank and Jaeger, 2008) holds that speakers tend distribute information uniformly across an utterance (in the limits of grammaticality). Information density is quantified in terms of the surprisal of each word (or other linguistic unit) in the utterance. These notions go back to Shannon (1948), who showed that conveying information uniformly close to channel capacity is optimal for communication through a (noisy) communication channel.

Frank and Jaeger (2008) investigated UID effects in the SWITCHBOARD corpus at a morphosyntactic level wherein speakers avoid using English contracted forms ("you are" vs. "you're") when the contractible phrase is also transmitting a high degree of information in context. In this case, n-gram surprisal was used as the information density measure. Related hypotheses have been suggested by Jurafsky et al. (2001), who related speech durations to bigram probabilities on the Switchboard corpus, and Aylett and Turk (2006), who investigated information density effects at the syllable level. They used a read-aloud English speech synthesis corpus, and they found that there is an inverse relationship between the pronunciation duration and the N-gram predictability. Demberg et al. (2012) also use the AMI corpus used in this work, and show that syntactic surprisal (i.e., the surprisal estimated from Roark's (2009) PCFG parser) can predict word durations in natural speech.

Our work expands upon the existing efforts in demonstrating the UID hypothesis by applying surprisal to the level of lexical semantics.

### 2.2 Distributional semantics

Given a means of evaluating the similarity of linguistic units (e.g., words, sentences, texts) in some numerical space that represents the contexts in which they appear, it is possible to approximate the semantics in distributional terms. This is usually done by collecting statistics from a corpus using techniques developed for information retrieval. Using these statistics as a model of semantics is justified in terms of the "distributional hypothesis", which holds that words used in similar contexts have similar meanings (Harris, 1954).

A simple and widely-used type of distributional semantic model is the vector space model (Turney and Pantel, 2010). In such a model, all words are represented each in terms of vectors in a single high-dimensional space. The semantic similarity of words can then be calculated via the cosine of the angle between the vectors in this manner: $\cos(\varphi) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$. Closed-class function words are usually excluded from this calculation. Until relatively recently (Erk, 2012), distributional semantic models did not take into account the fine-grained details of syntactic and semantic structure construed in formal terms.

## 3 Corpus

The AMI Meeting Corpus (Carletta, 2007) is a multimodal English-language corpus. It contains videos and transcripts of simulated workgroup meetings accompanied by various kinds of annotations. The corpus is available along with its annotations under a free license[1].

Two-thirds of the videos contain simulated meetings of 4-person design teams assigned to talk about the development of a fictional television remote control. The remaining meetings discuss various other topics. The majority of speakers were non-native speakers of English, although all the conversations were held in English. The corpus contains about 100 hours of material.

An important characteristic of this corpus for our work is that the transcripts make use of consistent English orthography (as opposed to being phonetic transcripts). This enables the use of natural language processing techniques that require the reliable identification of words. Grammatical errors, however, remain in the corpus. The corpus includes other annotations such as gesture and dialog acts. Most important for our work are the time spans of word pronunciation, which are precise to the hundredth of a second.

We removed interjections, incomplete words, and transcriptions that were still misspelled from the corpus, and we took out all incomplete sentences. This left 951,769 tokens (15,403 types) remaining in the corpus.

## 4 Semantic surprisal model

We make use of a re-implementation of the semantic surprisal model presented in Mitchell et al. (2010). As this paper does not provide a detailed description of how to calculate semantic surprisal, our re-implementation is based on the description in Mitchell's PhD thesis (2011).

In order to calculate surprisal, we need to be able to obtain a good estimate of a word given previous context. Mitchell uses the following concepts in his model:

- $h_{n-1}$ is the *history* and represents all the previous words in the sentence. If $w_n$ is the current word, then $h_{n-1} = w_1 \dots w_{n-1}$. The vector-space semantic representation of $h_{n-1}$ is calculated from the composition of individual word vectors, which we call $\vec{h}_{n-1}$.
- *context words* represent the dimensions of the word vectors. The value of a word vector's component is the co-occurrence of that word with a context word. The context words consist of the most frequent words in the corpus.
- we use *word class* and distinguish between content words and function words, for which we use open and closed classes as a proxy.

### 4.1 Computing the vector components

The proportion between two probabilities $\frac{p(c_i|w)}{p(c_i)}$ is used for calculating vector components, where $c_i$ is the $i$th context dimension and $w$ is the given word in the current position. We can calculate each vector component $v_i$ for a word vector $\vec{v}$ according to the following equation:

$$v_i = \frac{p(c_i|w)}{p(c_i)} = \frac{f_{c_i w} f_{total}}{f_w f_{c_i}} \tag{1}$$

where $f_{c_i w}$ is the cooccurrence frequency of $w$ and $c_i$ together, $f_{total}$ is the total corpus size, and $c_i$ represents the unigram frequencies of $w$. All future steps in calculating our language model rely on this definition of $v_i$.

### 4.2 Semantic probabilities

For the goal of computing $p(w|h)$, we use the basic idea that the more "semantically coherent" a word is with its history, the more likely it is. Cosine similarity is a common way to define this similarity mathematically in a distributional space, producing a value in the interval $[-1, 1]$. We use the following definitions, wherein $\varphi$ is the angle between $\vec{w}$ and $\vec{h}$:

$$\cos(\varphi) = \frac{\vec{w} \cdot \vec{h}}{|\vec{w}||\vec{h}|} \tag{2}$$

$$\vec{w} \cdot \vec{h} = \sum_i w_i h_i \tag{3}$$

Mitchell notes that there are at least three problems with using cosine similarity in connection with the construction of a probabilistic model: (a) the sum of all cosine values is not unity, (b) word frequency does not pay a role in the calculation, such that a rare synonym of a frequent word might get a high similarity rating, despite low predictability, and (c) the calculation can result in negative values.

[1] http://groups.inf.ed.ac.uk/ami/download/

This problem is addressed by two changes to the notion of dot product used in the calculation of the cosine:

$$\vec{w} \cdot \vec{h} = \sum_i \frac{p(c_i|w)}{p(c_i)} \frac{p(c_i|h)}{p(c_i)} \tag{4}$$

The influence of word frequencies is then restored using $p(w)$ and $p(c_i)$:

$$p(w|h) = p(w) \sum_i \frac{p(c_i|w)}{p(c_i)} \frac{p(c_i|h)}{p(c_i)} p(c_i) \tag{5}$$

This expression reweights the new scalar product with the likelihood of the given words and the context words. We refer the reader to Mitchell (2011) in order to see that this is a true probability. The application of Bayes' Rule allows us to rewrite the formula as $p(w|h) = \sum_i p(w|c_i)p(c_i|h)$. Nevertheless, equation (5) is better suited to our task, as it operates directly over our word vectors.

### 4.3 Incremental processing

Equation (5) provides a conditional probability for a word $w$ and its history $h$. To calculate the product $\frac{p(c_i|w)}{p(c_i)} \frac{p(c_i|h)}{p(c_i)}$, we need the components of the vectors for $w$ and $h$ at the current position in the sentence. We can get $\vec{w}$ from directly from the vector space of words. However, $\vec{h}$ does not have a direct representation in that space, and it must be constructed compositionally:

$$\vec{h}_1 = \vec{w}_1 \qquad \text{Initialization} \tag{6}$$

$$\vec{h}_n = f(\vec{h}_{n-1}, \vec{w}_n) \qquad \text{Composition} \tag{7}$$

$f$ is a vector composition function that can be chosen independently from the model. The history is initialized using the vector of the first word and combined step-by-step with the vectors of the following words. History vectors that arise from the composition step are normalized[2]:

$$h_i = \frac{\hat{h}_i}{\sum_j \hat{h}_j p(c_j)} \qquad \text{Normalization} \tag{8}$$

The equations (5), (6), (7), and (8) represent a simple language model, assuming calculation of vector components with equation (1).

### 4.4 Accounting for word order

The model described so far is based on semantic coherence and mostly ignores word order. Consequently, it has poor predictive power. In this section, we describe how a notion of word order is included in the model through the integration of an n-gram language model.

Specifically, equation (5) can be represented as the product of two factors:

$$p(w|h) = p(w)\Delta(w,h) \tag{9}$$

$$\Delta(w,h) = \sum_i \frac{p(c_i|w)}{p(c_i)} \frac{p(c_i|h)}{p(c_i)} p(c_i) \tag{10}$$

where $\Delta$ is the semantic component that scales $p(w)$ in function of the context. A word $w$ that has a close semantic similarity to a history $h$ should receive higher or lower probability depending on whether $\Delta$ is higher or lower than 1. In order to make this into a prediction, $p(w)$ is replaced with a trigram probability.

$$\hat{p}(w_n, h_{n-1}, w_{n-2}^{n-1}) = p(w_n|w_{n-2}^{n-1})\Delta(w_n, h_{n-1}) \tag{11}$$

However, this change means that the result is no longer a true probability. Instead, equation 11 can be seen as an estimate of semantic similarity. In order to restore its status as a probability, Mitchell includes another normalization step:

$$p(w_n|h_{n-3}, w_{n-2}^{n-1}) = \begin{cases} p(w_n|w_{n-2}^{n-1}) \\ \qquad \text{Function words} \\ \frac{\hat{p}(w_n, h_{n-3}, w_{n-2}^{n-1})}{\sum\limits_{w_c} \hat{p}(w_c, h_{n-3}, w_{n-2}^{n-1})} \\ \qquad \sum\limits_{w_c} p(w_c|w_{n-2}^{n-1}) \\ \qquad \text{Content words} \end{cases} \tag{12}$$

The model hence simply uses the trigram model probability for function words, making the assumption that the distributional representation of such words does not include useful information. On the other hand, content words obtain a portion of the probability mass whose size depends on its similarity estimate $\hat{p}(w_n, h_{n-3}, w_{n-2}^{n-1})$ relative to the similarity estimates of all other words $\sum_{w_c} \hat{p}(w_c, h_{n-3}, w_{n-2}^{n-1})$. The factor $\sum_{w_c} p(w_c|w_{n-2}^{n-1})$ ensures that not all of the probability mass is divided up among the content words $w_c$; rather, only the mass assigned by the n-gram model at position $w_{n-2}^{n-1}$ is re-distributed. The

[2] This equation is slightly different from what appears in Mitchell (2011). We present here a corrected formula based on private communication with the author.

probability mass of the function words remains unchanged.

Mitchell (2011) restricts the history so that only words outside the trigram window are taken into account in order to keep the n-gram model and the semantic similarity model independent. Thus, the n-gram model represents local dependencies, and the semantic model represents longer-distance dependencies.

The final model that we use in our experiment consists of equations (1), (6), (7), (8) and (12).

## 5 Evaluation Methods

Our goal is to test whether semantically reweighted surprisal can explain spoken word durations over and above more simple factors that are known to influence word durations, such as word length, frequency and predictability using a simpler language model. Our first experiment tests whether semantic surprisal based on a model trained using in-domain data is predictive of word pronunciation duration, considering the UID hypothesis. For our in-domain model, we estimate surprisal using 10-fold cross-validation over the AMI corpus: we divide the corpus into ten equally-sized segments and produce surprisal values for each word in each segment based on a model trained from the other nine segments. We then use linear mixed effects modeling (LME) via the lme4 package in R (Pinheiro and Bates, 2000; Bates et al., 2014) in order to account for word pronunciation length. We follow the approach of Demberg et al. (2012).

Linear mixed effects modelling is a generalization of linear regression modeling and includes both fixed effects and random effects. This is particularly useful when we have a statistical units (e.g., speakers) each with their own set of repeated measures (e.g., word duration), but each such unit has its own particular characteristics (e.g., some speakers naturally speak more slowly than others). These are the random effects. The fixed effects are those characteristics that are expected not to vary across such units. LME modeling learns coefficients for all of the predictors, defining a regression equation that should account for the data in the dependent variable (in our case, word pronunciation duration). The variance in the data that a model cannot explain is referred to as the residual. We denote statistical significances in the following way: *** means a p-value $\leq 0.001$, ** means $p \leq 0.01$, * means $p \leq 0.05$, and no stars means that the predictor is not significant ($p > 0.05$).

In our regression models, all the variables are centered and scaled to reduce effects of correlations between predictors. Furthermore, we log-transformed the response variable (actual spoken word durations from the corpus) as well as the duration estimates from the MARY speech synthesis system to obtain more normal distributions, which are prerequisite for applying the LME models. All conclusions drawn here also hold for versions of the model where no log transformation is used.

From the AMI corpus, we filter out data points (words) that have a pronunciation duration of zero or those that are longer than two seconds, the latter in order to avoid including such things as pauses for thought. We also remove items that are not represented in Gigaword. That leaves us with 790,061 data points for further analysis. However, in our semantic model, function words are not affected by the $\Delta$ semantic similarity adjustment and are therefore not analyzable for the effect of semantically-weighted trigram predictability. That leaves 260k data points for analysis in the models.

## 6 Baseline model

As a first step, we estimate a baseline model which does not include the in-domain semantic surprisal. The response variable in this model are the word durations observed in the corpus. Predictor variables include $D_{\text{MARY}}$ (the context-dependent spoken word duration as estimated by the MARY speech synthesis system), word frequency estimates from the same domain as well as the GigaWord corpus ($F_{\text{AMI}}$ and $F_{\text{Giga}}$, both as log relative frequencies), the interaction between estimated word durations and in-domain frequency, ($D_{\text{MARY}}$:$F_{\text{AMI}}$) and a domain-general trigram model ($S_{\text{AMI-3}}$). Our model also includes a random intercept for each speaker, as well as random slopes under speaker for $D_{\text{MARY}}$ and $S_{\text{AMI-3}}$. The baseline model is shown in Table 1.

All predictors in the baseline model shown in Table 1 significantly improve model fit. We can see that the MARY-TTS estimated word durations are a positive highly significant predictor in the model. Furthermore, the word frequency estimates from the domain general corpus as well as the in-domain frequency estimates are significant negative predictors of word durations, this means

| Predictor | Coefficient | t-value | Sig. |
|---|---|---|---|
| (Intercept) | 0.034 | 4.90 | *** |
| $D_{MARY}$ | 0.427 | 143.97 | *** |
| $F_{AMI}$ | -0.137 | -60.26 | *** |
| $F_{Giga}$ | -0.051 | -18.92 | *** |
| $S_{Giga\text{-}3gram}$ | 0.032 | 10.94 | *** |
| $D_{MARY}$:$F_{AMI}$ | -0.003 | -2.12 | * |

Table 1: Fixed effects of a baseline model including the data points for which we could calculate semantic surprisal.

that as expected, words durations are shorter for more frequent words. We can furthermore see that n-gram surprisal is a significant positive predictor of spoken word durations; i.e., more unexpected words have longer durations than otherwise predicted. Finally, there is also a significant interaction between estimated word durations and in-domain word frequency, which means that the duration of long and frequent words is corrected slightly downward.

## 7 Experiment 1: in-domain model

The AMI corpus contains spoken conversations, and is thus quite different from the written corpora we have available. When we train an n-gram model in domain (using 10-fold cross validation), perplexities for the in-domain model (67.9) are much lower than for a language model trained on gigaword (359.7), showing that the in-domain model is a better language model for the data[3].

In order to see the effect of semantic surprisal estimated based on the in-domain language model and reweighted for semantic similarity within the same sentence as described in Section 3, we then expand the baseline model, adding $S_{Semantics}$ as a predictor. Table 2 shows the fixed effects of this expanded model. The predictor for semantic surprisal is significant, but the coefficient is negative. This apparently contradicts our hypothesis that semantic surprisal has a UID effect on pronunciation duration, so that higher $S_{Semantics}$ means higher $D_{AMI}$. We found that these results are very stable—in particular, the same results also hold if we estimate a separate model with $S_{Semantics}$ as a predictor and residuals of the baseline model as a response variable, and when we include in-domain semantic surprisal in a model where there ngram surprisal on the out of domain corpus is not included as a predictor variable.

| Predictor | Coefficient | t-value | Sig. |
|---|---|---|---|
| (Intercept) | 0.031 | 4.53 | ** |
| $D_{MARY}$ | 0.428 | 144.06 | *** |
| $F_{AMI}$ | -0.148 | -59.15 | *** |
| $F_{Giga}$ | -0.043 | -15.10 | *** |
| $S_{Giga\text{-}3gram}$ | 0.047 | 14.60 | *** |
| $S_{Semantics}$ | -0.028 | -9.78 | *** |
| $D_{MARY}$:$F_{AMI}$ | -0.003 | -2.27 | * |

Table 2: Fixed effects of the baseline model with semantic surprisal (including also a random slope for semantic surprisal under subject).

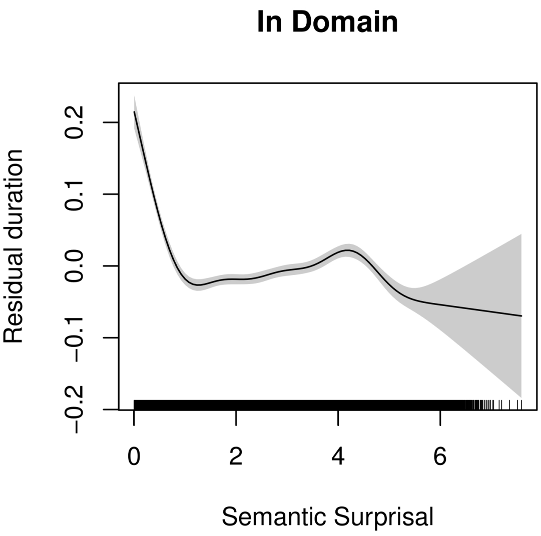

Figure 1: GAM-calculated spline for $S_{Semantics}$ for the in-domain model.

In order to understand the unexpected behaviour of $S_{Semantics}$, we make use of a generalized additive model (GAM) with the R package mgcv. Compared to LME models, GAMs are parameter-free and do not assume a linear form of the predictors. Instead, for every predictor, GAMs can fit a spline. We learn a GAM using the residuals of the baseline model as a response variable and fitting semantic surprisal based on the in-domain model; see Table 2.

In figure 1, we see that $S_{Semantics}$ is poorly fit by a linear function. In particular, there are two intervals in the curve. Between surprisal values 0

[3] Low perplexity estimates are reflective of the spoken conversational domain. Perplexities on content words are much higher: 357.3 for the in-domain model and 2169.8 for the out of domain model.

and 1.5, the curve falls, but between 1.5 and 4, it rises. (For high surprisal values, there are too few data points from which to draw conclusions.)

Therefore, we decided to divide the data up into datapoints with $\mathrm{S}_{\mathrm{Semantics}}$ above 1.5 and below 1.5. We then modelled the effect of $\mathrm{S}_{\mathrm{Semantics}}$ on the residuals of the baseline model, with $\mathrm{S}_{\mathrm{Semantics}}$ as a random effect. This is to remove a possible effect of collinearity between $\mathrm{S}_{\mathrm{Semantics}}$ and the other predictors.

| Interval of $\mathrm{S}_{\mathrm{Semantics}}$ | Predictor | Coef. | t-value | Sig. |
|---|---|---|---|---|
| $[0, \infty[$ | (Intercept) | 0 | 0 | |
| | $\mathrm{S}_{\mathrm{Semantics}}$ | -0.013 | -7.01 | *** |
| $[0, 1.5[$ | (Intercept) | 0 | 0 | |
| | $\mathrm{S}_{\mathrm{Semantics}}$ | -0.06 | -18.56 | *** |
| $[1.5, \infty[$ | (Intercept) | 0 | 0 | |
| | $\mathrm{S}_{\mathrm{Semantics}}$ | 0.013 | 5.50 | *** |

Table 3: Three models of $\mathrm{S}_{\mathrm{Semantics}}$ as a random effect over the residuals of baseline models learned from the remaining fixed effects. The first model is over the entire range.

Table 3 shows that the random effect of semantic surprisal is positive and significant in the range of semantic surprisal above 1.5. That low surprisals have the opposite effect compared to what we expect suggests to us that using the AMI corpus as an in-domain source of training data presents a problem. The observed result for the relationship between semantic surprisal and spoken word durations does not only hold for the semantic surprisal model, but also for the standard non-weight-adjusted in-domain trigram model. We therefore hypothesize that our semantic surprisal model is producing surprisal values that are low because they are common in this domain (both higher frequency and higher similarities), but speakers are coming to the AMI task with "models" trained on out-of-domain data. Thus, words that are apparently very low-surprisal display longer pronunciation durations as an artifact of the model. To test this, we conducted a second experiment, for which we built a model with out-of-domain data.

## 8 Experiment 2: out-of-domain training

In order to test for the effect of possible underestimation of surprisal due to in-domain training, we also tested the semantic surprisal model when trained on more domain-general text. As training data for our semantic model, we use a randomly selected 1% (by sentence) of the English Gigaword 5.0 corpus. This is lowercased, with hapax legomena treated as unknown words. We test the model against the entire AMI corpus. Furthermore, we also compare our semantic surprisal values to the syntactic surprisal values calculated by Demberg et al. (2012) for the AMI corpus, which we obtained from the authors. As noted above, the out-of-domain language model has higher perplexity on the AMI corpus—that is, it is a lower-performing language model. On the other hand, it may represent overall speaker experience more accurately than the in-domain model; in other words, it may be a better model of the speaker.

### 8.1 Results

Once again, the semantic surprisal model is only different from a general n-gram model on content words. We therefore first compare whether the model that is reweighted for semantic surprisal can explain more of the variance than the same model without semantic reweighting.

We again use the same baseline model as for the in-domain experiment, see table 1. As the semantic surprisal model represents a reweighted trigram model, there is a high correlation between the trigram model and the semantic surprisal model. We thus need to know whether the semantically reweighted model is **better** than the simple trigram model. When we compare a model that contains both trigram surprisal and semantic surprisal as a predictor, we find that this model is significantly better than the model including only trigram surprisal (AIC of baseline model: 618427; AIC of model with semantic surprisal: 618394; $\chi^2 = 35.8; p < 0.00001$). On the other hand, the model including both predictors is only marginally better than the model including semantic surprsial (AIC of semantic surprisal model: 618398). This means that the simpler trigram surprisal model does not contribute anything over the semantic model, and that the semantic model fits the word duration data better. Table 4 shows the model with semantic surprisal as a predictor.

Furthermore, we wanted to check whether our hypothesis about the negative result for the in-domain model was indeed due to an under-estimation of surprisal of in-domain words for the

| Predictor | Coefficient | t-value | Sig. |
|---|---|---|---|
| (Intercept) | 0.034 | 4.90 | *** |
| $D_{\text{MARY}}$ | 0.427 | 144.36 | *** |
| $F_{\text{AMI}}$ | -0.135 | -58.76 | *** |
| $F_{\text{Giga}}$ | -0.053 | -19.99 | *** |
| $S_{\text{Semantics}}$ | 0.034 | 11.70 | *** |
| $D_{\text{MARY}}$:$F_{\text{AMI}}$ | -0.003 | -2.09 | * |

Table 4: Model of spoken word durations, with random intercept and random slopes for $D_{\text{MARY}}$ and $S_{\text{Semantics}}$ under speaker.

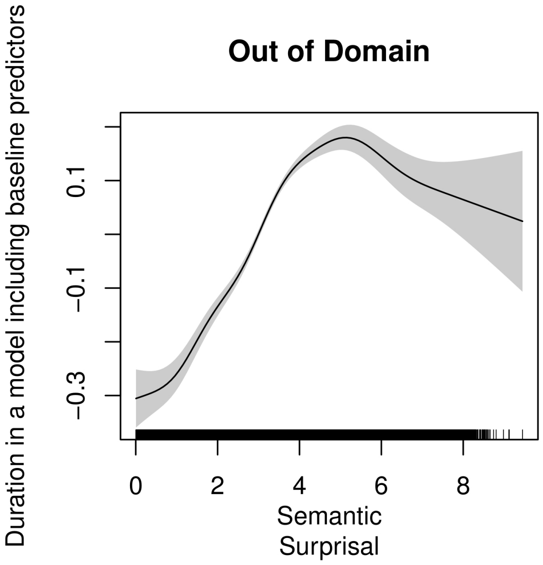

Figure 2: GAM-calculated spline for $S_{\text{Semantics}}$ for the ouf-of-domain model.

in-domain model. We again calculate a GAM model showing the effect of out-of-domain semantic surprisal in a model containing also the baseline predictors, see figure 2.

We can see that word durations increase with increasing semantic surprisal, and that there is in particular no effect of longer word durations for low surprisal words. This result is also confirmed by LME models splitting up the data in small and large surprisal values, as done for the in-domain model in Table 3; semantic surprisal based on the out-of-domain model is a significant positive predictor in both data ranges.

Next, we tested whether the semantic similarity model improves model fit over and above a model also containing syntactic surprisal as a predictor. We find that syntactic surprisal improves model fit over and above the model including semantic surprisal ($\chi^2 = 309.5; p < 0.00001$), and that semantic surprisal improves model fit over and above a model including syntactic surprisal and trigram surprisal ($\chi^2 = 28.5; p < 0.00001$). Table 5 shows the model containing both syntactic based on the Roark parser ((Roark et al., 2009); see also Demberg et al. (2012) for use of syntactic surprisal for estimating spoken word durations) and semantic surprisal.

| Predictor | Coefficient | t-value | Sig. |
|---|---|---|---|
| (Intercept) | -0.058 | -6.58 | *** |
| $D_{\text{MARY}}$ | 0.425 | 144.04 | *** |
| $F_{\text{AMI}}$ | -0.131 | -57.04 | *** |
| $F_{\text{Giga}}$ | -0.051 | -19.41 | *** |
| $S_{\text{Syntax}}$ | 0.011 | 17.61 | *** |
| $S_{\text{Semantics}}$ | 0.015 | 4.99 | *** |
| $D_{\text{MARY}}$:$F_{\text{AMI}}$ | -0.007 | -4.44 | *** |

Table 5: Linear mixed effects model for spoken word durations in the AMI corpus, for a model including both syntactic and semantic surprisal as a predictor as well as a random intercept and slope for $D_{\text{MARY}}$ and $S_{\text{Semantics}}$ under speaker.

Finally, we split our dataset into data from native and non-native speakers of English (305 native speakers, vs. 376 non-native speakers). Table 6 shows generally larger effects for native than non-native speakers. In particular, the interaction between duration estimates and word frequencies, and semantic surprisal were not significant predictors in the non-native speaker model (however, random slopes for semantic surprisal under speaker still improved model fit very strongly, showing that non-native speakers differ in whether and how they take into account semantic surprisal during language production).

## 9 Discussion

Our analysis shows that high information density at one linguistic level of description (for example, syntax or semantics) can lead to a compensatory effect at a different linguistic level (here, spoken word durations). Our data also shows however, that the choice of training data for the models is important. A language model trained exclusively in a specific domain, while a good language model, may not be representative of speaker's overall language experience. This is particularly relevant for the AMI corpus, in which groups of

|  | Native | Speaker |  | Non-native | Speaker |  |
|---|---|---|---|---|---|---|
| Predictor | Coefficient | t-value | Sig. | Coefficient | t-value | Sig. |
| (Intercept) | -0.1706 | -13.76 | *** | 0.035 | 3.42 | *** |
| $D_{\text{MARY}}$ | 0.4367 | 105.43 | *** | 0.415 | 104.09 | *** |
| $F_{\text{AMI}}$ | -0.1407 | -42.54 | *** | -0.122 | -38.66 | *** |
| $F_{\text{Giga}}$ | -0.0421 | -11.07 | *** | -0.063 | -18.70 | *** |
| $S_{\text{Syntax}}$ | 0.0132 | 14.22 | *** | 0.009 | 11.96 | *** |
| $S_{\text{Semantics}}$ | 0.0246 | 5.89 | *** |  |  | *** |
| $D_{\text{MARY}}$:$F_{\text{AMI}}$ | -0.0139 | -6.12 | *** |  |  | *** |

Table 6: Linear mixed effects models for spoken word durations in the AMI corpus, for native as well as non-native speakers of English separately. The models include both syntactic and semantic surprisal as a fixed effect, and a random intercept and slope for $D_{\text{MARY}}$ and $S_{\text{Semantics}}$ under speaker.

researchers are discussing the design of a remote control, but where it is not necessarily the case that these people discuss remote controls very frequently. Furthermore, none of the speakers were present in the whole corpus, and most of the $> 600$ speakers participated only in very few meetings. This means that the in-domain language model strongly over-estimates people's familiarity with the domain.

Words that are highly predictable for the in-domain model (but which are not highly predictable in general) were not pronounced faster, as evident in our first analysis. When semantic surprisal is however estimated based on a more domain-general text like Gigaword, we find a significant positive effect of semantic surprisal on spoken word durations across the complete spectrum from very predictable to unpredictable words.

These results also point to an interesting scientific question: to what extent to people use their domain-general model for adapting their language and speech production in a specific situation, and to what extent do they use a domain-specific model for adaptation? Do people adapt during a conversation, such that in-domain models would be more relevant for language production in situations where speakers are more versed in the domain?

## 10 Conclusions and future work

We have described a method by which it is possible to connect a semantic level of representation (estimated using a distributional model) to observations about speech patterns at the word level. From a language science or psycholinguistic perspective, we have shown that semantic surprisal affects spoken word durations in natural conversational speech, thus providing additional supportive evidence for the uniform information density hypothesis. In particular, we find evidence that UID effects connect linguistic levels of representation, providing more information about the architecture of the human processor or generator.

This work also has implications for designers of speech synthesis systems: our results point towards using high-level information about the rate of information transfer measured in terms of surprisal for estimating word durations in order to make artificial word pronunciation systems sound more natural.

Finally, the strong effect of training data domain raises scientific questions about how speakers use domain-general and -specific knowledge in communicative cooperation with listeners at the word pronunciation level.

One possible next step would be to expand this work to more complex semantic spaces which include stronger notions of compositionality, semantic roles, and so on, such as the distributional approaches of Baroni and Lenci (2010), Sayeed and Demberg (2014), and Greenberg et al. (2015) that contain grammatical information but rely on vector operations.

## Acknowledgements

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 “Information Density and Linguistic Encoding”.

## References

Aylett, M. and Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5):3048–3058.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Comput. Linguist.*, 36(4):673–721.

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2014). Fitting linear mixed-effects models using lme4. ArXiv e-print; submitted to *Journal of Statistical Software*.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Demberg, V., Sayeed, A., Gorinski, P., and Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 356–367, Jeju Island, Korea. Association for Computational Linguistics.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Frank, A. F. and Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In Love, B. C., McRae, K., and Sloutsky, V. M., editors, *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 939–944. Cognitive Science Society.

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2013). Word surprisal predicts n400 amplitude during reading. In *ACL (2)*, pages 878–883.

Greenberg, C., Sayeed, A., and Demberg, V. (2015). Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT)*.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language*, 45:229–254.

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

Mitchell, J., Lapata, M., Demberg, V., and Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206. Association for Computational Linguistics.

Mitchell, J. J. (2011). *Composition in distributional models of semantics*. PhD thesis, The University of Edinburgh.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer.

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.

Sayeed, A. and Demberg, V. (2014). Combining unsupervised syntactic and semantic models of thematic fit. In *Proceedings of the first Italian Conference on Computational Linguistics (CLiC-it 2014)*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(379-423):623–656.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.