KNOWING A THING IS "A THING": THE USE OF ACOUSTIC FEATURES IN MULTIWORD EXPRESSION EXTRACTION

BY

CASSANDRA L. JACOBS

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Adviser:

Research Associate Professor Margaret Fleck

Abstract

Speakers of a language need to have complex linguistic representations for speaking, often on the level of non-literal, idiomatic expressions like *black sheep*. Typically, datasets of these so-called multiword expressions come from hand-crafted ontologies or lexicons, because identifying expressions like these in an unsupervised manner is still an unsolved problem in natural language processing. In this thesis I demonstrate that prosodic features, which are helpful in parsing syntax and interpreting meaning, can also be used to identify multiword expressions. To do this, I extracted noun phrases from the Buckeye corpus, which contains spontaneous spoken language, and matched these noun phrases to page titles in Wikipedia, a massive, freely available encyclopedic ontology of entities and phenomena. By incorporating prosodic features into a model that distinguishes between multiword expressions that are found in Wikipedia titles and those that are not, we see increases in classifier performance that suggests that prosodic cues can help with the automatic extraction of multiword expressions from spontaneous speech, helping models and potentially listeners decide whether something is “a thing” or not.

Table of Contents

.

CHAPTER 1
INTRODUCTION

Speakers of a language are tasked with learning a number of different linguistic categories, from sounds, to words, to combinations of words and turns of phrase. The learning mechanism is thought by some to be the same, because speakers show sensitivity to the probabilistic properties of words and phrases. In the first section of this work, I discuss the language-level probabilistic factors that relate to how easily speakers process and produce sequences of words (phrases), as well as the acoustic side effects of this processing fluency in language production. Then, I discuss the ways that computational approaches have historically attempted to identify both literal and idiomatic expressions and propose a method for integrating prosodic information with language model information. In the final section, I present an experiment that integrates prosodic, lexical, and language model features to predict whether a phrase is an expression that exists in an existing ontology (Wikipedia).

1.1 Human language processing of sequences of words

Combinations of words can form expressions ranging from the literal (*strong coffee*) to the more idiomatic (*black sheep*). Expressions are a subset of a class called phrases, which includes novel combinations of words whose meanings are totally predictable from their parts like *yellow umbrella*. Despite this breakup, others have typically considered expression like *strong coffee* to be similar to phrases like *yellow umbrella*. Some have argued that literal expressions like strong coffee could be composed as needed, rather than needing to be stored in memory (Pinker, 1998). By contrast, figurative language is

stored in long-term memory precisely because the meaning is unpredictable. This proposal has generally been called the *words and rules hypothesis*.

There is a certain amount of evidence to suggest that the words and rules hypothesis may not be correct in its strictest form. Evidence that even compositional phrases are stored in long term memory comes from frequency effects, where common phrases are more easily processed than less common phrases (Janssen & Barber, 2012; Jacobs et al., 2016; Bannard & Matthews, 2008; Smith & Levy, 2013; Siyanova-Chanturia et al., 2011). The facilitatory effects of phrase frequency are seen at nearly every level of linguistic representation, from comprehension to production and language acquisition. Frequent phrases are safer from error (Bannard & Matthews, 2008; Choe & Redford, 2012), are more easily recognized (Arnon & Snider, 2010), and are better remembered (Tremblay & Baayen, 2010; Jacobs et al., 2016). These effects overall suggest that even though these word combinations could be created and understood without relying on existing phrase representations, speakers are sensitive to how common those phrases are in their language.

While there is still some debate about what it means for a phrase to be represented in long-term memory, there is a wealth of evidence to suggest that combinations of words that are more common, regardless of whether they are literal or idiomatic, are produced in a slightly different way from novel combinations of words. In the next section I discuss the acoustic properties of these sequences of words.

1.2 Acoustic properties of combinations of words

Word combinations differ from words by virtue of belonging to a broader prosodic phrase. Generally speaking, the way that combinations of words differ from individual

words is the constraints placed on them during sentence production, namely the prosodic realization of a phrase. Prosody may be treated as anything beyond the individual sounds or segments within a word, such as word duration, pitch in non-tone languages, or amplitude (volume).

Prosodic phrases can be defined perceptually by the concept of a phrase boundary, which occurs naturally in between syntactic units (Streeter, 1978; Scott, 1982), and which helps listeners parse sentences (Schafer et al., 2000; Schafer et al., 2005; Frazier et al., 2006; Milotte et al., 2008; Cole et al., 2010). Words produced in multiword utterances differ from words produced in isolation because they are often shorter and less hyperarticulated, especially when they occur in predictable contexts (Bell et al., 2009; Bell et al., 2003; Arnon & Cohen Priva, 2013; Howell & Kadi-Hanifi, 1991; Silverman et al., 1992; Seyfarth, 2014).

There are both local effects of language structure on word production and hierarchical effects. Familiar expressions often being shorter in duration than expected (Arnon & Cohen Priva, 2013; Strik et al., 2007). Like idiomatic expressions, compound words are thought to function as their own lexical items, even when they are relatively literal (e.g. *blackbird* or *outdoor*), with a substantial reduction in duration and a shift in stress that demonstrates that the two words are a compound (Farnetani et al., 1988; Saon & Padmanabhan, 2001; Plag et al., 2008).

A number of studies have looked into the acoustic realization of frequent phrases, but relatively few have been able to definitively analyze multiword expressions, particularly because identifying them is a non-trivial computational task, especially in light of the weak consensus as to the meaning of the term (Baldwin, 2006; Sag et al.,

2002). Previous studies focusing on the classification of multiword expressions into such categories have achieved mixed results, which are summarized below.

1.3 Challenges in multiword expression extraction

Multiword expressions (MWEs) are the blanket term used to describe a number of different types of word combinations, from relatively fixed but literal expressions like *strong coffee* to metaphorical or idiomatic expressions like *kick the bucket*. The challenge for the field of natural language processing has been to automatically identify these expressions in an unsupervised manner. While it is possible to extract MWEs by rule using a lexicon or database, these sources are usually expensive and incomplete, often serving a didactic purpose as for second language acquisition (Ellis, 2002).

The first challenge for a purely unsupervised approach to MWE extraction is in defining the term itself. Typically, researchers start with compositional expressions, which are those expressions that are roughly predictable from their parts (e.g. *strong coffee*), but for whatever reason occur more often than another similar hypothesized combination (e.g. *powerful coffee*). Successfully identifying sufficiently similar but "surprisingly common" phrases requires deciding in advance what it means for two words to be related, as well as whether substitution of one word for another results in "the same" basic meaning. This problem comes with a number of computational and representational challenges (e.g. Yazdani, Farahmand, & Henderson, 2015; Mitchell & Lapata, 2008; Yu & Dredze, 2015; Baroni & Zamparelli, 2010). First of all, solving the meaning-based approach to multiword expression extraction would require knowing that two words were synonyms, calculating some sort of distributional similarity metric, or

simply making a decision between vector-like or matrix-like representations of words and how they combine to form expressions or phrases.

One additional contributing factor is that it is not obvious that the compositionality of phrases is perceived categorically. Furthermore, speakers do not typically make such categorization judgments reliably or require considerable training to reliably rate phrases for their compositionality (Jacobs et al., 2016; Ramisch, personal communication; Mitchell & Lapata, 2010; Wieting et al., 2015; Williams et al., 2015).

Among computational methods for the induction of phrases, many rely heavily on a measure known as mutual information, which combines the probabilities of each of the words with the phrase probability. Phrases that occur more often than would be expected by the independent probabilities of the words occurring together (e.g. myocardial infarction) are considered to have high mutual information. Indeed, even more complex phrase-based tasks such as analogy completion or topic mining often start with high mutual information scores as their cutoff for what constitutes a multiword expression, rather than learning this category boundary explicitly (Yu & Dredze, 2015; Mikolov et al., 2013; Passos et al., 2014; Liu et al., 2015; El-Kishky et al., 2014).

One potential source of multiword expressions comes from existing ontologies of phrases such as titles on Wikipedia or WordNet, making the task no longer explicitly about unsupervised inference, but instead about classification. The advantage to using these resources is that they have been implicitly validated by a large community of researchers or editors, eliminating the need for human classification. One downside to this is that compositionality is completely ignored. Literal and idiomatic MWEs are treated in these tasks as equivalent, or would require additional annotation by researchers.

There is some precedent for using an existing ontology or lexicon to verify the results of phrase induction. For example, in analogy completion the training and test examples are constructed by hand (Mikolov et al., 2013). Similarly, models may be given phrases as input directly after extracting them from a lexicon, rather than having them be learned implicitly from sequences of words (e.g. Williams et al., 2015).

Wikipedia holds a number of advantages over WordNet for the study of multiword expressions. Wikipedia is an online encyclopedia that is generated entirely by editors, users, and which is typically free for any user to edit and is considered to be the largest freely available encyclopedic resource. Just as listeners must identify whether something they are hearing is a thing or not, the task of Wikification (Mihalcea & Csomai, 2007) takes strings as queries (e.g. "François Hollande, the President of France") and attempts to return the matching Wikipedia page for that string (e.g. François_Hollande). Despite the difficulties associated with this task, successfully identifying a matching Wikipedia page can improve performance on many tasks, such as coreference resolution. Wikipedia has been a crucial facilitator in providing an ontology for tasks in natural language understanding, especially in the closely related tasks of named entity recognition and coreference resolution.

As noted before, Wikipedia cannot distinguish between compositional expressions and non-compositional ones like *strong coffee* and *black sheep*. Another shortcoming of Wikipedia is that it contains ambiguous phrases like *White House* and *black cat*. While it is not entirely known whether phrases like *White House* influence the pronunciation of phonologically and lexically identical compositional phrases like *white house*, this should only increase the challenge of such a classification task. Some prior

research has suggested that this might be the case (Seyfarth, 2014; Dell, 1990). As such, the classification task in this experiment may be more difficult than speakers of a language typically encounter. At the same time, differentiating between ambiguous phrases *White House* and *white house* is also a source of difficulty for speakers and therefore classifiers, though this is outside the scope of the current work.

Importantly, Wikipedia does not contain compositional phrases like *green banana*, which are reasonable descriptions of real-world objects, but which are not themselves categories. Because these can be thought of as a proxy for non-unitary concepts, their acoustic properties should differ from those of phrases that are found in Wikipedia, such as *black sheep*.

Both experiments use data coming from the Buckeye corpus (Pitt et al., 2005), which contains monologues produced by 40 different native speakers of American English totaling approximately 300,000 words, each of which has been time-aligned on the phone and word level. Every speaker produced one to four monologues. One variant of the Buckeye corpus contains words along with their phonetic durations and part of speech tags. The corpus was tagged instead automatically using the C&C parser (Clark & Clark, 2003), achieving a reported 90% accuracy on the corpus. Ultimately this work focuses on the relationship between word frequency and phonetic duration in potential noun phrases as obtained by the Buckeye part of speech tags, and then transitions into the phonetic and lexical properties that distinguish noun phrases that can be matched to a title in Wikipedia versus those that cannot. A model that can integrate phonetic and prosodic features into identifying multiword expressions demonstrates the usefulness of at least one method for inducing multiword expression lexicons in the absence of supervision.

CHAPTER 2

PRELIMINARY EXPERIMENT

This study aimed to demonstrate that representations that are relatively atomic on a lexical level tend to have particular patterns of reduction. As discussed before, a statistic known as mutual information has served as a surprisingly reliable metric for identifying phrases or multiword expressions. The study here focuses on whether mutual information is predictive of acoustic duration beyond phrase frequency effects.

2.1 Data and preprocessing

We extracted all noun-noun (NN-NN) two-word pairs that the Buckeye database contained, which resulted in 4191 word pairs for analysis. For each word in a pair, we calculated the duration of the word, the expected duration of the word given all other tokens of that word outside of that word pair (e.g. the duration of the *sealed* in *hermetically sealed* is a function of all other observations in which *sealed* comes after another word), and frequency-related information.

Calculating the expected duration of a word removes all phrases with only a single observation in Buckeye, which would normally have posed a problem for estimating the duration of that phrase due to unreliable estimates. This calculation also excludes phrases which are composed of words that also only occur once, as well as phrases where the words always co-occur. After this filter, 3596 observations remained, with 959 unique first nouns and 921 unique second nouns.

2.2 Analysis

To account for the multiple observations of a word within a phrase as well as words occurring in many different phrases, as well as for the likelihood that some phrases would be more sensitive to reduction, the analysis below is a multilevel regression model on the amount of error on a given phrase after accounting for the average durations in the observed phrase duration. This amount was the sum of the observed durations of the two words in each noun-noun phrase minus the sum of the average durations of the two words.

The frequency of a phrase could be a better predictor than a different measure like mutual information, so the model fit was compared between a model that tested for the effect of frequency relative to the effect of mutual information. To compare these two models, the REML criterion at convergence was used. While the estimation of the effect of phrase frequency on phonetic duration reduction was larger, the higher probability model was the one testing for the influence of mutual information on duration, as evidenced by a lower REML criterion (β = -0.02, SE = 0.006, t = -3.66 versus β = -0.007, SE = 0.003, t = -2.26 having REML criteria of -1343 versus -1334 respectively). However, the relationships are relatively weak. This is demonstrated in part by the plot below in Figure 1, which shows the relationship between mutual information and phonetic reduction.

Given the relatively weak and linear relationship between a common metric used to identify collocations and duration reduction, the next experiment focuses on more fine-grained distinctions between phrases in an ontology and those outside of one.

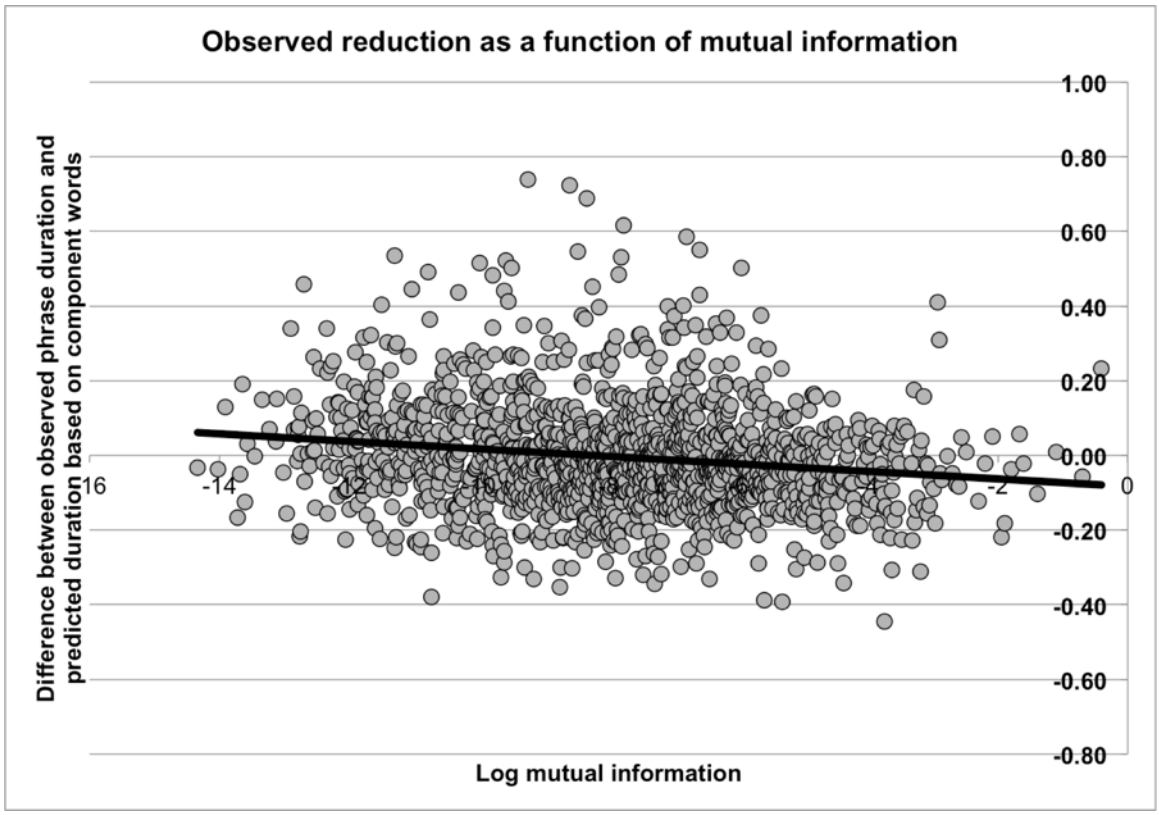

Figure 1: Mutual information is defined as the relative frequency of a pair of events (p(ab)) beyond what would be expected if the two events occurred together at chance (p(a)*p(b)). The higher the mutual information score, the shorter the phrase is in duration relative to what would be expected from the durations of those two words in every other context**.**

CHAPTER 3

EXPERIMENT

3.1 Data and preprocessing

This experiment was conducted using data from the Buckeye corpus (Pitt et al., 2005). Speech in the Buckeye monologues is broken up by questions from an interviewer, silence, coughs, noise, and unintelligible words. After removing these from the transcripts, we obtained only English words broken up into a single sentence stream without punctuation. Using the Illinois Chunker (Punkyanok & Roth, 2001), we obtained shallow parses of the monologues. This was done in part because shallow parsers are more robust to noise introduced by filled pauses and breaks (Osborne, 2002). Though is possible to obtain estimates of recall in principle via sampling and hand labeling, we instead focus below on an analysis of the chunker's noun phrase precision.

Many of the noun phrases that the chunker extracted contained errors (e.g. ***NP*** ***NN*** *Mom* ***NNS*** *lives*). Many of the errors described below can be attributed to differences in the training and test data that the Chunker uses. The Illinois Chunker relies on the Wall Street Journal corpus within the Penn Treebank data (Marcus et al., 1993), which is a marked departure from spoken language. The most relevant part of speech tags and chunking performance would most likely come from the Switchboard corpus (Godfrey et al., 1992), which is conversational instead of a monologue setup, but would have required retraining the shallow parsing model provided.

The training and test sets contain these errors due to the amount of labor needed to reclassify these examples by hand. To better understand how errors may have been introduced into the noun phrases that were included in the dataset, I analyzed the first 300

noun phrases that the parser identified. Among the first 300 noun phrases identified by the chunker, 74 were not complete noun phrases, with 33 not being noun phrases at all. The other 44 contained noun phrases but were parsed incorrectly. These were categorized as partial noun phrases, and either contained a noun phrase that was not properly chunked into its own NP, or contained the beginning of a noun phrase but not all of it, potentially with additional material.

Approximately 30% of all noun phrases identified were pronouns, which are actually titles in Wikipedia, but which are excluded after eliminating stop words. Below in Table 1 are examples of noun phrases that the chunker identified that were complete and incomplete.

Non-NPs (10%)	**Partial NPs (25%)**	**Non-pronoun NPS (30%+35%)**
NP JJ okay . . **NN** um	**NP DT** the **JJ** national **NN** news **FW** I	**NP JJ** high **NN** school **NNS** sweethearts
NP EX there **IN** that	**NP UH** uh **NN** um **JJ** different **NNS** bombings	**NP NN** ohio **NN** state **NN** football
NP DT an **RB** just	**NP JJR** more **NN** authority **FW** I	**NP RB** so **JJ** much **NN** negativity
NP DT all **RB** so	**NP DT** the **NN** hockey	**NP JJ** sunday **NNS** afternoons

Table 1: Noun phrases identified by the chunker. Of the noun phrases that did not match any titles in Wikipedia, 10% were not noun phrases, 25% contained noun phrases, and 30% of noun phrases were pronouns. The remaining 35% of noun phrases were complete noun phrases that were not found in Wikipedia titles.

The errorful noun phrases identified by the parser also demonstrated weaknesses in the part of speech tagging pipeline, which is well-documented in even noisier sources of corpora such as Twitter (Gimpel et al., 2011; Derczynski et al., 2013). For example, because "I" is consistently lowercased in Buckeye, it is tagged by the parser as a foreign word (FW), rather than as a pronoun. Additionally, the chunker fails to properly tag filled pauses; while "uh" is consistently tagged as a filled pause (UH), "um" is instead tagged as a noun (NN).

After additional processing to remove filled pauses (*um*, *uh*, *er*) and stop words including discourse markers (e.g. *I mean*, *yeah*, *wow*), noun phrases were then compared to Wikipedia titles. In some cases, the additional processing that removed stop words, especially disfluencies and pronouns, corrected the errors the parser initially had made (e.g. Table 1, column 2).

Noun phrases were not stemmed or lemmatized because a wealth of evidence suggests that there are acoustic differences between different word forms as a function of lemma frequency, even when the majority of the sounds or meaning are identical (Bien et al., 2005; Pluymaekers et al., 2005; but see Roelofs & Baayen, 2002; Gahl, 2008). After removing stop words, all noun phrases were constrained to be only two content words to be more analogous to work done in the preliminary experiment. These two-word noun phrases were then checked against titles in Wikipedia.

3.2 Wikification

The process of identifying matching Wikipedia titles was a naïve one compared to other approaches (e.g. Cheng & Roth, 2013). Each noun phrase extracted in the previous section was compared as a literal string match to all two-word Wikipedia titles. We

lowercased and removed underscores between words for each Wikipedia title (e.g. *African_American* is transformed into *african american*). The resulting string comparison left us with three classes of noun phrase. First of all, items that were not in Wikipedia could either be proper noun phrases (e.g. *better countries*) or could contain parser errors because of stop word removal or otherwise. These parser errors could have, if the stop words had been retained, continued to be noun phrases (e.g. *geometry (and) algebra*). Below in Table 2 is a set of three different types of noun phrases that surfaced after comparing all extracted two-word noun phrases from Buckeye with Wikipedia titles.

Not in Wikipedia (NP)	**Parser errors**	**In Wikipedia**
individual person	(a) lot looser	african american
methodist religion	sex (and) drugs	premarital sex
polaris area	things (in) columbus	catholic religion
right word	geometry (and) algebra	thirty three
many things	just sign	one year
another thing		united states
tremendous impression		
better countries		

Table 2: Example noun phrases extracted by the shallow parser with stop words removed that were found in Wikipedia and those that were not. Some phrases from the first column that were not found in Wikipedia are erroneously counted as noun phrases (e.g. “lot looser”) but others would have been correct noun phrases, albeit not in Wikipedia (e.g. “geometry [and] algebra”).

3.3 Features

Table 3 clarifies the names of each of the variables that were entered into the model. Figure 2 plots the correlations between each of the variables used in classification.

Stop words. We considered that having removed certain stop words could sometimes prevent matching noun phrases with a Wikipedia title, such as those where a central *and* was said and then removed (e.g. *sex and drugs*). To take this into account, the model includes bag-of-words features of these removed stop words. In addition, an additional feature represented the total number of all stop words that were removed in each of the phrases.

Lexical features. We calculated the frequencies of all words, and all bigrams omitting stop words across the entire Buckeye corpus. For classification, these features are log transformed.

Prosodic features. We calculated durations for each of the individual words, whether each word preceded a pause, whether each word followed a pause, the durations of the last three phonemes of each of the two words, and the average durations of the individual words. For classification, these features are log transformed.

Correlations. Figure 3 demonstrates the correlations between all of the non-lexical variables, including their correlations with the dependent measure of interest, which was whether the observed phrase was in an ontology or not. The strongest relationships are frequency and duration (more frequent words are shorter), while phrase frequency is strongly related to the durations of the words within the phrases. Upon visual inspection, phrase frequency appears to be the strongest indicator as to whether a phrase will be in Wikipedia or not.

Variable Name	Variable definition
inWikiOrNot	The variable we are predicting in the classifier. If a phrase is in Wikipedia, then this value is set to 1. Otherwise, it is 0.
precedesPause	Values from 0 to 2. If the first or word precedes a prosodic phrase boundary (w1 uh w2) or (w1 w2 uh), then 1 or both (w1 uh w2 uh), then 2.
durW1P1	The duration of the second to last phone of the first word of a phrase in milliseconds.
durW1P2	The duration of the penultimate phone of the first word of a phrase in milliseconds.
durW1P3	The duration of the last phone of the first word of a phrase in milliseconds.
durW2P1	The duration of the second to last phone of the second word of a phrase in milliseconds.
durW2P2	The duration of the penultimate phone of the second word of a phrase in milliseconds.
durW2P3	The duration of the last phone of the second word of a phrase in milliseconds.
durW1	The duration of the first word in milliseconds.
durW2	The duration of the second word in milliseconds.
avgW1Dur	The average duration of the first word in all other linguistic contexts.
avgW2Dur	The average duration of the second word in all other linguistic contexts.
freqW1	The number of times the first word occurs in the Buckeye corpus.
freqW2	The number of times the second word occurs in the Buckeye corpus.
freqPhrase	The number of times the first and second words occur together in the Buckeye corpus.

Table 3: Variable names and definitions plotted in Figure 3.

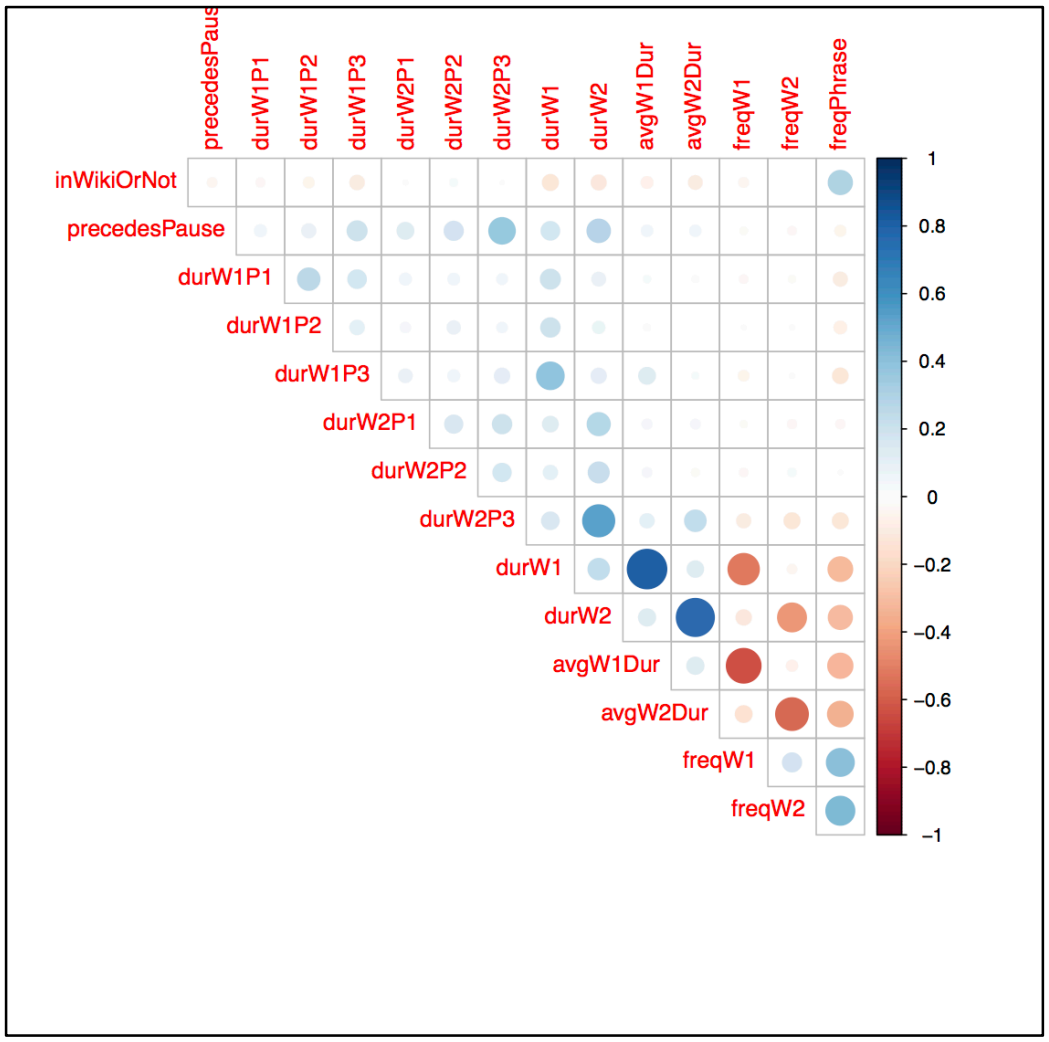

Figure 2: Plot of the correlations between each of the non-categorical variables in the automatically extracted noun phrase data. The strongest correlations are between how frequent the words are and how frequent the phrase is, as well as how short the words are relative to other words.

Generally speaking, the relationship between phrase frequency and phrase duration is much more tenuous than the relationship between word frequency and word duration. To demonstrate this relationship, Figure 3 shows how increasing word frequency leads to

significant reductions in phonetic duration, at points leading to a 300 millisecond difference in durations at the two extremes.

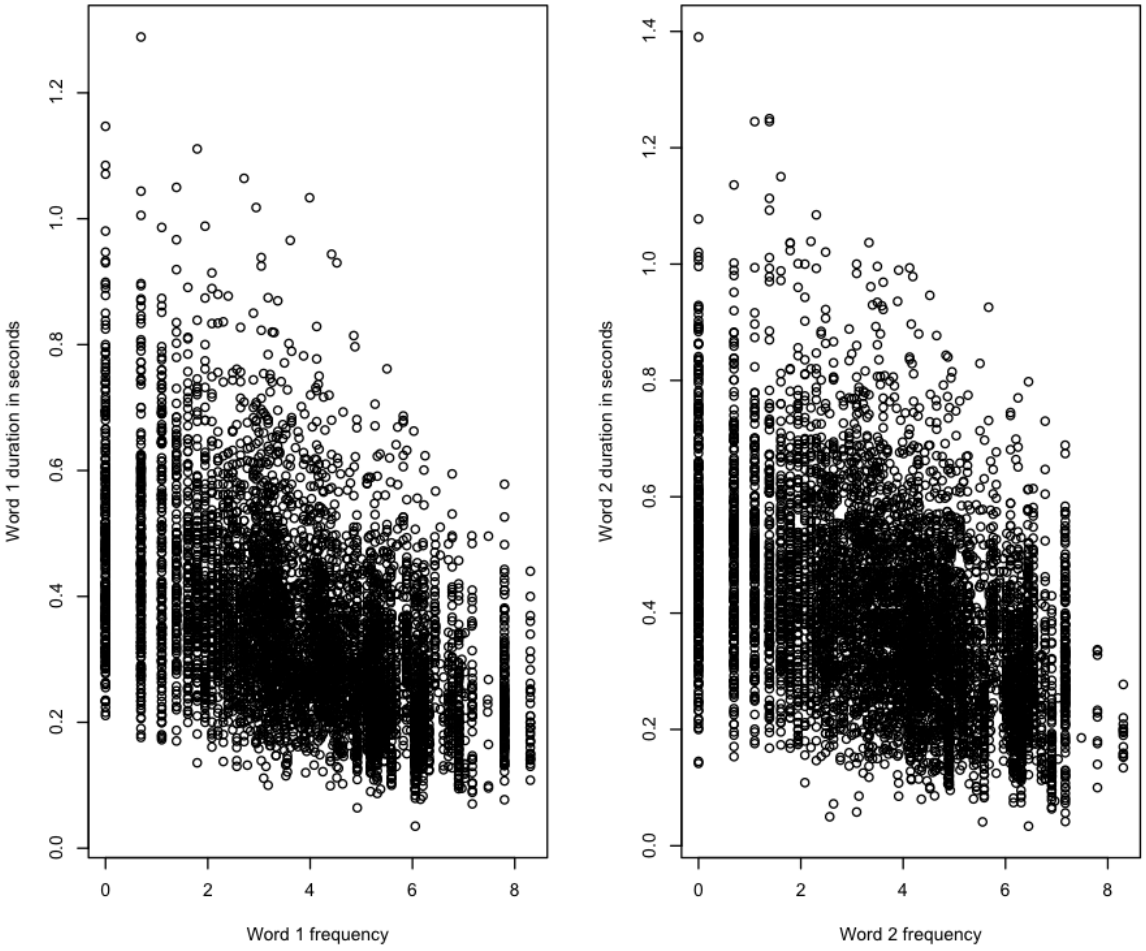

Figure 3: For both words within each two-word noun phrase, the more common the word is, the shorter it is in duration, though a number of other factors not plotted here also play a role in word duration. The most common words are almost 300 milliseconds longer than the least common words in the dataset. This pattern of reduction is similarly true for both the first word and the second word.

3.4 Classifier

The classifier here is attempting to distinguish between noun phrases that match a phrase in Wikipedia titles versus those that are not in Wikipedia titles, making this a two-class classification problem. For this, we use logistic regression and a ten-fold cross-validation scheme with random partitions of the data. For each fold, 90% of the data was used during training, and the other 10% during test. Results reported below are averages of each of the 10 test folds. We trained four general models:

1. The first model contained only pause and stop word features.
2. The intermediate models contained only one of the below sets of features:
    a. lexical and phrase frequency measures or
    b. additional prosodic features as described above.

The final model contained all of these variables simultaneously.

CHAPTER 4

RESULTS

Table 4 summarizes the performance of the different models tested. To perform above chance, the model would need to beat the majority class label (not in Wikipedia), which made up 66.7% of all identified noun phrases. The worst performing model contained only lexical information about the stop words that occurred in the noun phrases that the chunker included as well as the cue of whether the phrase preceded a pause or not, and performs worse than majority class. Beyond that, the model containing only prosodic information performed slightly better, but adding frequency information to the first model resulted in a gain in classifier performance. The best model contained pause features, prosodic features, frequency features, and stop word features. This model correctly classified 74.2% of test examples.

Model	Accuracy
Majority class	0.667
Pause + stop word features	0.660
+ Prosodic features	0.675
+ Frequency features	0.731
All features	**0.742**

Table 4: With increasing numbers of features, the model's performance increases.

Importantly, despite the clearly informative nature of all of the prosodic variables, one of which is theoretically unknown to a listener who is attempting to categorize a phrase into being in an ontology or not (phrase frequency), acoustic cues add additional signal that can help differentiate between phrases that are in Wikipedia from those that are not.

In the final model, the feature weights that were most strongly related to classifier decisions were both prosodic and lexical in nature. For both words in a two-word phrase, the lower frequency each word is, the more likely that word is to be in an ontology, suggesting that Wikipedia phrases tend to contain low frequency words. Additionally, relatively high frequency phrases tend to be in Wikipedia as titles. However, listeners do not necessarily know how common a phrase is the first time they hear it, so this variable may not be as informative in practice during human language comprehension.

Among the prosodic measures, some aspects of duration played a significant role in classifier performance. Typically, the phones at the end of the second word tended to be predictive of whether a phrase was in an ontology. For both the penultimate and the final phone, the longer the phone, the more likely the phrase was to be in Wikipedia. Similarly, the observed duration of the two words of each phrase strongly predicted whether the phrase was in Wikipedia. The shorter each of the words was, the more likely that phrase was to be in Wikipedia. Other prosodic information such as the final phone durations of the first word and whether the phrase preceded a pause or not were less informative. The results of this analysis are reported below in Table 5.

Stop word features, which seem to greatly influence the parser, did affect classification accuracy. *And*, *his*, *the*, *their*, and *your* were all significantly predictive of phrases being in Wikipedia. Other stop words such as *how* and *that* were more predictive of phrases not being in the ontology.

Feature	*r*	*SE*	*p*
Precedes pause	-0.06	-0.07	n.s.
Word 1 phone -3	0.12	-0.07	= 0.09
phone -2	-0.01	0.06	n.s.
phone -1	-0.05	0.07	n.s.
Word 2 phone -3	0.08	0.07	n.s.
phone -2	**0.26**	**0.06**	**< .001**
phone -1	**0.33**	**0.07**	**< .001**
Word 1 observed duration	**-2.53**	**0.51**	**< .001**
Word 2 observed duration	**-1.98**	**0.44**	**< .001**
Word 1 average duration	0.68	0.56	n.s.
Word 2 average duration	0.03	0.50	n.s.
Word 1 frequency	**-0.29**	**0.02**	**< .001**
Word 2 frequency	**-0.24**	**0.02**	**< .001**
Phrase frequency	**0.95**	**0.04**	**< .001**

Table 5: The most significantly predictive features of whether a phrase belonged to an ontology were frequency related variables. Generally, the less common the words were relative to the phrase, the more likely that phrase was to appear in Wikipedia. Despite this, prosodic variables like word durations provided additional predictive power.

The differences between the best and worst models were substantial. Typically, the previous model predicted 98.4% of test items to be in the majority class (not in

Wikipedia), resulting in an overall F1 score of .02. By adding prosodic and lexical features, the model predicted 21.4% of test items to be in Wikipedia, a substantial change, where F1 increased to .52. These class predictions between the worst model and the best model are below in Table 6.

	Was not in wiki	**Was in wiki**
Predicted not in	**60.3** -> 65.6	**18.3** -> 32.8
Predicted in	**7.4** -> 1.1	**14** -> 0.5

Table 6: Classification performance was significantly improved on all measures by adding in additional features containing information about the prosodic structure of the test phrases as well as lexical features such as word and phrase frequency. When these features are included, the model correctly extracts more expressions that are found in Wikipedia. The bold values demonstrate the performance of the better model.

CHAPTER 5

DISCUSSION AND FUTURE DIRECTIONS

The results of the experiment presented here demonstrate that acoustic and lexical features are critical for being able to identify phrases. When an addressee is listening to another person who may be introducing novel concepts into the conversation, they are faced with uncertainty about the things their interlocutor is saying. If a model can leverage very simply prosodic features and lexical frequency to correctly distinguish between phrases and non-phrases, then listeners may also be able to do this.

One interesting question that arises is why the prosodic features have any predictive value at all. Previous research has found that speakers are affected by the ease at which they can put words together, with familiar sequences being much easier to complete (Bannard & Matthews, 2008; Arnon & Cohen Priva, 2013). The fact that specific kinds of prosodic features are predictive of a phrase being in an ontology suggests that phrases have somewhat special representations.

In a number of domains of natural language processing, phrases are given similar representations to words, in that they are more complex, but still atomic linguistic events that happen to have relational features tied to the words that compose them (e.g. a phrase like *strong coffee* would have links to *strong* and *coffee*). In fact, many models treat phrases and their words hierarchically in what is known as *smoothing*. This is necessary to avoid problems with maximum likelihood estimates of phrase probabilities, which would assign 0 probability to all unique phrases, even when the words could reasonably co-occur (e.g. *individual person*).

There are a number of shortcomings to the results presented here. First of all, the noun phrases were extracted using a shallow parser. Shallow parsers tend to rely substantially on ontologies and word co-occurrences. It is therefore unsurprising that the shallow parser fails to identify noun phrases correctly, often containing phrases within them that would potentially be within an ontology. It is possible that the model's performance was so high after including prosodic features in part because a number of the phrases that were not matched to a Wikipedia title may have been erroneously left out (e.g. *sex and drugs*). Additionally, many of the "noun phrases" the parser identified contained errors, which are partly due to the differences between the training and test domain. All positive examples, however, were matched to a page in Wikipedia without error. As such, the task may simply have been distinguishing between correctly parsed noun phrases and potentially incorrectly parsed ones. Future work should eliminate all non-noun phrases extracted by the shallow parser.

Secondly, the model does not distinguish between compositional and non-compositional expressions. The phrase "White House" is idiomatic, referring to the American presidential building, but the phrase "catholic religion" is literal. Because the items are never treated differently by the model, regardless of their semantic compositionality, it is possible that acoustic differences such as the ones found here are the fault of differences in compositionality, where the positive examples actually belong to two categories, making the classification task more difficult. Currently it is unknown whether Wikipedia titles vary substantially in their compositionality, or whether most phrases in Wikipedia titles are non-compositional in nature. Future studies should focus on whether compositional expressions have different acoustics, and whether there are

informative cues to whether a phrase is non-literal or not, as has been explored in sarcasm detection (Tepperman et al., 2006; Rosenberg, 2009). Related to these studies, a more definitive study would look at finer grained phonetic features than simply segment and word durations.

Finally, a major shortcoming of this study is that, despite focusing on the cognitive representations that lead to specific phonetic events, it has not yet been determined whether phrases in an ontology are actually treated differently in the brain than ones that are not. In addition to this, even if a model can take advantage of these acoustic features for classification, it is not clear that speakers can accomplish such a task, or that they would often need to.

Altogether, the results of this study suggest that prosodic cues can affect the interpretation and classification of an unknown noun phrase into one that is "a thing" versus one that is not. When "White House" is said sufficiently quickly, we can infer that it is referring to where Barack Obama lives. Additional research needs to be conducted to identify whether semantic factors like compositionality, syntactic factors like the success of shallow parsers on continuous, unpunctuated spontaneous speech, or other acoustic factors may be at play before concluding that the acoustics reflect any cognitive truths. At the same time, the statistical tendencies presented here demonstrate that there are some cues available to listeners, if listeners were to attend to them.

References

Arnon, I., & Priva, U. C. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, *56*, 349-371.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, *62*, 67-82.

Baldwin, T. (2006, July). Compositionality and multiword expressions: Six of one, half a dozen of the other. In *Invited talk given at the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, July*.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning the effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241-248.

Baroni, M., & Zamparelli, R. (2010, October). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1183-1193). Association for Computational Linguistics.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, *113*, 1001-1024.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*, 92-111.

Bien, H., Levelt, W. J., & Baayen, R. H. (2005). Frequency effects in compound production. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 17876-17881.

Cheng, X. & Roth, D. (2013). Relational Inference for Wikification. In Proceedings of EMNLP, pages 1787– 1796.

Choe, W. K., & Redford, M. A. (2012). The distribution of speech errors in multi-word prosodic units. *Laboratory Phonology*, *3*, 5-26. doi:10.1515/lp-2012-0002

Cole, J., Mo, Y., & Baek, S. (2010). The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes*, *25*, 1141-1177.

Clark, S., & Curran, J. R. (2003, July). Log-linear models for wide-coverage CCG parsing. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 97-104). Association for Computational Linguistics.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*, 313-349.

Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013, September). Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *RANLP* (pp. 198-206).

El-Kishky, A., Song, Y., Wang, C., Voss, C. R., & Han, J. (2014). Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, *8*, 305-316.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in second language acquisition*, *24*, 143-188.

Farnetani, E., Torsello, C. T., & Cosi, P. (1988). English compound versus non-compound noun phrases in discourse: An acoustic and perceptual study. *Language and Speech*, *31*, 157-180.

Frazier, L., Clifton, C., & Carlson, K. (2004). Don't break, or do: prosodic boundary preferences. *Lingua*, *114*, 3-27.

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, *84*, 474-496.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., ... & Smith, N. A. (2011, June). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* (Vol. 2, pp. 42-47). Association for Computational Linguistics.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992, March). SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on* (Vol. 1, pp. 517-520). IEEE.

Howell, P., & Kadi-Hanifi, K. (1991). Comparison of prosodic properties between read and spontaneous speech material. *Speech Communication*, *10*, 163-169.

Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. *Journal of Memory and Language*, *87*, 38-58.

Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. *PloS one*, *7*, e33202.

Liu, J., Shang, J., Wang, C., Ren, X., & Han, J. (2015, May). Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (pp. 1729-1744). ACM.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, *19*, 313-330.

Mihalcea, R., & Csomai, A. (2007, November). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 233-242). ACM.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Millotte, S., Wales, R., & Christophe, A. (2007). Phrasal prosody disambiguates syntax. *Language and Cognitive Processes*, *22*, 898-909.

Mitchell, J., & Lapata, M. (2008, June). Vector-based Models of Semantic Composition. In *ACL* (pp. 236-244).

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, *34*, 1388-1429.

Osborne, M. (2002). Shallow parsing using noisy and non-stationary training material. *Journal of Machine Learning Research*, *2*, 695-719.

Passos, A., Kumar, V., & McCallum, A. (2014). Lexicon Infused Phrase Embeddings for Named Entity Resolution. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*.

Pinker, S. (1998). Words and rules. *Lingua*, *106*(1), 219-242.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*, 89-95.

Plag, I., Kunter, G., Lappe, S., & Braun, M. (2008). The role of semantics, argument structure, and lexicalization in compound stress assignment in English. *Language*, *84*, 760-794.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, *118*, 2561-2569.

Punyakanok, V. and Roth, D. (2001) The use of classifiers in sequential inference. In *NIPS 13*, pages 995–1001.

Roelofs, A., & Baayen, H. (2002). Morphology by itself in planning the production of spoken words. *Psychonomic Bulletin & Review*, *9*, 132-138.

Rosenberg, A. (2009). *Automatic detection and classification of prosodic events.* Doctoral dissertation, Columbia University.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002, February). Multiword expressions: A pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 1-15). Springer Berlin Heidelberg.

Saon, G., & Padmanabhan, M. (2001). Data-driven approach to designing compound words for continuous speech recognition. *IEEE transactions on Speech and audio processing*, *9*, 327-332.

Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, *29*, 169-182.

Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2005). Prosodic influences on the production and comprehension of syntactic ambiguity in a game-based conversation task. *Approaches to studying world situated language use: Psycholinguistic, linguistic and computational perspectives on bridging the product and action tradition*, 209-226.

Scott, D. R. (1982). Duration as a cue to the perception of a phrase boundary. *The Journal of the Acoustical Society of America*, *71*, 996-1007.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*, 140-155.

Silverman, K., Blaauw, E., Spitz, J., & Pitrelli, J. F. (1992). A prosodic comparison of spontaneous speech and read speech. In *Second International Conference on Spoken Language Processing*.

Siyanova-Chanturia, A., Conklin, K., & Van Heuven, W. J. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword

sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 776.

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

Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception.*The Journal of the Acoustical Society of America*, *64*, 1582-1592.

Strik, H., Hulsbosch, M., & Cucchiarini, C. (2010). Analyzing and identifying multiword expressions in spoken language. *Language Resources & Evaluation*, *44*, 41-58.

Tepperman, J., Traum, D. R., & Narayanan, S. (2006, September). " yeah right": sarcasm recognition for spoken dialogue systems. In *INTERSPEECH*.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. *Perspectives on formulaic language: Acquisition and communication*, 151-173.

J. Wieting, M. Bansal, K. Gimpel, and K. Livescu. (2015). From paraphrase database to compositional paraphrase model and back," *Transactions of ACL*, *3*, pp. 345–358.

Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A., & Spasić, I. (2015). The role of idioms in sentiment analysis. *Expert Systems with Applications*, *42*, 7375-7385.

Yazdani, M., Farahmand, M., & Henderson, J. Learning Semantic Composition to Detect Non-compositionality of Multiword Expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, number September* (pp. 1733-1742).

Yu, M., & Dredze, M. (2015). Learning composition models for phrase embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 227-242.