

# Discourse-Driven Pitch Accent Prediction



Jennifer Moore

Department of Computational Linguistics

Universität des Saarlandes

Thesis submitted for the degree of

*Master of Science*

2009

Advisors

Magdalena Wolska

Bistra Andreeva

---

## Acknowledgements

First and foremost, I would like to thank *Magdalena Wolksa*, without whose encouragement, guidance, and support, from the first nebulous ideas, to the dried ink on the very last page, this thesis would not have been possible. Every conversation with her was a jumpstart to my flagging spirits, in which ideas would spiral one from another, while her feedback always kept me on the straight and narrow. She also has my gratitude for introducing me to *Bistra Andreeva*, whose expertise in phonological systems in both English and German was invaluable for such multi-disciplinary investigations. She directed me to many different areas of research in the field, allowing me to narrow in on a clear set of problems to address, and really strengthen the key points of my research.

Secondly, there are no words to express my indebtedness to *John Hörnqvist*, who was always there for me through the best and worst times, who read every draft, who listened patiently to every idea and every problem. I would not be where I am today without his love and support these past many years.

Finally, I would like to thank the many friends, faculty, and researchers that have contributed to this thesis in various ways: *Dr. Kim Silverman*, who first inspired me with the idea over a cup of coffee at the Apple campus; *Prof. Dr. Manfred Pinkal*, who was my mentor throughout the program; *Michael Roth*, who provided the German translation of this abstract and gave me feedback on some of my early drafts; and *Roland Albrecht*, who proof-read the more mathematical sections of this thesis, among many other things, for which I am eternally grateful.

## Abstract

Pitch accent is a component of prosody that is often used to convey information beyond the intrinsic linguistic meaning of a spoken utterance, such as highlighting words that correspond to important information, or in signaling a contrast with information that was previously conveyed. This information-bearing aspect of pitch accent is therefore important for effective communication in spoken applications.

Recent work has looked into statistical modeling techniques for automatic pitch accent prediction as a component of speech technologies like Text-to-Speech (TTS). Many of these systems, however, have largely overlooked the dimension of discourse context in driving pitch accent placement; others simply introduced more complex models of discourse-level phenomena into the accent prediction component.

We investigate a model for discourse-driven statistical pitch accent prediction that makes use of a dynamically-updated semantic space as a means of introducing context-sensitive features into the prediction model. This approach has the advantage of being trainable on a large corpus of unannotated data, making it less prone to corpus domain bias (i.e. distributions estimated from a given corpus that reflect the genre of that corpus only) inherent in purely probabilistic variables for accent prediction. Moreover, this approach does not require additional modeling of complex discourse processes, but relies solely on shallow analysis of the input text.

## Abstract

Betonungsakzente bilden eine Prosodie-Komponente, die in gesprochener Sprache häufig verwendet wird, um Informationen zu übermitteln, die über intrinsische linguistische Bedeutung hinausgehen. Beispiele hierfür sind das Betonen von Worten, die eine besondere Wichtigkeit haben, oder das Signalisieren eines Kontrasts zu vorherigen Informationen. Dieser informationstragende Aspekt von Betonung ist daher wichtig für das effektive Kommunizieren in sprachlichen Anwendungen.

Neuste Forschungen haben versucht, statistische Modelle zu entwickeln, um automatische Betonungsakzente in sprachtechnologische Anwendungen wie Text-to-Speech-Systeme zu integrieren. Viele dieser Systeme berücksichtigen allerdings gar keinen Diskurskontext, um Betonungsakzente zu setzen; andere Systeme führen komplexe Diskurs-Modelle ein, um Phänomene auf der Ebene in einer Betonungsvorhersage nachzubilden.

In dieser Arbeit untersuchen wir ein Modell zur statistischen, Diskurs-getriebenen Betonungsvorhersage, in dem ein dynamisch aktualisierter semantischer Raum als kontext-sensitives Attribut verwendet wird. Diese Vorgehensweise hat den Vorteil, dass das Modell mit einem großen Korpus unannotierter Daten trainiert werden kann und dadurch weniger anfällig für Domänen-spezifische Tendenzen wird (d.h. die aus einem Korpus berechneten Verteilungen spiegeln stets nur das Genre des Korpus wider). Darüber hinaus ist dieser Ansatz nicht auf ein zusätzliches Modellieren von komplexen Diskursprozessen angewiesen, sondern bedient sich lediglich einer flachen Analyse des Eingabetexts.

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Purpose . . . . .	3
1.2	Method . . . . .	4
1.3	Scope . . . . .	4
1.4	Related Work . . . . .	5
1.5	Overview of the Thesis . . . . .	7
<b>2</b>	<b>Theoretical Foundations</b>	<b>9</b>
2.1	Accent and Prominence in Spoken Language . . . . .	10
2.2	Information and Accent . . . . .	22
2.3	Computational Models of Accent Variation . . . . .	34
<b>3</b>	<b>Model Design</b>	<b>40</b>
3.1	The Stochastic Model . . . . .	41
3.2	Feature Design . . . . .	44
3.2.1	Semantic Space . . . . .	45
3.2.2	Local and Global Discourse . . . . .	52
3.2.3	Discourse Structure . . . . .	55
<b>4</b>	<b>Data Acquisition</b>	<b>57</b>
4.1	Corpora for CRF Training . . . . .	57
4.1.1	The MULI Corpus . . . . .	57
4.1.2	The IViE Corpus . . . . .	64
4.2	Corpora for LSA Training . . . . .	68
4.2.1	The Wikipedia XML Corpus . . . . .	68

<b>5</b>	<b>Experimental Setup</b>	<b>69</b>
5.1	Preliminaries . . . . .	69
5.2	Results . . . . .	73
5.3	Discussion . . . . .	79
<b>6</b>	<b>Conclusions</b>	<b>80</b>
	<b>References</b>	<b>92</b>

# List of Figures

2.1	Spectral analysis and fundamental frequency of the word <i>May</i> . . . . .	11
2.2	Prince’s taxonomy of given/new information . . . . .	31
3.1	A linear-chain CRF . . . . .	43
3.2	Word and document factors as coordinates in a two-dimensional space .	51
4.1	The complete MULI transcription of ‘ <i>Exporte in den Libanon sichert Bonn derzeit nur kurzfristig ab</i> ’ . . . . .	60
4.2	The complete IViE transcription of ‘ <i>We arrived in a limo</i> ’ . . . . .	66



The English teacher W. Jones, on attending a concert in Japan, heard his Japanese host announce, “Tonight, we shall be singing a programme of French SONGS and German SONGS; I am sorry that we shall not be singing any English SONGS.” To which address, the baffled Jones mused, “His grammar was faultless, his pronunciation unambiguous, but since I knew when I bought my ticket that the business of choirs is to sing songs, I wondered, if only for a moment, why he had stressed the word ‘songs’: what was this meant to tell us?”

– *Studies in Culture*

# Chapter 1

## Introduction

Pitch accent, a component of prosody in spoken natural language, can be used to convey information beyond the intrinsic linguistic meaning of a spoken utterance. This form of accentuation in speech is often employed as a means of alerting the listener to new or important information within an utterance, or in signaling a contrast to information which was previously conveyed. In such cases, accent is clearly situated within the context of the immediate discourse, and in the absence of discourse, can even evoke a specific context. For example, the accent in the utterance ‘JOHN went to the store’ presumes a discourse context in which the question ‘WHO went to the store?’ was posed. In this way, accents can be simultaneously information-bearing and discourse-bound.

It is this information-bearing nature of accent that makes it a key element of effective spoken communication, in itself an important goal of human-computer information systems. In many applications of speech technology, proper accenting strategies are not only integral to the correct interpretation of a message, but misplaced accents can cause miscomprehension and confusion on the part of the listener. The classic example is that of an in-car spoken navigation system. In a situation in which the driver comes upon two roads of the same name, but suffixed differently, a qualifying accent on ‘Ocean AVENUE’ would signal to the driver the importance of the street kind, whereas an accent in ‘OCEAN avenue’ might cause the driver to derail to the nearest road of that name.

Unfortunately, while much theoretical work has been done to investigate the dimension of discourse in pitch accent placement, accent modeling for practical applications of speech technology have largely overlooked it. In particular, incorporating informational

accent into general domain-independent speech technologies, such as Text-To-Speech (TTS), is challenging. Intuitively, some representation of the discourse and semantic meaning of the intended output is needed for deriving context-sensitive accent. In modern TTS systems, however, the intended output is typically generated from unrestricted text. Although many systems have been developed for deep linguistic processing of text, to our knowledge, there have been no solutions for adequate modeling of discourse and semantics that scale well for general purpose applications.

Secondly, many state-of-the-art approaches to accent prediction, such as those based on statistical models of accentuation, face another challenge: discourse context is fundamentally dynamic, yet current methods are inherently limited by the fixed nature of a prediction model. From fixed rule sets to parameters estimated from a known distribution, many of these approaches make use of static measures of accentability in order to predict occurrences of accent ultimately bound to a dynamic discourse context.

For these reasons, current systems often fall back on shallow, sentence-bound and word-level indicators of accentability, such as lexical class, word position, or probabilistic measures of collocation and accent likelihood. More recently, some research has attempted to cast certain aspects of discourse into discrete “features” for predicting accent status. These include elements like focus structure and information status, which have been identified in the literature as possible indicators of accent. While this may be a step in the right direction, such attributes nevertheless introduce additional layers of complexity by demanding a model proper of focus and informativeness, many of which have yet to be fully explored.

## 1.1 Purpose

We investigate a model for discourse-driven accent prediction that does not rely on additional complex models of discourse objects. In particular, we explore the use of dynamic, context-sensitive features for predicting accent within the static confines of a statistical model. By introducing predictors of informativeness (the static parameters of our statistical model) in a given discourse as measures of a dynamically-updated semantic space, we hope to incorporate an element of the changing aspect of discourse into our model for context-sensitive accent prediction.

## 1.2 Method

For the purposes of this thesis, we implement a system for automatic discourse-driven pitch accent prediction. In particular, we make use of a statistical model based on Conditional Random Fields (CRFs), a sequence-learning discriminative method for the probabilistic prediction of a sequence of accent labels given a sequence of words. To investigate discourse-driven accent prediction for domain-independent input, we introduce a number of features into the model as predictors of accent designed to capture elements of a dynamically-updated discourse context.

Our model is then trained on corpora of transcribed speech that have been manually annotated with pitch accent labels at the word level. We test our model on held-out annotated data through seven-fold cross-validation, and present our results using measures of precision, recall, and an F1 score – a combined measure of precision and recall – as an average over seven runs.

## 1.3 Scope

We are primarily interested in investigating potential semantic and discourse variables that are predictors of accent status. To this end, we focus on automatic pitch accent prediction in German and English, two languages which exhibit discourse-sensitive prosodic accentuation. We will not be concerned with modeling pitch contours (e.g. ‘high’ versus ‘low’ accents), but rather pitch *prominence*. In other words, we are only interested in predicting the presence or absence of a pitch accent on a given word. Also referred to as *pitch prominence detection*, we will hereafter use the terms ‘detection’ and ‘prediction’ interchangeably to discuss the automatic placement of pitch accent in spoken utterances. Likewise, we will use the terms ‘accent’ and ‘pitch accent’ interchangeably to refer to the prosodic prominence given to certain words in an utterance.

Finally, pitch accent in practice often manifests at the syllable level in a spoken utterance, and indeed some systems have investigated pitch accent prediction for syllable-based input (Gregory & Altun, 2004). For our purposes, however, we implement pitch accent prediction on word-based input in order to capture certain word-level accentuation phenomena as discussed in 2.1.

## 1.4 Related Work

This work follows Gregory & Altun (2004) and Levow (2008) in modeling pitch accent placement via discriminative sequential learning models using CRFs. In contrast to their work, which introduces syntactic, probabilistic, and acoustic variables as predictors for pitch accent, this thesis investigates potential variables for approximating semantic and discourse information in the prediction model.

Notwithstanding the abundant theoretical work on the importance of semantics and discourse in prosodic prominence, relatively little has been done to incorporate this information in online TTS-based prosodic systems. The biggest challenge lies in extracting discourse entities (as developed in various theoretical frameworks, roughly falling into the categories of *focus* and *givenness*) at the text analysis stage which can be passed into the prediction component.

One of the first studies in this direction was from Hirschberg (1993) and consisted of hand-crafted rules aimed at approximating the ATTENTIONAL, INTENTIONAL, and LINGUISTIC structure trifecta of Grosz & Sidner (1986). In her system, Hirschberg modeled the attentional structure as a stack of FOCUS SPACES<sup>1</sup>, in which objects, properties, and relations were not abstract concepts, but were represented simply by their lexical roots. In addition, she defined the LOCAL FOCUS as a collection of focus spaces on the current phrase, which is continually updated through push and pop of the stack, with orthographic and lexical cues being used to hierarchically relate focus spaces. GLOBAL FOCUS was concurrently defined as the concepts essential to the main purpose of the discourse, always accessible. Finally, contrastive stress in turn was modeled by tracking the presence of an item in both global and local focus, with the observation that:

“...items in global focus but not currently in local focus were frequently observed to be uttered with special emphasis, as if these items were being reintroduced into the discourse.”

This concrete implementation of abstract concepts of discourse was interesting, but ultimately did not result in substantial gains in the overall accuracy of her system. Hirschberg admitted, however, that some of the rules used to identify givenness often

---

<sup>1</sup>INTENTIONAL structure was not defined.

produced incorrect results<sup>1</sup>, thus leaving inconclusive the question of significance of discourse in determining pitch accent placement.

More recent work has focused on the use of discourse objects in statistical models of accent. The aims of this body of research have been two-fold: the first in finding ways of approximating discourse given lightweight syntactic information at the text analysis phase; the second in relating discourse information to other variables previously identified as relevant in the literature.

Regarding the former, Sridhar *et al.* (2008a) report on work in which the goal was to automatically detect *givenness* and separately, *focus*, for the eventual purpose of incorporating them as predictors of pitch accent. For the task of detecting givenness, a corpus was annotated using the hierarchy of Prince (1992)<sup>2</sup> in which first mentions are marked as *new*, subsequent mentions as *old*, and implicitly known or contextually evoked entities as *mediated*. Using a decision tree classifier based on a noun/pronoun distinction of items, their model achieved 88.29% accuracy on a binary classification of *new+med* and *old* items<sup>3</sup>.

In a separate experiment, they modeled contrastive elements by detecting various forms of focus. Focus annotations included *adverbial* for certain cue phrases like “only” or “just”; *contrastive* for comparison of two explicit lexical items; *subset* for hyponymy relations; *other* for all focused items not belonging to the above-mentioned categories; and *background* for all remaining entities. Analyses of the distribution of focus and accent status over three classes of words (nouns, adjectives, and function words) showed that while noun and adjectival background items are more likely to be accented in the absence of focus, the rate of accenting increases given focus of some kind. Classification results for the binary focus distinction given POS features, however, was much lower, at 72.95%.

At least two studies have investigated the use of information status and focus, along with several other semantic, topical, and discourse information. Nenkova *et al.* (2007) incorporated the same annotations for givenness and focus as described above, as well as additional indicators of *animacy* of the referent, and *dialog act* (by specifying the

---

<sup>1</sup>Contrastive elements were generally correctly identified and accented.

<sup>2</sup>The reader is referred to 2.2 for discussion.

<sup>3</sup>They likewise demonstrate an interesting correlation between acoustic parameters and givenness. While advantageous in speech understanding applications, we focus only on textually-extractable features in this work.

function of an utterance for declarations, or the type of question), and several other well-known probabilistic and positional features, including their own definition for *accent ratio*. Brenier *et al.* (2006) further added to this list variables for *lead word value*, which measured the ratio of how often a word occurs at the beginning of a discourse compared to elsewhere in the text (motivated by the empirical observation that the most important information, in news, occurs in the first paragraph), as well as a measure for verb specificity, indicative of the degree of relation between a verb and its subject (e.g. the relationship between “*arrest*” and “*police*” versus “*you*” and “*have*”).

In their results, the authors report relatively little gain in accuracy to be had from linguistic variables. In fact, Brenier *et al.* (2006) further claim, given the information gain for each feature, that discourse lends little to accent prediction, concluding that approximations of information status and contrast are “unlikely to be helpful in pitch accent prediction.”

It is important to point out, however, that their choice of model for accent prediction ignored important contextual behavior of accent, such as dropping accents to preserve rhythm. The choice of a decision-tree classifier, over other sequence-learning models, results in a measure of significance based on absolute counts of accented and unaccented elements in an utterance. These counts, however, will have effectively been skewed by productive deletion processes. While information status and focus might have in fact been useful in predicting accent for a given word, the significance of that decision was lost when accents are subsequently dropped in order to preserve prosodic rhythm. Indeed, Brenier *et al.* (2006) list incorrect accent in premodified nouns (i.e. “...when I was in HIGH school”<sup>1</sup>) as one of the three main sources of classification errors.

## 1.5 Overview of the Thesis

We begin with a discussion of some of the theoretical foundations for pitch accent placement, exploring some of the factors that can affect accent placement. We then introduce a model for discourse-driven pitch accent prediction, along with our experiments on data sets in German and English, followed by an evaluation of our results.

In Chapter 2, we discuss the prosodic aspects of pitch, including how it is produced and perceived, and motivate our treatment of pitch accent as a word-level process in

---

<sup>1</sup>Here, capitalization indicates true accent, while underline represents predicted accent.

spoken language. We also introduce some of the important theories regarding where and why accents occur in speech, in particular the elements of discourse and information structure that can affect accent placement. We finish with a discussion of state-of-the-art approaches to automatic detection of accent for use in speech technology. In Chapter 3, we propose a model for discourse-driven automatic pitch accent prediction using discourse structure and semantic space features. We continue in Chapter 4 with details of the data used to train our system, and end in Chapter 5 with an evaluation and analysis of our results. We conclude with a look at future directions for our work.



## Chapter 2

# Theoretical Foundations

Prominence plays a vital role in prosodic strategies within spoken language. Many factors may be involved in its contrivance, yet the wholesale perception of prominence figures highly in the facilitation of spoken communication.

In languages like German and English, prominence at the syllable level (in which a syllable may be perceived as somehow stronger than another) might, for example, distinguish lexical meaning, such as in *content* (noun) as opposed to *content* (verb). Prominence at the word level might distinguish different groups of words in a meaningful way, so as to disambiguate syntactic structures like ‘Gott vergibt Django nie’ (*God will never forgive Django*) from ‘Gott vergibt | Django nie’ (*God forgives, but Django never will*). At the utterance level, prominence can convey certain attitudes of the speaker, such as disbelief: ‘JOHN went to the game’ (*But John hates baseball!*); or highlight specific elements within a phrase such that additional information is imparted by this very emphasis, such as the answer to a question: ‘JOHN went to the game’ (*Who went to the game?*); or alternatively, signaling a contrast: ‘John went to the GAME’ (*Not the cinema*).

It is this utterance-level prominence of the latter example that is of primary interest, and in particular the highlighting capacity of emphasis, with implications reaching far beyond the scope of the purely linguistic meaning of an utterance. There are many facets to this particular manifestation of prominence. Globally, it is perceived in the domain of the tune – the overall intonational movement – of an utterance, and is generally produced with some motivating purpose.

These factors will be explored in the following chapter, beginning with a discussion of the perception and acoustic correlates of prominence, as well as phonological underpinnings in terms of well-known theoretical frameworks. From there, we will look at some of the motivations behind focal prominence in terms of information structure and discourse. Finally, we take a look at various attempts to computationally model phrase-level accent for applications in speech technology.

## 2.1 Accent and Prominence in Spoken Language

### The Production and Perception of Prominence

As previously mentioned, *tune* refers to the overall intonational movement of an utterance and is primarily driven by the perception of *pitch* – that phonetic impression of speech melody whose acoustic source lies in the fundamental frequency, or  $F_0$ . Physiologically, the fundamental frequency is a measure of the phonation process. More precisely, when laryngeal muscles contract and cause an adduction of the vocal cords, the approximated cords form a resistance to air expelled from the lungs. This in turn increases the sub-glottal pressure until the cords are forced apart, allowing increased air flow until such time as aerodynamic forces intervene to snap close the glottis ('t Hart *et al.*, 1990). The oscillation, or vibration resulting from an open-close cycle is measured in Hertz (Hz), where 1 Hz corresponds to one cycle per second. The range in the frequency of vibration is what we perceive as pitch – higher frequencies are perceived as higher pitch, and conversely, lower frequencies as lower pitch<sup>1</sup>.

Certain physiological factors, such as elasticity of the folds, and their length and mass, can contribute to the rate of vocal fold vibration. Whereas the normal speaking voice of men ranges between 80 and 200 Hz, the speaking voice of women, whose vocal folds are often shorter and lighter, allowing more frequent vibrations, fall in the ranges of 150 and 400 Hz. Other factors, however, serve to modulate  $F_0$ , such as muscle tension, which attenuates the vocal folds allowing higher  $F_0$  values, or intensified sub-glottal air pressure, which increases the amplitude of vocal fold vibrations (the *intensity* of which is perceived as *loudness*), and are largely under the speaker's control. More importantly, a speaker may choose to vary pitch levels over time.

---

<sup>1</sup>The term *pitch* is sometimes used to refer directly to  $F_0$ .

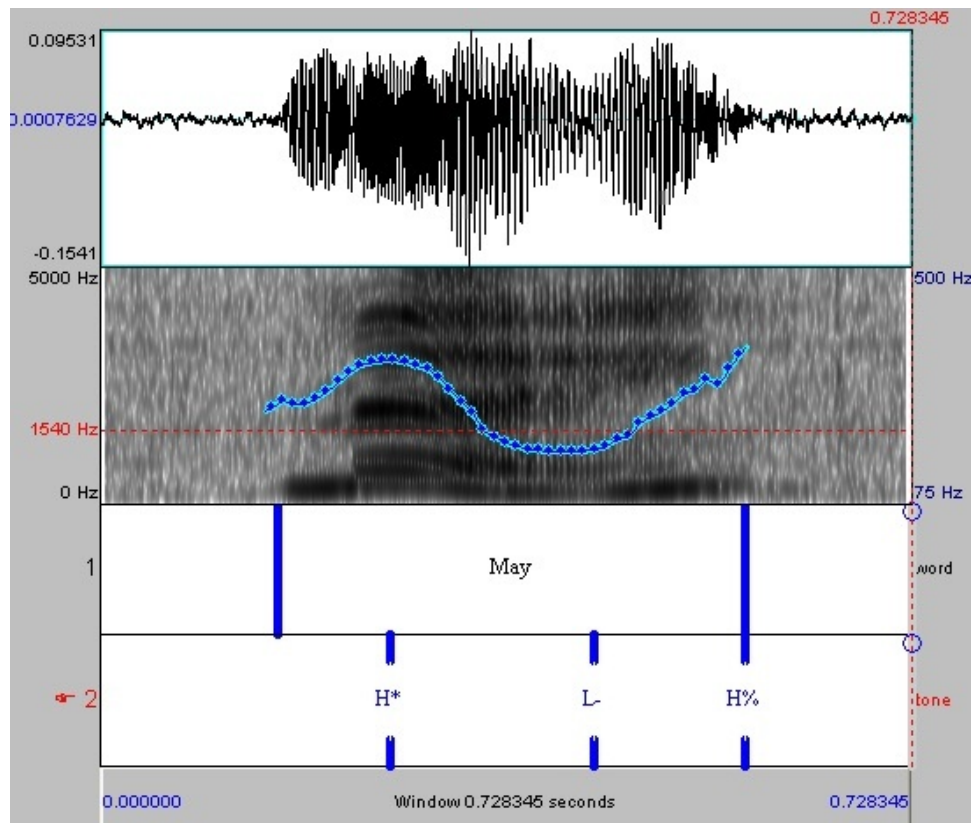


Figure 2.1: Spectral analysis and fundamental frequency of the word *May*

From figure 2.1<sup>1</sup>, it can be seen that as pitch is increased or decreased with time, distinct patterns of pitch rises and falls are created. Likewise, when pitch takes an abrupt and steep excursion relative to the rest of the  $F_0$  line, it thereby “accents” the given part of the speech signal<sup>2</sup>.

<sup>1</sup>Used with permission: [http://en.wikipedia.org/wiki/File:May\\_H\\_peg.jpg](http://en.wikipedia.org/wiki/File:May_H_peg.jpg).

<sup>2</sup>As a note, pitch is discontinuous as it is constantly peppered with voiceless consonants like /k/, /p/, /t/. Humans as listeners do not hear these interruptions, perceiving instead only the unbroken speech melody (for interruptions shorter than around 200ms). Moreover, changes in pitch after intervals of silence are simply perceived as continuous rises or falls, “as if human perception unconsciously bridges the silent gap by filling in the missing part of the pitch contour,” Nooteboom (1997).

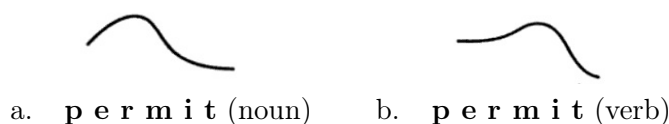
### Perceptual Components of Prominence

Bolinger (1958) called this sudden inflection in the  $F_0$  a *pitch accent*, a term which has been widely adopted since his time. Early investigations into the acoustic correlates of prominence identified pitch as the main contributing factor to lexical stress. Indeed, experiments by Fry (1958) showed that, by varying the parameters of ‘change in fundamental frequency’ and ‘change in intensity’, as well as duration in synthesized minimal pairs like *contract* (noun) and *contract* (verb), the perception of stress was largely due to changes in  $F_0$ . Isačenko & Schädlich (1966) found similar results for German after isolating the parameters of  $F_0$  and intensity in the minimal pair *übersetzen* (‘to ferry’) and *übersetzen* (‘to translate’). In spite of increased intensity in minor syllables, subjects overwhelmingly identified the syllables with increased  $F_0$  as accented.

Bolinger’s (1958) theory of pitch accent was based on the findings by Fry, and aimed to give an account of prominence solely based on pitch movements. According to him, accent constituted the (concrete) post-lexical realization of an (abstract) lexical property of stress and suggested that, if a syllable was lexically stressed, then that syllable would be post-lexically accented if that word was important enough to be accented (Bolinger, 1951).

Others, however, believed there might be more to the question of prominence than simply a peak in pitch. As (Ladd, 1996) points out, this view is challenged as soon as one explores prominence in more varied contexts. While pitch might be a significant factor for words spoken in citation form<sup>1</sup>:

(2.1)

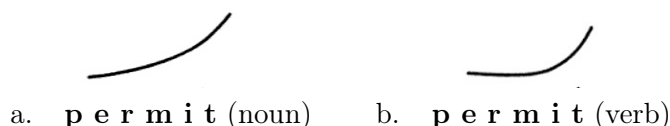


the story is quite different when the same words are given as questions:

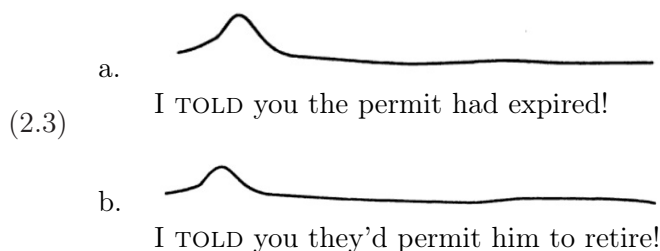
---

<sup>1</sup>The following examples are adapted from Ladd (1996).

(2.2)



In the examples of 2.2, not only do the contours of the question form differ wholly from their statement counterparts, but the pitch peak is no longer on the stressed syllable. Moreover, stress differences are plainly evident even without pitch variation, such as when the word occurs after an intonational peak, as in:



Subsequent studies investigated further acoustic parameters in hopes of explaining this contrast. Using automatic stress recognition on recordings of naturally spoken (American) English, Beckman (1986) found that in addition to fundamental frequency and intensity (each garnering around an 80% recognition rate), duration (whereby syllables are perceived as lengthened relative to their neighbors) was ranked just below (around 70%), and the parameter with the highest recognition rate (at 94%) was in fact ‘total amplitude’, a category combining duration and intensity. More recent work from Batliner *et al.* (2001) on (American) English and German, as well as from Kochanski *et al.* (2005) on (British and Irish) English found that not only was combined duration and intensity the most relevant prosodic feature of prominence and the most robust across the dialects/languages in question, but that fundamental frequency contributes very little to classification.

Other indicators that have been pointed out include the degree of articulation in vocalic quality, which is often directly observable through phonological processes of vowel reduction (Beckman, 1986) Beckman referred to English and German as ‘stress-accent’ languages in that they combine variations in pitch with factors of loudness,

length, and alternations of full and unreduced vowels to induce accentual prominence. These were contrasted with ‘pitch-accent’ languages like Japanese which make sole use of pitch for accentuation.

### Stress Shift

The story of stress-accent is somewhat compromised, however, in the so-called rhythmic clash contexts. In these cases, familiar stress patterns in words like ‘*MassaCHUsetts*’ suddenly exhibit a shift in accent when paired with certain other kinds of words, as in ‘*MASSachusetts MIRacle*’ (Hayes, 1984).

The same phenomenon can be seen in German, especially with regards to cardinal numbers, where the stress in the typical citation form of ‘*Vier-und-zWANzig*’ is shifted to an earlier syllable in ‘*VIER-und-zwanzig ROsen*’ (Wagner & Paulson, 2006). An acoustic study by Shattuck-Hufnagel *et al.* found that in the (English) cases, a shift in  $F_0$  was not accompanied by an additional increase in duration, observing that this offered “some evidence for rhythmic reorganization apart from but not necessarily independent of pitch accent placement” (1994). In other words, clash contexts induced only a shift in pitch, while other properties of the lexically-stressed syllable remained more or less the same<sup>1</sup>, suggesting a distinction between the two processes.

### Stress and Accent

It is clear that, contrary to Bolinger, many factors contribute to lexical prominence. Still, his claim might only have been limited in scope: if one factors out the notion that accent is the indicator of lexical stress, then pitch accents might be said to ‘fall’ on the stressed syllable of a word if that word happens to be important enough to bear accent. This effectively separates the process of accent from that of stress.

Such an account is more likely given that any word in an utterance can bear accent in the appropriate context. Consider the examples, in which the same utterance differs only by a single emphasis:

---

<sup>1</sup>Shattuck-Hufnagel *et al.* (1994) did report a decrease in duration on the stressed syllable in clash contexts, however this occurred regardless of a shift in pitch.

1. JOHN gave Mary the book.
2. John GAVE Mary the book.
3. John gave MARY the book.
4. John gave Mary THE book.
5. John gave Mary the BOOK.

In this way, *stress* can be distinguished as a purely lexical phenomenon, whereas *accent* occurs at the level of the utterance. We therefore adopt, for the remainder of this work, the definition of *lexical stress*<sup>1</sup> as the abstract potential for stress, with *post-lexical stress* denoting the concrete realization of stress. Additionally, the notion of *accent* (hereafter synonymous with *pitch accent*) will be used with the observation that, although accent and stress may co-occur, they are nevertheless separate events.

### The Phonology of Prominence

There have been many attempts to provide a mapping between the acoustic level of prominence (consisting of  $F_0$  and other properties) and the phonological level of categorical events (Taylor, 1992). While acoustic properties can be directly measured and extracted (and are, in a sense, concrete), phonology as an abstract layer can account for systematic variational phenomena. Intonational phonology, in particular, describes the fundamental components of intonation, such that phenomena like tonal variance (rise and fall patterns in  $F_0$ ), or prominence events within and across defined segments (lexical or compound stress) can be adequately explained.

One item which was subtly alluded to, but never fully articulated, is the notion of an ‘abstract’ lexical representation. This idea follows the tradition of Generative Grammar, the fundamental principles of which specifically posit an underlying (abstract) form of a linguistic item (such as a word or an entire utterance), which is subject to various parameters and rules resulting in its surface (observable) form (Chomsky, 1957; Chomsky & Halle, 1968). In spoken language, the atomic unit of this underlying form was considered to be the sound segment, and each segment bore a set of binary-valued features with the potential of being realized in its surface form. These features included

---


<sup>1</sup>Following Bolinger’s original distinction.

## 2.1 Accent and Prominence in Spoken Language

---

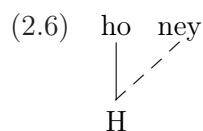
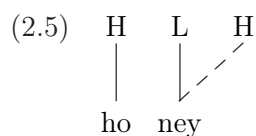
properties of the sound segment ([±voiced], [±palatal]), but also the so-called ‘suprasegmental’ features (properties spanning several sound segments) such as [±high(tone)], [±low(tone)] or [±stress].

This representation, however, failed to account for rich intonation patterns. Consider the complex intonation on the single syllable in 2.4 (adapted from Ladd (1996)):

- A. I hear Sue’s taking a course to become a driving instructor.  
(2.4)   
B. Sue?!

### Autosegmental Phonology

The desire to represent complex intonation led Goldsmith (1976) to decompose this consolidated linear representation of sound segments into several levels; in his model, each feature is represented on its own ‘tier’ as independent segments (or ‘autosegments’), with each tier related to the others via association lines. In this way, each linear string of underlying tone units (high H or low L) could be associated with an underlying string of tone-bearing units (vowels or syllables), as long as those association lines followed certain well-formedness conditions (for example, that associations must occur left-to-right, one-to-one maximally, and that association lines may not cross). Given these conditions, certain effects such as ‘tone-spreading’ and ‘tone-dumping’ appear to occur, resulting in either complex or drawn-out tones, as illustrated in:

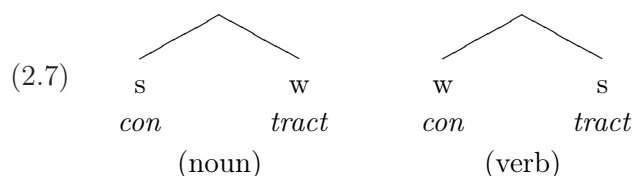


Of particular interest is the representation of complex intonation as a linear sequence of primitive tones, as illustrated in the tone-dumping of 2.5. Note that this representation makes no mention of the motivations behind the underlying form.

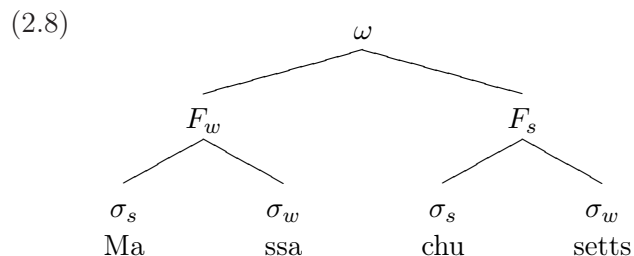


### Metrical Phonology

Apace with autosegmental phonology was another branch of phonology that moves away from strictly linear representations of sounds. Metrical phonology, first developed by (Lieberman, 1975; Liberman & Prince, 1977), holds that phonological units are hierarchically ordered into binary (sometimes n-ary) trees according to different domains. The success of this theory lies in its ability to account for prominence relations among several kinds of units. For example, tree nodes relating to syllables are either structurally strong (*s*) or weak (*w*) at each hierarchical level, as illustrated below:



The hierarchy of nodes is generally said to combine two syllables ( $\sigma$ ) into a metrical foot ( $F$ ), which themselves combine into a prosodic word ( $\omega$ ).



Prosodic words might then combine into higher units (such as phonological phrases, intermediate, or intonational phrases), but importantly, prominence is borne down through the tree in such a way that syllables carry different degrees of prominence. Syllables wholly dominated by *s*-nodes are considered to be the Designated Terminal Element (DTE), thereby uniquely accounting for primary, secondary and even lower degrees of stress.

Separately, the metrical tree can also be represented as a metrical grid which denotes strong nodes in terms of beats ( $\times$ ). In this view, each syllable gets a single beat, and may receive more beats given its relative strength in the tree. In particular, Selkirk (1984) proposed two set of rules governing the assignment of beats. Her ‘text-to-grid alignment’ rules add beats for heavy syllables (i.e. long vowels (CVV) or consonantal

## 2.1 Accent and Prominence in Spoken Language

---

onset+coda (CVC)) via the BASIC BEAT LEVEL and on a third pass, the last syllable with a second level beat gets another beat via the MAIN STRESS RULE (MSR), resulting in the following:

(2.9)

			×	
	×		×	
	×	×	×	×
	Ma	ssa	chu	setts

Note that while *-setts-* is a heavy syllable, it is considered ‘extrametrical’ and does not get a beat at the second level. In addition to these are high-level alignment rules, such as the NUCLEAR STRESS RULE (NSR), which assigns a beat to the rightmost lexically-stressed constituent in a phrase, and the COMPOUND STRESS RULE (CSR), which assigns a beat to the left-hand constituent of compound words.

In her second set of ‘grid-euphony’ rules, an attempt is made to produce the ‘ideal grid’. These rules are motivated by the *Principle of Rhythmic Alternation*, which holds that strong segments must be followed by weak segments, and that no more than two weak segments should occur at the same level. These rules serve to add, delete, or shift beats in the grid to preserve this alternation. Thus, when the compound ‘*Massachusetts miracle*’ is analyzed, the MSR assigns a beat to *-chu-* and *-mi-* in each isolated word, while the NSR additionally assigns a beat to ‘*miracle*’ given that it contains the last lexical stress in the phrase. The extrametrical *-setts-* cannot prevent the clash between *-chu-* and *-mi-*, therefore a movement rule shifts a beat leftward to the next stressed syllable<sup>1</sup>.

(2.10)

		×		×			×			×					
	×		×		×		×		×		×				
	×	×	×	×	×	×	×	×	×	×	×	×			
	Ma	ssa	chu	setts	Mi	ra	cle	⇒	Ma	ssa	chu	setts	Mi	ra	cle

One final rule in Selkirk’s model consists of the PITCH ACCENT PROMINENCE RULE (PAPR), which ensures that any syllable bearing a pitch accent is more prominent than any non-accent-bearing syllable. This rule reflects the dichotomy between pitch accent and rhythmic stress, and is most notable at the post-lexical level. Consider examples like 2.11 and 2.12 in which capitalization denotes the presence of a pitch accent<sup>2</sup>:

<sup>1</sup>Uhmann (1991) proposed similar rules for German.

<sup>2</sup>Examples adapted from Goldsmith (1996).

## 2.1 Accent and Prominence in Spoken Language

---

$\begin{array}{ccc} & \times & \\ & \times & \times \end{array}$

(2.11) FIREmen are available (NSR  $\rightarrow$  PAPR)

$\begin{array}{ccc} & & \times \\ \times & & \times \\ \times & \times & \times \end{array}$

(2.12) VOLunteer FIREmen are available (NSR  $\rightarrow$  PAPR  $\rightarrow$  NSR)

While the NSR calls for the prominence on the last lexical stress of the phrase *-vail-*, the PAPR takes precedence by assigning an additional beat on accented syllables *-fire-* and *-vol-*. In 2.12, we see the cyclic effect of these rules when the PAPR makes both *-vol-* and *-fire-* equally prominent, and NSR is again applied on the last most prominent syllable in the phrase.

From the discussion thus far, it might be said that stress and accent operate independently, yet influence each process to some extent. Although there are some notable exceptions<sup>1</sup>, pitch accent largely aligns with prominent syllables if a word is accented, and rhythmic prominence ensures euphony among stressed and accented syllables.

### Pierrehumbert Theory of Intonation

Developing the trend of generative phonology, and building on the theoretical foundations of Autosegmental-Metrical (AM) theory<sup>2</sup>, Pierrehumbert devised a system of intonation that provides something of a mapping between the  $F_0$  contour and categorical tonal events (Pierrehumbert, 1980). Her system proposed two primitives, the H and L tones, allowing tonal events to be categorized as sequences of tones. These tonal events can be thought of as pitch ‘targets’ and have the dual function of accentuation (in the form of pitch accents) and delimitation (edge tones).

The highest level domain of the intonational system is the intonational phrase (IP)<sup>3</sup>, with a secondary level called the intermediate phrase (ip) (Beckman & Pierrehumbert, 1986). Within these domains, pitch accents may associate with lexically-stressed syllables. These may be complex, in which case they are conjoined with a ‘+’, but the peak of the inflection is marked with a star ‘\*’, as in H\* or L+H\*<sup>4</sup>. Outside of these

<sup>1</sup>Consider *chiNESE*  $\rightarrow$  *a CHiNEse PERson* as opposed to *oBESE*  $\rightarrow$  *an oBESE PERson*.

<sup>2</sup>Term coined by Ladd (1996).

<sup>3</sup>As proposed by Selkirk (1984).

<sup>4</sup>In the most recent revision of Pierrehumbert’s system, complex tones consisted only of bitonal combinations of dissimilar tones (*i.e.* H\*+L, H+L\*, L\*+H, L+H\*), and were joined with the monotonal accents H\* and L\*.

## 2.1 Accent and Prominence in Spoken Language

---

domains, monotones associate with the edges of the phrase. Phrase tones or phrase accents, denoted with ‘-’, associate with the edge of intermediate phrases, and boundary tones, indicated by ‘%’, associate with the edge of intonational phrases.

Given the aforementioned inventory of tonal elements, a fully specified intonation target might look like:

$$(2.13) \quad \begin{array}{cccccccc} [ & ( & \text{John} & \text{gave} & \text{Mary} & ) & \text{ip} & ( & \text{the} & \text{book} & ) & \text{ip} & ] & \text{IP} \\ & & & & \downarrow & & \downarrow & & & & & \downarrow & & \downarrow \\ & & & & \text{H}^* & & \text{H-} & & & & & \text{L-} & & \text{L}\% \end{array}$$

Realization of the  $F_0$  contour based on these tonal targets is achieved via interpolation rules. For this reason, the actual contour may vary significantly between utterances given different sound segments and duration properties, yet still represent the same basic pattern.

$$(2.14) \quad \begin{array}{ccc} [ & \text{HE} & \text{lied} & ] \\ & \text{H}^* & & \text{L}\% \end{array}$$

$$(2.15) \quad \begin{array}{ccccccc} [ & \text{I} & \text{TOLD} & \text{her} & \text{to} & \text{leave} & ] \\ & & \text{H}^* & & & & \text{L}\% \end{array}$$

In 2.14 and 2.15, the target specification is the same, but factors of duration, down-drift, down-step, and pitch range, all influence the final  $F_0$  contour.

### Prosodic Annotation: ToBI

In a revised version of the Pierrehumbert system, an attempt was made to create a standard for labeling prosodic features of digitized speech through the Tones and Break Indices (ToBI) framework (Silverman *et al.*, 1992; Beckman & Hirschberg, 1994; Beckman & Ayers, 1994). The ToBI transcription system follows in the AM tradition by having several related tiers of transcription. These tiers may include phonetic and syllabic transcription, disfluencies, and more. The two most important tiers are factors of its name: these are the *tone* tier and the *break indices* tier. The tone tier specifies tone targets for pitch accent, phrase tones, and boundary tones as previously described. The break indices tier specifies additional information about the strength or type of boundary: 0 - for clitic boundaries (*e.g. who's*); 1 - for normal word boundaries; 2 - for boundaries with no apparent intonational movement; 3 - for an intermediate phrase;

## 2.1 Accent and Prominence in Spoken Language

---

and 4 - for full intonational phrases. This multi-tiered transcription system thus provides machine-readable prosodic symbols which can be used in speech technologies like Automatic Speech Recognition (ASR) and TTS.

ToBI was specifically developed for English given the extensive theoretical background of intonation available for the language. Other work has been carried out to adapt ToBI for the analysis of variants of British English with IVIE (Grabe *et al.*, 2001a), and German with GToBI (Baumann *et al.*, 2001).

Although these systems have become the de-facto standard for prosodic annotation, there has nevertheless been some criticism of the model, especially with regards to inter-transcriber reliability (Pitrelli *et al.*, 1994; Grice *et al.*, 1996). As a theory-driven framework, there is no inverse mapping from the acoustic speech signal to the abstract prosodic symbols. Notwithstanding the inadequacies of a categorical framework for describing the characteristics of a continuous acoustic signal, it remains a useful tool for learning where and which kinds of accents occur in speech.

### Patterns of Intonation: Accent Deletion

One advantage of a distinct theory of intonation is the emergence of observable phonological patterns of accent. In one particular case, the principle of rhythmic alternation seems to reappear at the word level, inducing an effect of ‘dropped’ accents in sequences of pitch-accented words within an intonational phrase. This happens frequently in compound nouns in English, but the position of the dropped accent is often idiosyncratic to the compound<sup>1</sup>:

	apple PIE	APPLE cake
	sunday NIGHT	GARDEN hose
(2.16)	Madison AVENUE	MADISON street
	student UNION	TRADE union
	city HALL	TAX office

Interestingly, when such compounds are joined into larger compounds, these familiar accent patterns change, presumably to preserve rhythmic alternation:

	SUNDAY night FOOTBALL
(2.17)	CITY hall TAX office
	FAIR trade STUDENT union

---

<sup>1</sup>The majority of English compounds are left-headed (Lieberman & Sproat, 1992).

The same holds to a certain extent in German, for example in ‘*montag MORGEN*’ versus ‘ANFANG *Januar*’. It is more easily observed in names and quantities, however. Consider:

- |        |                     |                 |
|--------|---------------------|-----------------|
|        | cap GEMINI          | SEMA group      |
| (2.18) | deutsche POST       | DEUTSCHE Bank   |
|        | Verbundnetz GAS     | ROTE Armee      |
|        | tausender HAITIANER | TAUSENDE Häuser |

While typical German compounds differ somewhat from their English counterparts<sup>1</sup>, strings of nouns nevertheless exhibit something of the same behavior, as in:

- |        |                                    |
|--------|------------------------------------|
| (2.19) | PLO-CHEF Yassir ARAFAT             |
|        | MARBURGER SPARKASSEN-CHEF Udo GÜDE |

The examples above have sometimes been referred to as instances of ‘Compound Stress’ (cf. Ladd (1984)). However, as these can be regarded as primarily pitch-driven events, we will hereto call this a case of ‘Compound Accent’ in accordance with the definition previously laid out in 2.1. Moreover, we argue that this can be regarded as an instance of accent deletion. The reason for this is two-fold: first, the deaccented words in all of these examples can be contrasted with productive instances of fully accented phrases<sup>2</sup>; secondly, in the general case of focal highlighting<sup>3</sup>, the focus can be said to span the entire entity referenced by the full name (e.g. ‘*Yassir Arafat*’ the individual), suggesting that the accent was dropped where one might expect an accent to be. For these reasons, we will therefore refer to this as an instance of *compound deletion* for the purposes of preserving prosodic rhythm.

## 2.2 Information and Accent

### The Purpose of Accent

Intuitively, the motivation for accent often centers around the notion of *information*; as Prince (1981:223) observes: “It is a truism that, when people use language naturally,

<sup>1</sup>German compounds are generally realized as a single word (e.g. ‘Apfelkuchen’ versus ‘apple pie’). Consequently, lexical stress rules apply and accent is generally realized on the first syllable of the word.

<sup>2</sup>German deaccented names like those in 2.19 are contrasted in our corpus with fully accented names (cf. section 4.1.1 for an introduction to the MULI corpus). Most often, this occurs between commas, such as ‘*Bundesminister, Günter Rexrodt, ...*’, or in cases where the name refers to a well-known individual (e.g. ‘*James Bond*’ was always fully accented).

<sup>3</sup>As opposed to contrastive emphasis as found in: ‘*Yassir ARAFAT, not Yassir ARUBAT*’, wherein focus lies solely on the last name.

they are usually attempting to convey information.” Information, however, is not simply the strictly semantic meaning that arises from a particular linguistic expression. Rather, it is a function of a fluid system of communication. Lambrecht (1994), for example, describes the informational value of an utterance in terms of its capacity to change the ‘mental representation’ of the world for the hearer. Information, according to him<sup>1</sup>, is inherently propositional (e.g. one can inform someone of the cost of a book, but not of a book, etc.) and it is the sum of all propositions known or believed to be valid between the interlocutors that constitute this representation, also referred to as the *common ground* (Stalnaker, 1978). Thus, to inform someone of something is to “induce a change in that person’s knowledge state by adding one or more propositions” (Lambrecht, 1994:44).

This additive, forward progression of mental states necessarily introduces the dichotomy of the established from the novel, as Dahl (1976:38) notes:

[T]he speaker assumes that the addressee has a certain picture – or model – of the world and he wants to change this model in some way. We might then identify THE OLD or THE GIVEN with the model that is taken as a point of departure for the speech act and THE NEW with the change or addition that is made in this model. [...] We can say that the addressee receives “new information” in the sense that he comes to know or believe more about the world than he did before.

Lambrecht (1994:47) argues, moreover, that information cannot be said to be syntagmatic in nature. In other words, old and new information is not necessarily equivalent to old and new referents, as illustrated in the following exchanges:

- (2.20) A. Where did you go last night?  
B. I went to the movies.

- (2.21) A. When did you move to Switzerland?  
B. When I was seventeen.

While one could argue that the syntactic constituent ‘*the movies*’ in 2.20 might constitute new information in the mental representation of the hearer, the same could not

<sup>1</sup>Lambrecht departs from Dahl’s (1976:38) characterization of propositional information in that propositions as mental representations merely exist or do not exist, but have no logical value of truth.

be said of the linguistic elements in 2.21. The information conveyed in 2.21 is more accurately characterized as the relation between an act, an individual involved in the act, and the timing of the event.

More precisely, information is primarily pragmatic; utterances are never made in a contextual void, but are always situated. Wennerstrom (2001) illustrates this idea through the following example:

(2.22) Watch out! That CHÍMNEY's falling down.

In describing the situational context, Wennerstrom makes the following observations, “From the warning *Watch out!* it is evident that the speaker sees a danger that he supposes the hearer does not yet see. After this phrase, the hearer is on the alert, looking for something amiss. The word that conveys the deictic structure of the situation: a particular chimney must be visible, at some distance from the hearer and speaker, but near enough to pose a danger, hence the warning” (2001:5).

The example in 2.22 illustrates several important issues regarding how information is conveyed, and how it is construed. First, intonational markings play a role in conveying the message. The accent on ‘*chimney*’, for example, ensures that only one chimney is of interest and that its current state is of primary interest – an accent on ‘*that*’ would have suggested the presence of more than one (e.g. ‘*THAT chimney*’ as opposed to the another one). The lack of accent on other elements suggests that the speaker regards certain aspects of the situation as uncontroversial. An accent on the auxiliary, for example in ‘*that chimney IS falling down*’, would signal prior doubt among the interlocutors, while ‘*that chimney's falling DOWN*’ would eliminate the possibility of the chimney falling in any other direction.

Halliday (1967) called this the result of *information structure*, a term which has since come to broadly incorporate the linguistic organization in an utterance for the purpose of successfully conveying information. A related notion is *information packaging* (Chafe, 1976), which is based on the active role played by the speaker in gauging the hearer’s state of mind, described as “the tailoring of an utterance by a sender to meet the particular assumed needs of the intended receiver. That is, information packaging in natural language reflects the sender’s hypotheses about the receiver’s assumptions and beliefs and strategies” (Prince, 1981:224).



### Information Structure

Although we adopt the terminology of (Halliday, 1967), the notion of information structure has in fact been described through a bewildering number of concepts, many of which are inter-related and none of which is generally accepted. Most theories, however, are based on the notion that each utterance is driven by some form of *new* information, which can be contrasted with *old* or *given* information, available to the interlocutors of the discourse.

Halliday called *new* information the *information focus*, or the “main burden of the message” (Halliday, 1967:204) which is intonationally marked with a tonic component (or, as previously defined, pitch accent). According to him, accent falls on the focused item of an utterance, as illustrated in the examples below:

(2.23) JOHN painted the shed yesterday.

(2.24) John PAINTED the shed yesterday.

(2.25) John painted the shed YESTERDAY.

(2.26) John painted the SHED yesterday.

The examples in 2.23–2.25 reveal the information focus which presupposes a certain context, for example as the answer to questions like ‘*Who painted the shed yesterday?*’, ‘*What did John do to the shed yesterday?*’ and ‘*When did John paint the shed?*’. The example in 2.26, on the contrary, is ambiguous: it might presuppose the question<sup>1</sup> ‘*What did John paint yesterday?*’, but it could as easily be an answer to the question ‘*What happened?*’.

Ladd (1980, 1996) distinguishes the focus in 2.26 in terms of *narrow*, that is, restricted to a single word, versus *broad*, spanning an entire informational unit, as illustrated by the focus items below<sup>2</sup>:

(2.27) John painted [the SHED]<sub>F</sub> yesterday.

(2.28) [John painted the SHED yesterday.]<sub>F</sub>

<sup>1</sup>Setting aside, for the moment, the case of contrastive stress, in which this could be an answer to the question ‘*Did John paint the barn yesterday?*’.

<sup>2</sup>Following the generative tradition of focus semantics, in which focus is typically marked as a binary feature.

This analysis of focus, however, precludes the direct mapping of focus and accent. Importantly, one cannot say that all but the item ‘*shed*’ are somehow given in the answer to the question ‘*What happened?*’. Selkirk (1984, 1995) nevertheless accounts for the accent in broad focus through the syntactic rule of *focus projection*<sup>1</sup> in which a focus item can license the f-marking of higher constituents in the syntactic tree, as shown in 2.28.

### Focus Accent

Gussenhoven (1983) called this the *Focus-to-Accent* (FTA) approach, and much debate has centered on the placement of accent in a focus context. Halliday (1967) follows the traditional constituent-based approach, claiming only one main accent per utterance falling on the last lexical item in the phrase (according to Chomsky & Halle’s (1968) *Nuclear Stress Rule* (NSR)).

Others have argued for similar structure-based approaches, for example on the basis of predicate-argument structure. Schmerling (1976) suggests that an argument-stress rule can account for differences in German and English, wherein verb-final German phrases often violate NSR, while English sentences typically have argument-final constructions<sup>2</sup>. In a similar vein, Gussenhoven’s *Sentence Accent Assignment Rule* (SAAR) (1983; 1992) attempts to construct accent domains applicable for English and Dutch, but also German by extension, in which predicate-argument structures make up a single focus domain with adjuncts forming a separate domain.

Of note in these constituent-based theories is the idea of the singular notion of focus: each phrase contains only a single item that is highlighted in any given utterance. In accordance with his *one-new-idea hypothesis*, Chafe (1994) supports the idea that intonational units express one new concept at a time for purposes of cognitive efficiency given the temporal constraint on processing information in the mind. This might also be a sufficient explanation for why not all elements in a broad focus receive accent, which one would expect if all focused items must be accented.

In practice, however, several items may be presented as new (accented) within an utterance. Cruttenden (1997) in particular reported that listeners were divided as to which word was most prominent in examples like 2.29 and 2.30.

---

<sup>1</sup>Focus projection has been independently shown for German by Uhmman (1991).

<sup>2</sup>Consider ‘*Ede drove to [Frankfurt]<sub>NP</sub>*’ and ‘*Ede ist nach Frankfurt [gefahren]<sub>VP</sub>*.’

(2.29) It's NOT quite the RIGHT shade of BLUE.

(2.30) Her FACE USED to be much FATTER.

Some listeners regarded the ‘*not*’ in 2.29 as most prominent, even as most preferred ‘*blue*’. In 2.30, however, the majority found ‘*face*’ to be the most pronounced, evidence he readily admits undermines his position on a phrase-final nucleus, and indeed, calls into question the basis for a nuclear accent.

Still, Büring (2006), another proponent of structure-based focus domains, contends that the distribution of accents does not necessarily coincide with focus structure.

(2.31) The LAWYER sent the REQUEST to the OFFICE.

Büring maintains that the intonation pattern in 2.31 would satisfy not only the general-purpose question ‘*What happened?*’, but equally ‘*Who did the lawyer send the request to?*’ or ‘*What did the lawyer do?*’. He further points out that 2.31 would likewise be an answer to the question ‘*Where did the lawyer send the request?*’, in which case the focus in 2.31 would be on the prepositional phrase, as illustrated below:

(2.32) The LAWYER sent the REQUEST [to the OFFICE]<sub>PP</sub>.

In 2.32, the accent on ‘*lawyer*’ and ‘*request*’ obtain in spite of the fact that these words constitute backgrounded items. Büring, however, claims that these accents are merely “ornamental” and optional, only assigned to maintain a “pattern of relative prominence” across a constituent, but importantly, are only ever pre-nuclear and pre-focal (2006).

### Deaccentuation

Another interesting case of focal accent involves the apparent deaccenting of items within a focus domain. One example which has often been discussed in the literature is the following (in Büring, 2006):

- I know that John drove Mary’s red convertible. But what did Bill drive?
- (2.33) A. Bill drove [her BLUE convertible.]<sub>F</sub> (Büring, 1996)  
 B. Bill drove [her MOTORCYCLE.]<sub>F</sub>  
 C. \*Bill drove [her blue CONVERTIBLE.]<sub>F</sub>

Even as Büring argues for a rule of unrestricted vertical focus projection, he suggests that the deaccentuation on ‘*convertible*’ is due to Schwarzschild’s (1999) definition of ‘given’ – namely, that the antecedent of the item is salient.

Ladd (1996) also argues that if the lexical item is somehow semantically available from the local discourse (i.e. *given* or *backgrounded*), then it will not be accented. Rather, it appears to shift onto its neighbor, as in the following examples:

(2.34) A. Why didn’t you read the article I gave you?  
B. I can’t read GERMAN.

(2.35) A. The only article on this is in German.  
B. I can’t READ German.

(2.36) A. Where did you go just now?  
B. I took the GARBAGE out.

(2.37) A. What happened to all the garbage?  
B. I took the garbage OUT.

In the above examples, 2.34 and 2.36 show the default pattern of accent in the absence of givenness, while 2.35 and 2.37 display a left-ward and right-ward shift in stress due to the repeated mention of ‘*German*’ and ‘*garbage*’, respectively. Ladd (1980, 1996) accounts for this syntagmatically via a theory of sentential metrical stress, but admits some form of “relative prominence within a metrical structure” (1996:231), suggesting for instance that heavier constituents tend to form intermediate phonological phrases. In other words, accent is also affected by considerations of rhythm and melody across an utterance.

### **Theme-Rheme (Topic-Comment)**

The deaccentuation of ‘given’ items runs rampant in both English and German, however, and the process is never really straightforward. In the first place, several dimensions of givenness can be observed. It has been claimed that sentence structure reflects given and new propositions, respectively. Halliday described this in terms of a functional split in the clause, in which the initial element is the *theme* (or, “what is being talked about, the point of departure for the clause as a message” (1967:212), and its complement, the *rheme*. Prevost (1995) supports this division and further contends that a correlation exists between thematic/rhematic propositions and accent type, as illustrated below:

- Q. What kind of music does your older brother prefer?
- (2.38) A. [ My OLDER brother prefers ]<sub>theme</sub> [ BAROQUE music. ]<sub>rheme</sub>  
 L+H\* L(H%) H\* LL\$

Related to givenness in information structure is that of *topic-comment* relations (Gundel, 1985; Sgall *et al.*, 1973). The topic is rather informally defined as what the sentence is about, while the comment is what is actually said in the utterance. The topic is established through the discourse and can be thought of as that which answers the question ‘*What about X?*’. Jackendoff (1972:261) also relates the prosodic tune to the topic-comment structure via the (background) B pitch accent, and the corresponding (answer) A accent, as illustrated below:

- Q. Well, what about FRED? What did HE eat?
- (2.39) A. [ FRED ]<sub>T</sub> ate the [ BEANS. ]<sub>F</sub>  
 B A

- Q. Well, what about the BEANS? Who ate THEM?
- (2.40) A. [ FRED ]<sub>F</sub> ate the [ BEANS. ]<sub>T</sub>  
 A B

### Givenness of Discourse Referents

Givenness can also be said to operate on the representation of referents in the mind of the interlocutors. Ladd (1980:52) gives the following example in which the deaccented ‘books’ represents an accessible referent:

- (2.41) A. Has John read Slaughterhouse-Five?  
 B. No. John doesn’t READ books.

In this case, ‘books’ is a referent to ‘*Slaughterhouse-Five*’, not through explicit mention, but through a hypernym relation, which determines its information status as Given.

In practice, there are many degrees of givenness for referents. Prince (1981) perhaps gives the most thorough account of the degrees of givenness in a taxonomy of given/new information. She first acknowledges three basic levels of givenness:

**Predictability/Recoverability** The speaker assumes the hearer CAN PREDICT OR COULD HAVE PREDICTED that a PARTICULAR LINGUISTIC ITEM will or would occur in a particular position WITHIN A SENTENCE.

The speaker makes the assumption that the semantic meaning of a lexical item is predictable based on the preceding situation or context. The situation or context may be felt syntactically, or pragmatically.

(2.42) John punched Bill and then he insulted him.

(2.43) John punched Bill and then HE insulted HIM.

The lack of accent on ‘*he*’ in 2.42 is due to the predictability of ‘*John*’ as the antecedent, given that he was the subject of the immediately preceding clause. An accent on ‘*he*’ and ‘*him*’ signals an unpredictable (and thus unrecoverable) change in the meaning. Similarly, we do not want to say that because the focus is on these respective pronouns that ‘*insulted*’ is therefore given (or constitutes the background), and neither would we want to say that ‘*punch*’ presupposes ‘*insult*’. Rather, it is the pragmatic context that makes this idea recoverable or predictable for the hearer.

**Saliency** The speaker assumes that the hearer has or could appropriately have some particular thing/entity/... in their consciousness at the time of hearing the utterance.

Again, an entity may be made salient through explicit or pragmatic mention. Consider the following examples:

(2.44) We got some beer out of the trunk. **The beer** was warm.

(2.45) [*Ai* to *Bj* as *Ck* passes by, in view and out of earshot]  
How old do you think **he<sub>k</sub>** is?

In each of these cases, the bold-font items are *given* due to having been explicitly referenced, as in 2.44, or due to the *saliency* of the referent given the immediate pragmatic context, as in 2.45.

**Shared Knowledge** The speaker assumes that the hearer “knows,” assumes, or can infer a particular thing (but is not necessarily thinking about it).

In the following examples, the bold-font referents are assumed to be known between both speaker and hearer:

(2.46) Where were **your grandparents** born?

(2.47) A. Hi, **I**’m home.  
B. Where’s **Daddy**?

In these cases, one might say that they are perpetually given in the sense that they belong to some “permanent registry” between the interlocutors<sup>1</sup> (Prince, 1981).

These levels of givenness notwithstanding, Prince argues that the terminology is misleading in the sense that all three levels are essentially founded on the speaker’s assumptions. She therefore proposes the following taxonomy (represented in illustrated in figure 2.2) of given/new information based on the principle of “assumed familiarity” of entities (1981:237).

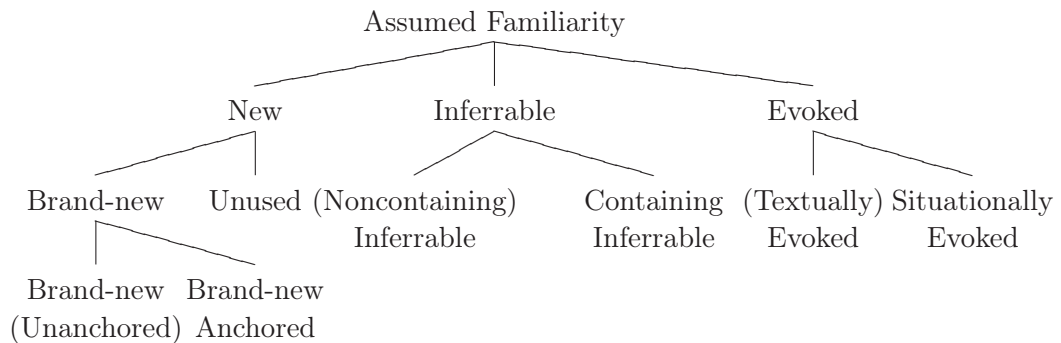


Figure 2.2: Prince’s taxonomy of given/new information

Empirical investigations into the accentibility of given/new entities has led to varying results. Using classifications derived from Prince’s taxonomy in a simplified dialog setting, Brown (1983) found that speakers accented 87% of “brand new” entities (i.e. assumed unknown to the hearer) and 77% of “new inferrable” entities (i.e. assumed inferrable from previously evoked entities), while 96-100% of “evoked” entities were deaccented, suggesting that the simple given/new distinction is adequate in accounting for accent. Terken & Hirschberg (1994), in a similar experiment, found just the opposite; according to them, simple givenness, defined as mere mention in a context, is not sufficient for deaccent, but that grammatical function and surface position play an important role. They suggest that these might be additional constraints on deaccentuation.

In German, Becker *et al.* (2006) reported that while new entities were consistently accented (“brand new” 91%, “unused” 93%), deaccentuation could not be confirmed

<sup>1</sup>From Kuno’s distinction of anaphoric-nonanaphoric, in which he states that an NP is anaphoric if “[its] referent . . . has been mentioned in the previous discourse” or is “in the permanent registry” (Kuno, 1972:270).

since accenting held for “inferrable” (89%) and “evoked” entities (91%) alike. Baumann (2006), on the other hand, investigating the same German corpus, found that while deaccentuation occurred to some extent (especially in non-contrastive cases), given or accessible information is more often represented by the H+L\* accent rather than H\*<sup>1</sup>.

Accents thus can occur both in and outside of a focus domain, but not all accents in a focus domain are accented (e.g. in the case of broad focus, as well as in deaccentuation). Furthermore, while new entities are often accented in a discourse, the extent to which given information is accented/deaccented is in question. Aspects of rhythm and relative prominence likewise seem to play a role in the overall accenting strategy, calling into question the adequacy of a purely syntactic, rule-based approach to accent assignment.

A major opponent to rule-based accent assignment was Bolinger (1972), whose famous paper entitled “*Accent is Predictable (If You’re a Mind-Reader)*” sums up many of the difficulties in predicting the placement of accent. In his view, accent is a product of the speaker’s intentions (emotional highlighting) within a discourse; more precisely, syntax cannot account for what is clearly a semantic issue. What counts in accenting strategies, if anything, is *relative* semantic weight; in phrases like ‘*bóoks to write*’, ‘*wórk to do*’, ‘*clóthes to wear*’, the verb is highly predictable and thus is often deaccented, whereas the semantically richer lexical items in phrases like ‘*pròblems to cómputerize*’ (compare with ‘*próblems to solve*’) and ‘*pòint to émphasize*’ (as opposed to ‘*póint to make*’) are typically accented. “The point is,” continues Bolinger, “that the speaker adjusts the accents to suit his meaning. *Weed and fertilize* can be deaccented [from *I still have most of the gárden to wèed and fértilize*]; *clean and oil* can be accented [from *I have a clóck to clean and oil*]. It is in the nature of the case that our examples can show probabilities, rarely certainties” (1972:635). While semantic likelihood cannot account for all instances of accent, it too factors in accenting strategies.

## Discourse Structure

Many aspects of information structure, such as lexical likelihood, focus items, and givenness, clearly result from some discourse context, but are often examined in settings where such context is assumed rather than made explicit. Citation-form examples and query pairs are useful for emphasizing important characteristics within an utterance,

---

<sup>1</sup>It should be noted that the corpus in question represents read news articles (cf. MULI corpus, section 4.1.1). The lack of deaccentuation in these cases may be genre-specific.



but may ignore important trends across an entire coherent passage. The example in 2.28, reproduced in 2.48, illustrates how accent clearly distinguishes the focus item from several possible entities:

(2.48) [John]<sub>e</sub> painted [the SHED]<sub>e</sub> yesterday.

In practice, focus may be expressed in a number of ways, such as through the choice of referring expression, the spread of focus items over several utterances, and even the ordering of utterances across the passage.

- (2.49) a. John woke up early yesterday.  
b. He wanted to paint the shed.  
c. By the end of the day, it was a horrid lime green.

The passage in 2.49 illustrates a shift in focus from ‘*John*’ to the ‘*shed*’, primarily achieved through the pronominalization of ‘*John*’. Towards the end of the passage, the discourse continues to center around the shed, but as this has already been mentioned, it is likewise pronominalized, thereby emphasizing new information given by ‘*horrid lime green*’.

Grosz & Sidner’s (1986) theory of discourse structure has perhaps been the most comprehensive in nature for processing the utterances in a discourse structure. In particular, they proposed three separate but related components of discourse structure: the INTENTION, ATTENTION, and LINGUISTIC structure.

In their model, a discourse is viewed as an aggregation of *discourse segments* – typically consisting of one or more linguistic utterances – each segment is said to be associated with a discourse segment purpose (DSP). It can also be viewed as *global discourse* factored by smaller segments of *local discourse*. The INTENTIONAL structure describes the purposes and the relations among them that ultimately satisfy the purpose of the entire discourse. The LINGUISTIC structure, on the other hand, describes the linguistic devices, such as cue phrases and the choice of referring expression, that are used to signal the shift from one discourse segment to the next. Importantly, the linguistic structure is constrained by the discourse segment; the use of pronouns or reduced definite noun phrases are limited to the objects, properties, or entities within a single segment. Finally, the ATTENTIONAL state captures the dynamic transition between focus objects as the discourse unfolds. This is achieved by various sets of *focus spaces*, each of which is associated with a discourse segment. These focus spaces contain

not only those entities which are salient at any given moment in the discourse segment, but the purpose of the discourse segment as well.

Entities are considered salient by explicit mention or because they have somehow been evoked in the process of producing or comprehending the current discourse. Moreover, the structure for focus spaces is defined as a stack, wherein entities of the top-most focus spaces are considered the most salient. In this way, the depth of the focus space reflects relative salience. Dominance relations also hold among DSPs, reflecting a partial ordering of intentions, and determine the application of push and pop to the stack. Processing the attentional state depends on the INTENTIONAL structure, just as the LINGUISTIC structure is restricted by the salient items of the ATTENTIONAL state.

A related notion is *centering*, a theory for modeling coherence, focus, and linguistic expressions at the discourse segment level of ATTENTIONAL structure (Grosz *et al.*, 1995). Centering, in particular, imposes constraints on each of these elements to ensure that entities are smoothly linked across all utterances of a discourse segment. Centers are defined as semantic objects rather than words or syntactic phrases, yet they constrain the realization of linguistic form to preserve coherence. For example, once a center has been realized as a pronoun, it must continue as a pronoun across the discourse segment. Also, coherence of a discourse segment is relative to the transitions between centers: continuing the link of the same center over a sequence of utterances is preferred over shifting to a new center.

We have so far introduced commonly observable patterns of accent, including accent deletion, accent shift, and deaccent, and discussed several theoretical accounts for the placement of accent, including focus, givenness, and discourse structure. In the next section, we will explore several attempts to model accent assignment for use in speech technology, and investigate the extent to which constraints of focus, givenness, structure, and rhythm affect prosodic accent in real-world models.

## 2.3 Computational Models of Accent Variation

### Empirical Aspects of Pitch Accent

Empirical investigations of pitch accent variation offer interesting insights into the nature of accent processes, as well as potentially significant predictors of accent. Part-of-speech (POS), for example, has long been considered the single most useful predictor

## 2.3 Computational Models of Accent Variation

---

for pitch accent. This is because most analyses of corpora show that certain classes of words, such as nouns and adjectives, are much more likely to be accented, while function words, such as prepositions and articles, are largely unaccented. These findings are in line with theoretical assumptions claiming that accent highlights important information in an utterance.

Accordingly, later work focused on capturing *information* in the prediction model. Pan & Mckeown (1999) looked into the relevance of Information Content (IC) as a statistical predictor of accent status. In particular, they used a standard IC measure (defined as the log probability of a lexical item), as well as TF\*IDF (a popular score in information retrieval applications) to test their predictive power over POS tags in pitch accent prediction. Their results showed that while IC was roughly equivalent to POS in their models<sup>1</sup>, the performance of the combined POS+IC model was above and beyond that of individual predictors alone. Other probabilistic measures were likewise explored by (Gregory & Altun, 2004), including log-scaled unigram, bigram, and joint probability distributions. The addition of these variables produced a slight improvement in accuracy: 73.94% as compared to their equivalent POS+IC score of 72.56%<sup>2</sup>.

Syntactic category and information content is not enough, however, to account for the deaccentuation patterns in compounds. Hirschberg (1993) tried to model this phenomenon explicitly using a complex nominal analyzer which, combining semantic rules with table look-up, predicts citation form stress patterns. Later work by Pan & Hirschberg (2000) introduced several collocation measures into a rule-based system, including bigram predictability, mutual information, and the Dice coefficient. While all three measures were found to be significantly correlated, bigram predictability alone accounted for most of the variation across both read and conversational speech domains. N-gram statistics, while useful for idiosyncratic cases of compound accent, are nevertheless extremely sensitive to corpus size. The authors noted, for example, that while ‘*street*’ is commonly deaccented in compounds like ‘FIFTH *street*’, this compound never occurred in their particular corpus. Consequently, the bigram probability was extremely low.

---

<sup>1</sup>IC proved more effective than TF\*IDF, and thus figures solely in comparisons with POS.

<sup>2</sup>This score is further increased when adjustments to the particular model are made. For example, increasing the window size to 5 in their sequence learning model boosted the performance to 74.51%.

Another class of accent predictors is acoustic-based, and has been explored in conjunction with pitch accent detection within speech. Duration, speech rate, number of syllables and phones, and pausing information (Gregory & Altun, 2004) have all been incorporated for marginal improvement. Spectral characteristics, and in particular spectral balance, indicative of the vocal effort put into syllable articulation, was also explored (Ren *et al.*, 2004). Levow (2008) further investigated co-articulation effects that have demonstrable impact on tone and pitch accent recognition. These include differences in pitch ranges of neighboring syllables, which may influence whether an accent is perceived as high or low, and syllable production, which affects the point at which maximum pitch height will be reached. In her system, contextual articulatory features included mean pitch across the syllable, maximum and mean pitch as well as intensity from neighboring syllables, and changes between these values.

Until very recently, very little work has been done to approximate information status in TTS applications. Focus, the given/new distinction, and contrastive stress can be difficult to obtain in the absence of deep linguistic processing of the text. Rather, several light-weight syntactic elements (such as POS) are combined with a decision-making algorithm to predict potentially highlight-able elements in the input text.

### Modeling Pitch Accent Variation

There are two families of approaches to pitch accent placement, mainly defined by their particular application in various speech technologies. On the one hand, prosodic components in a Natural Language Generation (NLG) system can benefit from clearly defined and obtainable syntactic, semantic, discourse, and information structure to derive carefully-crafted accent patterns. Prosodic components in more general-purpose TTS or ASR systems, on the other hand, do not have such information available, and must therefore rely on approximating linguistic structure as a distribution.

### Rule-based Systems for Accent Placement

Meaning-to-Speech, also known as Concept-to-Speech (CTS) systems are generally implemented as extensions to language generation components in spoken dialog information systems. They differ from TTS systems in that source input is not *textual*, but *conceptual* – that is, the input consists of a deep-structure semantic representation. In this way, discourse, information structure, and surface structure of the text can be

## 2.3 Computational Models of Accent Variation

---

directly accessed during parameterization for wave synthesis, and consequently, accent assignment.

One system proposed by Prevost & Steedman (1994) employs Combinatory Categorical Grammar (CCG) as a grammatical framework for unifying syntax with prosody. In their model for database-driven query applications, prosodic categories are defined as combinatorial functions in parallel with syntactic categories. Specifically, intonational boundaries are designated as arguments, with pitch accents as the (typed) functions over them. Pitch accents consist of  $L+H^*$ , whose type identifies the constituent as the *theme*, and  $H^*$ , whose type identifies the *rheme*. Thus, information structure is constructed as an analysis of semantic propositions rather than syntactic constituents, and is tightly integrated with prosody. As such, the system is better poised to handle contrastive stress and discourse-elicited emphasis (Prevost & Steedman, 1993). The result is a much more context-appropriate intonational pattern.

Other systems take a more flexible approach via rule-based pitch accent assignments based on deep linguistic parses of generated content. In particular, Pan & Mckeown (1999) utilize the FUF/SURGE grammar – useful for supplying semantic roles in a description which might answer questions such as when/where/how/why – then feeding semantic information into a rule induction system (RIPPER) for rule-based intonation. Alter *et al.* (1996) likewise introduced a system for German that generates strategic focus types which are fed into a tactical generator based on FUF/HPSG. A phonological component is then used to interpret grammatically derived pitch accents into tones such as High-Low ( $H^*L$ ) or Low-High ( $LH^*$ )<sup>1</sup>.

These systems, while better able to construct contextually appropriate pitch patterns, are nevertheless extremely limited by their domain of application. Not least is the inability of deep grammars to robustly handle unrestricted text as demanded by general purpose TTS applications (i.e. uses other than expert systems). CTS systems, moreover, have been more or less limited in the scope of discourse types and intonational tunes, an area which statistical methods can improve.

### Statistical Accent Detection and Prediction

Recently, focus has shifted to developing more robust systems for use in TTS. These systems often make use of machine learning techniques to provide general purpose prosody

---

<sup>1</sup>The reader is referred to section 4.1.1 for discussion of the GToBI annotation scheme for German.

## 2.3 Computational Models of Accent Variation

---

for open-domain, unrestricted input. In general, deep linguistic processing is abandoned in favor of ‘lightweight’ linguistic variables that can be used indicators of accent.

Given the variety of potential variables as predictors, early work investigated the use of models which can handle many correlated features. Most of these approaches have centered on predicting the accent of each word independently given a set of features, in particular using machine learning techniques such as Decision Trees (Hirschberg, 1993), bagging and boosting (Sun, 2002), Gaussian Mixture Models (Chen *et al.*, 2004), and other more modern discriminative frameworks such as Maximum Entropy Markov Models (MEMMs) (Sridhar *et al.*, 2008b) and Support Vector Machines (SVMs) (Levow, 2005). These stochastic models, however, inherently fail to account for context between accent labels, which can be indicative of deletion processes from compound collocations.

Rule-based approaches, including hand-written rules (Hirschberg, 1993), as well as rules learned via rule-induction systems (Pan & Mckeown, 1999) overcome this limitation, yet have their own set of drawbacks as well. Specifically, rule-based systems make absolute decisions in an area which is non-rigorous. While distinct patterns of deaccentuation exist, some non-conformist patterns may still be judged acceptable by human listeners (Hirschberg, 1993; Liberman & Sproat, 1992). This suggests that a weighted approach to rule application may be more suitable.

Other statistical approaches manage to incorporate contextual accent with weighted application using sequence learning models. A sequence learning model optimizes the likelihood of a sequence of accent labels given a corresponding input sequence of text, such that label context is incorporated into the model in a principled way. Pan & Mckeown (1998), for example, showed that using Hidden Markov Models (HMM) for pitch accent prediction outperforms a rule-induction based system (textscRIPPER) given the same set of variables. As a probabilistic formalism, however, textscHMMs are limited to very few variables in the model. This is due in part to the training method (models are trained non-discriminatively to model the joint probability of input and label sequences), as well as certain underlying assumptions (observation input is assumed to be independent to achieve efficient inference).

More recently, researchers in this area have adopted discriminative learning methods because of their ability to incorporate a large set of correlated, dependent features within a sequence learning model. In particular, Conditional Random Fields (CRF) were introduced as a comprehensive model for pitch accent prediction in conversational and

## 2.3 Computational Models of Accent Variation

---

read speech, respectively (Gregory & Altun, 2004; Levow, 2008). As a machine learning technique for accent prediction, CRFs were shown to outperform HMMs given the same set of input variables (Gregory & Altun, 2004). In addition, apart from superior performance at the baseline, CRFs have the advantage of optimizing the entire sequence of accent labels given a rich set of correlated, inter-dependent, potentially long distance features on the observation input.

In the following chapters, we investigate the use of a CRF-based model for predicting pitch accent in German and English. In particular, we will explore the use of features approximating discourse structure and semantic context.

## Chapter 3

# Model Design

The design of this model is fundamentally based on two underlying assumptions which have emerged from the discussion on accent variation. The first is that pitch accent is context sensitive; while many kinds of word characteristics, from word category to its position and length, contribute to accented status, contextual information about intonational boundaries, word categories, lengths, and even the accent status of neighboring words, consistently influence the overall accenting strategy. The second is that discourse and semantic meaning play an important role in the outcome of accent patterns.

In order to capture aspects of context and discourse in a statistical framework, we outline an approach to discourse-driven statistical accent prediction based purely on textual analysis. In particular, this model adopts a sequence-learning approach to accent likelihood, making use of discriminative modeling techniques which allow a large set of targeted, inter-dependent, long-distance features to be used as parameters in the probability model. Furthermore, we introduce a set of features specifically designed to capture both explicit and implicit elements of the discourse for the purpose of predicting context-appropriate accent.

This chapter begins with an overview of the statistical framework in question, with subsequent discussion on the design of discourse-based features to be used in the model. In particular, we introduce several features related to semantic meaning, as well as local and global discourse, which can be easily extracted from the input text.



### 3.1 The Stochastic Model

Following recent work in statistical pitch accent prediction (Gregory & Altun, 2004; Levow, 2008), this system takes a sequence learning approach to modeling accent variation. Specifically, this approach models an entire sequence of accent labels from an input sequence of words. Apart from this, discriminative modeling techniques, which have been shown to outperform traditional generative models in this task (Gregory & Altun, 2004), form the basis for accent classification.

In particular, our model is based on Conditional Random Fields, first introduced by Lafferty *et al.* (2001). In recent years, CRFs have been used in a wide variety of natural language processing applications, including POS-tagging (Lafferty *et al.*, 2001), NP-chunking (Sha & Pereira, 2003), and Named-Entity Recognition (McCallum, 2003), among others. Although many other discriminative or regression modeling techniques might be considered, most notably MEMMs or SVMs, and are often better at classifying individual events (e.g. predicting a letter given a set of pixels from an image), the strength of CRFs lies in its ability to classify labels across an entire sequence (e.g. predicting the word from likely letter candidates) (Hoefel & Elkan, 2008).

#### Conditional Random Fields

CRFs are a framework for defining a conditional distribution as an undirected graphical model (Lafferty *et al.*, 2001). Graphical models have proven a useful tool in natural language processing, representing what Jordan (1999:1) describes as “a marriage between probability theory and graph theory,” in which graph nodes stand in as the random variables in a system (such as linguistic entities in a natural language grammar) and edges, or the lack of edges, represent the dependencies and independencies, respectively, between variables. Although the associated graph in a CRF can be generalized to any form, the intuitive choice for modeling sequences is the special case in which the graphical model is acyclic and in the form of a chain, often referred to as a *linear-chain CRF*<sup>1</sup> (McCallum, 2003).

---

<sup>1</sup>We will, however, simply refer to this as a CRF for the remainder of this section.

### Generalized CRF

Adopting the original notation of (Lafferty *et al.*, 2001), let  $X$  be a set of random variables (the observed input), and  $Y$  be a set of random variables which are to be inferred by the model (the predicted output). Further, let  $G = (V, E)$  be a graph whose vertices in  $V$  are the random variables connected by undirected edges of  $E$ , and let  $C(X, Y)$  be the set of (maximal) cliques in the graph. By the Hammersley-Clifford Theorem (1971), if the distribution is strictly positive and the graph encodes conditional independencies, then the conditional distribution  $P(Y|X)$  is simply the product of potentials (or potential functions) on the cliques of the graph:

$$P_{\Lambda}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \prod_{c \in C(\mathbf{y}, \mathbf{x})} \Phi_c(\mathbf{y}_c, \mathbf{x}_c) \quad (3.1)$$

where  $\Phi_c(\mathbf{y}_c, \mathbf{x}_c)$  is the potential on clique  $c$ . Here, the potential function is defined as an exponential of the weighted sum of features over a clique, such that

$$\Phi_c(\mathbf{y}_c, \mathbf{x}_c) = \exp(\sum_{k=1}^K \lambda_k f_k(\mathbf{y}_c, \mathbf{x}_c))$$

Features are represented by a set of *feature functions*, defined as  $f_{y'}(y, x) = 1_{y'=y}$ , each of which is indexed by  $f_k$ , with corresponding feature weights  $f_{y',j}(y, x) = 1_{y'=y}x_j$ , each indexed by  $\lambda_k$  (Sutton & McCallum, 2007). The potential function is not normalized, however. Therefore, a normalization factor,

$$Z_{\mathbf{x}} = \sum_{\mathbf{y}'} \prod_{c \in C(\mathbf{y}', \mathbf{x})} \Phi_c(\mathbf{y}'_c, \mathbf{x}_c)$$

known as the *partition function*, is needed to arrive at a valid probability (McCallum, 2003).

### Linear-Chain CRFs

Now, let  $X = \{X_1, X_2, \dots, X_n\}$  be a sequence of words corresponding to a natural language sentence, and  $Y = \{Y_1, Y_2, \dots, Y_n\}$  be a corresponding sequence of pitch accent labels. Then, if the variables in  $Y$  are connected by edges in a linear chain, such that they obey the Markov<sup>1</sup> property with respect to the graph, then the conditional probability  $P(Y|X)$  is a linear-chain *conditional random field* (Lafferty *et al.*, 2001).

<sup>1</sup> $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$ , where  $w \sim v$  means that  $w$  and  $v$  are neighbors in  $G$ .

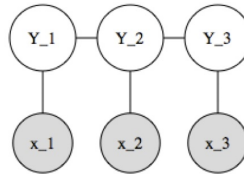


Figure 3.1: A linear-chain CRF

The potential over a clique includes state-state features,  $f_{ij}(y, y', \mathbf{x}) = \mathbf{1}_{y=i} \mathbf{1}_{y'=j}$ , which is the transition  $(i, j)$ , as well as state-observation features,  $f_{io}(y, y', \mathbf{x}) = \mathbf{1}_{y=i} \mathbf{1}_{x=o}$ . Thus, given the set of feature weights  $\Lambda = \{\lambda_k\}$  and the set of feature functions  $\{f_k(y, y', \mathbf{x}_t)\}_{k=1}^K$ , the conditional distribution of a linear-chain CRF becomes

$$P_{\Lambda}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \exp\left\{\sum_{k=1}^K \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\} \quad (3.2)$$

with the normalization function

$$Z_{\mathbf{x}} = \sum_{\mathbf{y}} P_{\Lambda}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \exp\left\{\sum_{k=1}^K \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\} \quad (3.3)$$

which is a summation over all possible label sequences (Sutton & McCallum, 2007). The challenge then is to estimate the parameters of the CRF model,  $\Lambda = \{\lambda_k\}$ , so as to maximize the log-likelihood of labeled data within a training set.

### Parameter Estimation

Several methods exist for parameter estimation, including iterative scaling (Lafferty *et al.*, 2001), and more efficient methods based on gradient descent (Sha & Pereira, 2003). In particular, limited-memory BFGS (Byrd *et al.*, 1994) has been shown to be

many times faster during training than iterative scaling or conjugate gradient methods (Sha & Pereira, 2003), and has become the method of choice for most implementations.

### Feature Functions

Features in a CRF model may be real-valued, but are typically binary, representing hand-crafted observational tests of the data. An example feature function might be

$$\mathbf{feature\ 1} \begin{cases} \text{current-word}=\text{"Bill"} \text{ and } \text{isNamedEntity} & 1 \\ \text{else} & 0 \end{cases} \quad (3.4)$$

which would return true if the current word ‘*Bill*’ is a named entity.

In this way, weighted features, expressing a wide variety of observations, effectively compete against each other across an entire observation sequence. Moreover, since the probability is determined across the entire sequence, input at all points in the sequence is available in the scope of a feature function. A common approach to feature design is therefore to make use of feature templates that effectuate observational tests in a sliding window<sup>1</sup>.

$$\mathbf{feature\ 2} \begin{cases} \text{current-word}=\text{"Bill"} \text{ and } \text{previous-word}=\text{"President"} & 1 \\ \text{else} & 0 \end{cases} \quad (3.5)$$

It is important to note that a feature is associated with a parameter weight for each label type, a value which is static in nature, as it represents static, reproducible phenomena. In contrast to this, discourse is inherently dynamic in nature, always defining a contextual environment.

## 3.2 Feature Design

We investigate potential features for approximating discourse in a static, observational manner suitable for our model. Discourse, as an abstract system, is dynamic and ever-changing. In order to capture elements of the discourse given the static parameters

<sup>1</sup>The typical length of a sliding window is 5:  $(x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2})$  where  $x$  is the current observation input.

of our prediction model, we must first quantify the domain. We therefore introduce a set of variables designed to approximate a constant contextual state from constantly updated discourse.

### 3.2.1 Semantic Space

For each element added to the conversation, there is more contextual information (i.e. “shared knowledge”) available to speaker and hearer which can potentially modify the expectations of pitch. We therefore introduce a fixed semantic space which is constantly updated apace with discourse. As a fixed space, it is similar to Grosz & Sidner’s (1986) stack of focus spaces, yet differs from focus in that the semantic space does not consist of abstract semantic objects, but textually-salient concepts and entities. In particular, the semantic space, which is constantly updated, consists of explicitly and implicitly defined (non-entity) lexical items, and explicitly defined entities.

Let  $L$  be the set of words  $w$  that occur in the language of the texts, and  $N \subset L$  contain all words that are names of entities. A semantic space  $\mathcal{S} = (E, C)$  contains a set of entities  $E = \{e_0, e_1, \dots, e_n\}$  and a set of concepts  $C = \{c_0, c_1, c_2, \dots, c_n\}$  where  $E \subset N$  and  $C \subset \mathcal{P}(L)$ <sup>1</sup>. Entities specifically refer to person names and are always given explicitly in the text. Concepts, in contrast, are defined as sets of semantically related terms, and are joined in the concept context ( $C$ ) of the semantic space as the discourse unfolds. Concepts consist of both explicit (*immediate*) terms (directly available from the text) and implicit (*extended*) terms (semantically evoked from an explicit term).

We then define the following feature functions regarding concepts and entities in the semantic space:

**Definition 1.** *Immediate Concept ( $C_I$ )*

*An immediate concept tests whether a given word  $w$  exists in the current context of concepts  $C$ .*

$$C_I(w, C) = w \in \bigcup_{c \in C} c \quad (3.6)$$

**Definition 2.** *Extended Concept ( $C_E$ )*

*An extended concept tests whether any word in a concept  $c'$  exists in the current context of concepts  $C$ .*

$$C_E(c', C) = \exists_c (c \in C \wedge c \cap c' \neq \emptyset) \quad (3.7)$$

---

<sup>1</sup> $\mathcal{P}(L)$  is the power set of  $L$ .

Thus, the semantic space in our model contains all immediate and extended lexical items that have accumulated at the time of discourse. In this way,  $C_I$  checks whether the input word at a given point in discourse already exists in the semantic space, while  $C_E$  checks whether this word evokes similar terms which exist in the current semantic space.

The definition of an immediate concept in our model is fairly straightforward: it is simply the current word  $w_i$  during text analysis which is neither a named entity, nor a function word. The definition of an extended concept is somewhat more arbitrary. For our purposes, the set of extended items are defined as the *k most related words of  $w_i$* .

The next step is to determine the set of  $k$  related words that make up an extended concept to  $w_i$ . Importantly, ‘relatedness’ need not entail part-of-speech matching, but can be defined as merely ‘associated relatedness’. In other words, the elements within a set of concepts must be related by their association with, or ability to evoke a particular semantic concept.

There are many ways to define lexical relatedness, each of which can be compared not only in terms of their respective strengths, but also in terms of their relative cost. A ‘cheap’ method, for example, might employ distribution-based clustering on a large set of documents, in which the similarity of words is calculated based on co-occurrences (cf. Brown *et al.*, 1991). Clustering is relatively inexpensive given the current processing power of computer chips and the ease in obtaining large amounts of digitized text<sup>1</sup> for analysis. An ‘expensive’ method might take advantage of expertly-crafted knowledge databases such as WordNet in English (Fellbaum, 1998), or equivalently, GermaNet in German (Hamp & Feldweg, 1997). Knowledge bases are costly in the sense that much is needed in terms of time and effort to develop such resources, making it difficult to obtain, as well to expand the same principle to other languages.

There are, of course, advantages and disadvantages associated with the cost. Expensive resources can make use of well-defined word relations such as synonymy/antonymy, hyponymy/hypernymy, meronymy/holonymy, etc., whereas distributional methods are subject to noise and are sensitive to domain. For our purposes, we chose a method that lies somewhere in between these two poles.

---

<sup>1</sup>For example, through resources like Wikipedia. <http://www.wikipedia.org>

### Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a mathematical-statistical method for inferring relations between words across a great number of ‘meaningful’ bodies of text. It has been used successfully in a number of NLP applications, including Information Retrieval (Dumais *et al.*, 1988), document classification (Zelikovitz & Hirsh, 2001), and word-sense disambiguation (Buckeridge & Sutcliffe, 2002). LSA has also been shown to simulate a variety of human cognitive abilities, such as acquiring subject matter knowledge<sup>1</sup> (Landauer *et al.*, 1998) and assessing the quantity and quality of essays (essay-scoring)<sup>2</sup> (Foltz *et al.*, 1999). LSA is also useful for modeling a variety of word relations, especially synonymy and polysemy (Landauer *et al.*, 1998).

Rather than calculating surface co-occurrences of words, LSA approximates the meaning of words by averaging the *effect* of a given word on the meaning of the document in which it occurs. In this way, two words may be quite similar even if they have never occurred together in the same text. These are the ‘latent’ similarities which are available as part of some hidden semantic structure.

LSA is fundamentally based on a term-document matrix, in which rows represent individual terms (words) and columns represent coherent bodies of text (paragraphs, articles, essays, etc.). In its most basic construction, the cells of the matrix represent the absolute counts of each word in a given document. Singular Value Decomposition (SVD), a well-known matrix-algebra transformation, is used to decompose the matrix into three separate matrices, one of which roughly corresponds to term coordinates, and another to document coordinates. Finally, dimension reduction of the tripartite decomposition casts these term and document vectors into a  $k$ -dimensional space, allowing the strongest relations between terms and documents to emerge (Landauer & Dumais, 2008).

Formally, given a rectangular matrix  $X = t \times d$  of terms and documents, then

$$X = T * S * D^T$$

<sup>1</sup>An LSA model trained on psychology textbooks was assessed on multiple choice tests against both an expert in the field and a novice who had read the same books. LSA results had a higher correlation with expert than novice scores.

<sup>2</sup>LSA models of student essays were assessed against expert graders with roughly the same correlation as between expert graders.

is the SVD of  $X$ , where  $T$  is an orthonormal matrix of left singular vectors (term vectors),  $D$  is an orthonormal matrix of right singular vectors (document vectors), and  $S$  is a diagonal matrix of singular values. If only the  $k$  largest singular values are retained with their corresponding singular vectors, then it is the rank  $k$  *approximation* to  $X$  with the smallest error. We therefore compute the SVD to the  $k^1$  reduced dimension

$$X \approx T_k * S_k * P_k^T$$

which effectively translates the term and document vectors into a *concept space* to which common similarity measures may be applied.

As a simple illustration of this process, consider the following set of documents from which we might build our initial term-document matrix<sup>2</sup>:

- d1: Shipment of gold damaged in a fire.
- d2: Delivery of silver arrived in a silver truck.
- d3: Shipment of gold arrived in a truck.

Ignoring for the moment punctuation and capitalization, the corresponding counts of words across all documents would be the following:

---

<sup>1</sup>The value of  $k$  is arbitrary, and often decided through trial and error. An optimal value has been noted at 300, with a useful range between 200–2000, with the suggestion that corpus sizes exceed  $\sim 20k$  unique words or documents (Landauer & Dumais, 2008).

<sup>2</sup>Example from (Grossman & Frieder, 2004:71).



	$d_1$	$d_2$	$d_3$
a	1	1	1
arrived	0	1	1
damaged	1	0	0
delivery	0	1	0
fire	1	0	0
gold	1	0	1
in	1	1	1
of	1	1	1
shipment	1	0	1
silver	0	2	0
truck	0	1	1

Table 3.1: Term-document matrix of absolute counts.

From this original table of counts, we construct a matrix which can be decomposed via the SVD into the components

$$T \begin{bmatrix} -0.4201 & 0.0748 & -0.0460 \\ -0.2995 & -0.2001 & 0.4078 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.1576 & -0.3046 & -0.2006 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.3151 & -0.6093 & -0.4013 \\ -0.2995 & -0.2001 & 0.4078 \end{bmatrix}$$

$$S \begin{bmatrix} 4.0989 & 0.0000 & 0.0000 \\ 0.0000 & 2.3616 & 0.0000 \\ 0.0000 & 0.0000 & 1.2737 \end{bmatrix} D^T \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \\ -0.5780 & -0.2556 & 0.7750 \end{bmatrix}$$

In order to reduce noise and artifacts from word usage in documents, the dimensionality of these components are reduced such that together they represent an approximation of  $X$ . If only the first two singular values are kept ( $k = 2$ ), along with their

corresponding left and right values, then this is the *Rank 2 Approximation*

$$T_k \begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & -0.2001 \\ -0.1206 & 0.2749 \\ -0.1576 & -0.3046 \\ -0.1206 & 0.2749 \\ -0.2626 & 0.3794 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3794 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix} S_k \begin{bmatrix} 4.0989 & 0.0000 \\ 0.0000 & 2.3616 \end{bmatrix} D_k^T \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \end{bmatrix}$$

Each word and each document can now be described as a 2-factor analysis, where ‘*silver*’ corresponds to the factors  $(-0.3151, -0.6093)$  and  $d_3$  corresponds to  $(-0.4945, 0.6492)$ . These factors can further be viewed as coordinates in a two-dimensional space.

In this way, standard distance measures, such as Euclidean distance or cosine similarity, may be used to measure the similarity of term-term, term-document, and document-document vectors. Standard clustering techniques such as K-Nearest-Neighbor may also be used to find sets of synonyms, sets of related documents, and more.

In our model, we define the relatedness of words in terms of term-term cosine similarity, as given in 3.8:

$$\text{sim}(\hat{\mathbf{t}}_i, \hat{\mathbf{t}}_j) = \frac{\hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_j}{|\hat{\mathbf{t}}_i| |\hat{\mathbf{t}}_j|} \quad (3.8)$$

The term-document counts listed in 3.1 represent the most basic assumptions about the contextual usage of words<sup>1</sup> More sophisticated frequency measures are often used

<sup>1</sup>More sophisticated models often make use of preprocessing steps such as punctuation stripping, stemming, and stop-list word removal to compensate for noise due to word variation and counts of extremely common words.

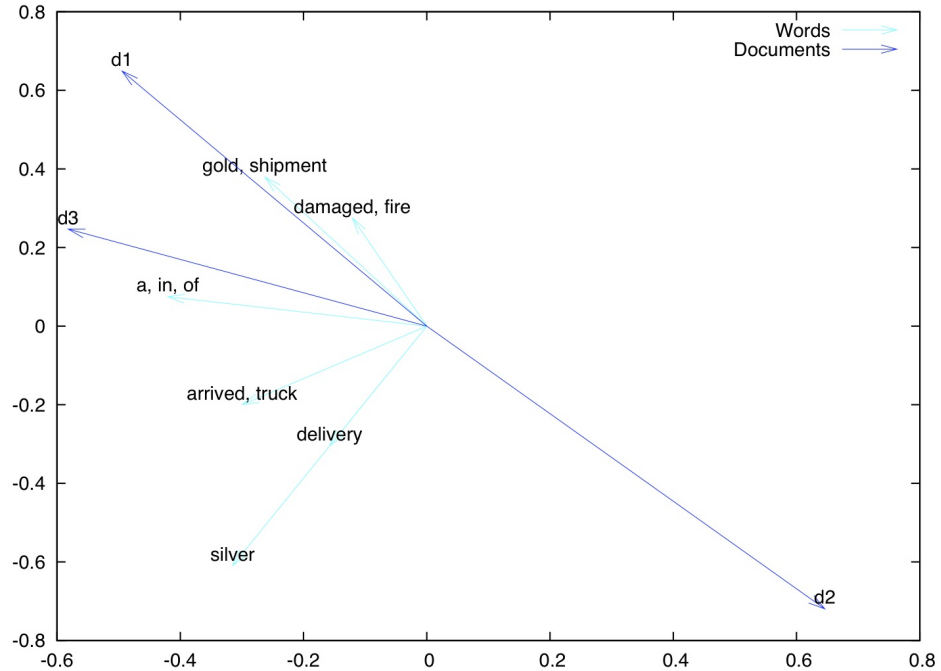


Figure 3.2: Word and document factors as coordinates in a two-dimensional space

to account for biases in document length and contextual word usage. Specifically, term-weighting schemes are used in the place of absolute counts, and typically consist of the components

$$L_{ij}G_iN_j$$

where  $L_{ij}$  is a measure of the local frequency of a term  $i$  in document  $j$ ,  $G_i$  is a global weight of the word  $i$  across all documents, and  $N_j$  is a normalization factor to compensate for different lengths in texts (Chisholm & Kolda, 1999). In our model, we used the weighting scheme

$$L_{ij} = \begin{cases} 0.2 + 0.8 \log(f_{ij} + 1) & f_{ij} > 0 \\ 0 & f_{ij} = 0 \end{cases} \quad (3.9)$$

$$G_i = 1 + \sum_{j=1}^N \frac{\frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i}}{\log N} \quad (3.10)$$

$$N_j = 1 \quad (3.11)$$

which includes a local weighting score, introduced by (Chisholm & Kolda, 1999:9), that uses a log score rather than within-document frequency, as well as a global weight based on a measure of entropy (which gives higher weights to rare words), with no normalization of document lengths.

In this way, as a term  $t_i$  is processed in the discourse, we extract the  $k$  most similar terms to  $t_i$  as its extended concept.

### 3.2.2 Local and Global Discourse

A semantic space is a progression over the discourse of a text, yet is asymmetric with respect to the discourse. For one thing, lexical ambiguity restricts the relation of elements within a concept to a single topic or subject<sup>1</sup>. A word such as *pupil* entails a different set of related terms in the context of school (e.g.  $\{student, education, study, \dots\}$ ) than it would in the context of anatomy (e.g.  $\{eye, iris, see, \dots\}$ )<sup>2</sup>. Conversely, entities have a fixed interpretation across all subjects in a discourse. For example, while the discussion change from ‘*John’s dogs*’ to ‘*Mary’s party, to which John is invited*’ evokes different sets of concepts, the interpretation of ‘*John*’ remains constant in both. Also, accenting patterns are different in each case: entities are often consistently accented as they tend to represent new or re-enforced information, with subsequent mentions being pro-nominalized before deaccent (cf. section 2.2 for the discussion on *centering*); subsequent mention of concept items, on the other hand, does not induce a change of form before deaccentuation.

<sup>1</sup>We use the term *subject* to side-step the conflicting definitions of *topic*. On the one hand, it is a structural term (e.g. *topic-comment*); on the other, it can refer to the subject matter of an entire text (e.g. as used in *Topic-Detection-and-Tracking* (cf. Allan *et al.*, 1998)).

<sup>2</sup>Equally in German, the word *Tau* used in the nautical sense (‘rope’) might have the associated concept  $\{Schiff, Poller, segeln, \dots\}$  while in the atmospheric sense (‘morning dew’), the concept might be more like  $\{Morgen, Gras, benetzen, \dots\}$ .

Adapting the definition from (Grosz & Sidner, 1986), we distinguish *local* from *global* discourse, such that the semantic space persists across the global discourse, while its contents are bound to the local discourse.

Let  $T = \langle d_1, d_2, d_3, \dots, d_n \rangle$  be a text of ordered<sup>1</sup> discourse subjects<sup>2</sup> where a discourse subject  $d = \langle s_1, s_2, \dots, s_n \rangle$  is an ordered list of sentences, and a sentence  $s = \langle w_1, w_2, \dots, w_n \rangle$  is an ordered list of words. A semantic space  $\mathcal{S} = (E, C)$  can be amended by adding entities to  $E$ , which we denote as  $E(e)$ , or by adding concepts to  $C$ , denoted as  $C(c)$ . As a shorthand, we use  $\mathcal{S}(e)$  and  $\mathcal{S}(c)$  for the state  $\mathcal{S}$  modified by adding the entity  $e$  to  $E$  and the concept  $c$  to  $C$ , respectively.

A text  $T$  is processed by iterating over the list of discourse subjects, each of which in turn is processed by iterating over the list of sentences within the discourse subject. Similarly, each sentence is processed by iterating over the list of words within the sentence. As a text is processed, a discourse subject is evaluated in the space  $\mathcal{S}$ , which we write as

$$\Gamma \vdash \mathcal{S}() : d \Rightarrow \Gamma \vdash \mathcal{S}(d) \quad (3.12)$$

Here,  $\Gamma$  denotes the environment with memory<sup>3</sup> where the evaluation of each fragment affects the environment for the evaluation of the next. The process is thus a simple progression in which elements can affect the space

$$\begin{aligned} \Gamma \vdash \mathcal{S} : \langle d_1, d_2, \dots, d_n \rangle \\ \rightarrow \Gamma \vdash \mathcal{S}(d_1) : \langle d_2, \dots, d_n \rangle \\ \rightarrow \dots \\ \rightarrow \Gamma \vdash \mathcal{S}(d_1, d_2, \dots, d_{n-1}) : \langle d_n \rangle \end{aligned} \quad (3.13)$$

Importantly, concepts and entities may be added to the semantic space while iterating over words, but only entities may be added to the space while iterating over

<sup>1</sup>Grosz & Sidner (1986) proposed nested discourse; we adopt an ordered list for simplicity.

<sup>2</sup>We use the term *discourse subject* in the sense of ‘aboutness’ of the text, as opposed to the intention behind it suggested in the use of *discourse purpose* (Grosz & Sidner, 1986).

<sup>3</sup>This notation is commonly used in computer science.

sentences or discourse subjects.

$$\begin{aligned}
\Gamma \vdash (C, E) : \langle d_1, d_2, \dots, d_n \rangle \\
\rightarrow \Gamma \vdash (C, E(d_1)) : \langle d_2, \dots, d_n \rangle \\
\rightarrow \dots \\
\rightarrow \Gamma \vdash (C, E(d_1, \dots, d_{n-1})) : \langle d_n \rangle
\end{aligned} \tag{3.14}$$

$$\begin{aligned}
\Gamma \vdash (C, E) : \langle s_1, s_2, \dots, s_n \rangle \\
\rightarrow \Gamma \vdash (C(s_1), E(s_1)) : \langle s_2, \dots, s_n \rangle \\
\rightarrow \dots \\
\rightarrow \Gamma \vdash (C(s_1, \dots, s_{n-1}), E(s_1, \dots, s_{n-1})) : \langle s_n \rangle
\end{aligned} \tag{3.15}$$

Thus changes in discourse subject cause the set of concepts to be reset, while the set of entities remains intact. In this way, the set of concepts may be said to grow locally with every new discourse subject, just as the set of entities grows globally across an entire text.

$$\Gamma_d \vdash \mathcal{S}() : d \Rightarrow \mathcal{S}(d) \tag{3.16}$$

Here,  $\Gamma$  denotes the environment with memory<sup>1</sup> where the evaluation of each fragment affects the environment for the evaluation of the next. The process is thus a simple progression in which elements can affect the space

$$\begin{aligned}
\Gamma_d \vdash \mathcal{S} : \langle d_1, d_2, \dots, d_n \rangle \\
\rightarrow \Gamma_d \vdash \mathcal{S}(d_1) : \langle d_2, \dots, d_n \rangle \\
\rightarrow \dots \\
\rightarrow \Gamma_d \vdash \mathcal{S}(d_1, d_2, \dots, d_{n-1}) : \langle d_n \rangle
\end{aligned} \tag{3.17}$$

Importantly, concepts and entities may be added to the semantic space while iterating over words in  $\Gamma_s$ , but only entities may be added to the space while iterating over

---

<sup>1</sup>Notation as is commonly used in computer science.

sentences or discourse subjects in  $\Gamma_d$ .

$$\begin{aligned}
 \Gamma_d \vdash (C, E) : \langle d_1, d_2, \dots, d_n \rangle \\
 \quad \rightarrow \Gamma_d \vdash (C, E(d_1)) : \langle d_2, \dots, d_n \rangle \\
 \quad \rightarrow \dots \\
 \quad \rightarrow \Gamma_d \vdash (C, E(d_1, \dots, d_{n-1})) : \langle d_n \rangle
 \end{aligned} \tag{3.18}$$

$$\begin{aligned}
 \Gamma_s \vdash (C, E) : \langle s_1, s_2, \dots, s_n \rangle \\
 \quad \rightarrow \Gamma_s \vdash (C(s_1), E(s_1)) : \langle s_2, \dots, s_n \rangle \\
 \quad \rightarrow \dots \\
 \quad \rightarrow \Gamma_s \vdash (C(s_1, \dots, s_{n-1}), E(s_1, \dots, s_{n-1})) : \langle s_n \rangle
 \end{aligned} \tag{3.19}$$

When  $\mathcal{S}$  is modified by a word  $w$  at the base level, then the set of entities  $E$  is changed if the word is an entity, or the set of concepts  $C$  is changed if there is a mapping from  $w$  to an immediate concept  $c$ , and there is a mapping from  $c$  to an extended concept  $c'$ .

$$\mathcal{S}(w) = \begin{cases} (C, E(w)) & w \in N \\ (C(c, c'), E) & \text{if } (\mathcal{S}, w) \mapsto c \text{ and } (\mathcal{S}, c) \mapsto c' \end{cases} \tag{3.20}$$

In our system, the mapping from an immediate concept to an extended concept is achieved through our LSA model, as described in section 3.2.1.

This manner of resetting the semantic space thus associates ‘given’ information with the given discourse, while allowing new discourse subjects to have few presuppositions. An online system must therefore clearly demarcate the boundaries of discourse subject to effectively make use of the semantic space. In an NLG component, this can be easily supplied through context as the domain is fairly well controlled. In unrestricted TTS domains, however, an estimation of discourse subject boundary would be necessary. Recent work in Topic Detection and Tracking (TDT) shows promising results for statistically tracking changes in discourse, such as in online news streams. See (Allan *et al.*, 1998) and (Allan, 2002) for more detailed discussion on the topic.

### 3.2.3 Discourse Structure

Sentence structure and discourse timing can affect accenting strategy in noticeable ways. Syntactic structures like noun phrases (NP) and conjunctive phrases (CP) can be additional indicators for accenting phenomena such as compound accent deletion (i.e.

consider the sentence ‘*He went to [the CITY hall TAX office]<sub>NP</sub> to file papers.*’) and contrastive accent (e.g. ‘*Are you [going to the STORE or grabbing LUNCH?]<sub>CVP</sub>*’), respectively. The relative point in the discourse can also indicate the potential for accent; aggressive accenting, for instance, often occurs at the very beginning of a discourse when less background context is available, whereas accent deletion tends to occur towards the end of a discourse.

We therefore introduce into our model shallow structural and positional information which can easily be extracted from the text. Phrase structure information, for example, can be easily estimated through *chunking*, a shallow parsing method that is used to identify simple syntactic phrases.

Additionally, we introduce discourse-level positional features to track the newness of an utterance in relation to the current discourse. Specifically, we include the position of the utterance with respect to previous utterances in the discourse (that is, the position of the sentence in which the word occurs relative to previously uttered sentences), along with the usual variable for the position of the word in the utterance.



## Chapter 4

# Data Acquisition

In order to investigate pitch accent prediction, we train a CRF model on a corpus of transcribed speech that has been annotated with pitch accent labels. Similarly, in order to investigate the use of a semantic space in accent prediction, we train an LSA model on a collection of topically-coherent documents designed to approximate “world knowlege” in our system.

This chapter introduces the corpora used in training each of these models for German and English. We will first present two sets of annotated transcribed speech, the first consisting of single-speaker read news articles in German, and the second of spontaneous dialog in (British) English. We discuss the origins of these data sets and their prosodic annotation, and give an analysis of the data therein. We also look into annotations beyond the prosodic as potential features in the CRF model, while illustrating how these are extracted in an online system. We further describe a set of annotations in our German dataset that explicitly define discourse referents and their information status, which can be used to provide a standard with which to compare statistical approximations of an equivalent model. Finally, we introduce the corpora used in training LSA components for German and English.

### 4.1 Corpora for CRF Training

#### 4.1.1 The MULI Corpus

The MULti-Lingual Information structure (MULI) project (Baumann *et al.*, 2004a) aimed at providing researchers with empirical evidence for comparative studies of in-

formation structure in German and English. MULI builds on top of existing linguistic resources by enriching standard treebanks with additional annotation for information structure. In particular, the MULI corpus offers a rich set of syntactic, discourse-level, and prosodic structure for transcribed read speech of news articles.

The MULI corpus is based on extracts from both the Penn Treebank (Marcus *et al.*, 1994) in English, with articles derived from the *Wallstreet Journal*, and from the TIGER Treebank (Brants *et al.*, 2002) in German, with articles selected from the economics section of the *Frankfurter Rundschau*. Prosodic annotation, however, was only completed for German; accordingly, we concentrate solely on the German corpus for the remainder of this section. This corpus consists of about 250 sentences (approx. 3,500 words) read aloud by a native speaker. A subsection of this, comprised of 22 texts, averaging 9 sentences apiece and 170 words per text, was prosodically labeled following the MULI annotation system.

### The MULI Annotation System

The MULI annotation system features more varied and diverse prosodic labels than its English counterpart. In particular, annotations exist not only at the prosodic level, but at the syntactic and discourse level as well (Baumann *et al.*, 2004b).

**syntactic** At the syntactic level, information such as POS, morphology and syntactic structure are encoded as part of the original TIGER treebank. Beyond these, the MULI annotation scheme covers interesting syntactic phenomena on the clausal unit, typically involving syntactic structures designed to put focus on certain elements. These include *clefts*, *pseudo-clefts*, *reversed pseudo-clefts*, *extraposition*, *expletives*, and *active*, *medio-passive*, and *passive* voice.

**discourse** Information structure is encoded on the level of discourse referents and their properties, information status, as well as anaphoric associations across expressions. Regarding discourse referents, annotations capture *type* (intensional or extensional object, property, eventuality, or textuality); *semantic sort*; *delimitation* (unique, existential, variable, or non-denotational use) and *quantification* (uncountable, unspecific, non-singular, specific-nonsingular, or specific-singular). Other properties include the *form* of an expression and *information Status* (new, unused, inferable, evoked). *coreference* and *bridging* is distinguished, as well as

links between anaphora and antecedents, including *set-containment*, *part-whole composition*, *property-attribution*, *possession*, *causality*, and *lexical-argument filling*.

**prosodic** Prosodic annotation of the German corpus follows the conventions of GToBI, the de-facto standard for annotating German intonation (Baumann *et al.*, 2001). GToBI is slightly modified from the original ToBI prosodic labeling system. Table 4.1 illustrates the GToBI inventory of tones and breaks. Annotation tiers include word boundaries and pauses, punctuation, pitch accent and boundary tone, position and strength of phrase breaks, and rhythmic phenomena such as non-canonical word stress.

Pitch Accents	H*, L*, L+H* L*+H, H+!H*, H+L*
Force Accents	H(*), L(*)
Boundary Tones	L-, H-, L-% H-%, H-Ĥ% L-H%, %H
Break Indices	2r, 2t, 3, 4

Table 4.1: GToBI prosodic labels

As an example of MULI annotation, the complete transcription<sup>1</sup> of the utterance ‘*Exporte in den Libanon sichert Bonn derzeit nur kurzfristig ab*’ is given in figure 4.1, illustrating all tiers of prosodic information.

### The German Data Set

The dataset used to train the German system for automatic pitch prediction consists of the 22 read news articles from the MULI corpus for which semantic and discourse-level annotations are available. For our purposes, we only extracted word-level transcriptions with their corresponding pitch accent and intonational phrase boundary annotations, discarding all other prosodic information.

Moreover, several modifications were made to the dataset to ensure that each word bears only a single accent. First, in cases where a bi-tonal pitch accent straddled two

<sup>1</sup>Taken from Baumann *et al.* (2004b). <http://www.coli.uni-saarland.de/projects/muli>.

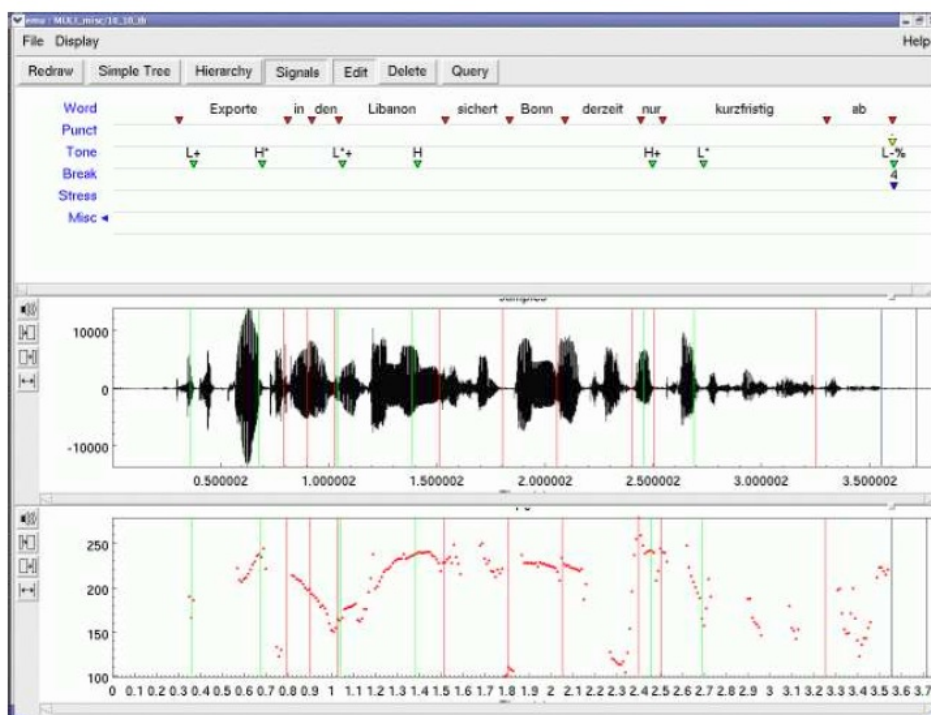


Figure 4.1: The complete MULI transcription of ‘*Exporte in den Libanon sichert Bonn derzeit nur kurzfristig ab*’

words (i.e. each word was marked as having only half an accent), the accent tone was re-joined and placed on the word bearing the peak of the accent (as denoted by ‘\*’). In this way, *Nach/L+ Strafaktion/H\** becomes *Nach/O Strafaktion/L+H\**, and *Vertrauen/L\* in/+H* becomes *Vertrauen/L\*+H in/O*. Secondly, acronyms or long numbers often carried two or more pitch accents. For this reason, long words were decomposed into hyphenated units, with pitch accents being placed on the prominent unit according to the audio signal. Thus singular instances of *rtr/{H\*  $\hat{H}$ +L\*}* become *r-/H\* t-/O r/ $\hat{H}$ +L\** and similarly *dreihundertvierzigtausendsiebenhundert/{L+H\* L+H\*}* becomes *dreihundertvierzigtausend-/L+H\* siebenhundert/L+H\**.

At the semantic and discourse structure level, the MULI corpus contains a great deal of annotation. We distinguish between shallow annotations which can be derived from off-the-shelf components<sup>1</sup> and used in an online system, and annotations rooted in a theoretical framework for which automatic parsers do not yet exist.

Regarding the former, these annotations include STTS-style POS tags (Schiller *et al.*, 1999), phrase structure based on the TIGER treebank (Brants *et al.*, 2002), and named entities derived from the POS tag set. In addition, the discourse subject is defined as the scope of the article in our model. Table 4.2 and 4.3 show the distribution of lexical categories and linguistic entities, respectively, in the training set.

	Accented	Unaccented
Nouns	84%	16%
Verbs	28%	72%
Function	5%	95%
Other	34%	61%

Table 4.2: Percentage of accented words by category in MULI

Annotations for more detailed syntactic information, including grammatical roles and sentence structure types, as well as discourse referents and their properties make up the set of theoretical features available. For our purposes, we focus on the presence of discourse delimitation and its information status (cf. section 2.2 for a description of information status in terms of Prince’s taxonomy of given/new). Table 4.4 gives counts of the different delimitations of discourse objects in the data set, along with the

<sup>1</sup>For German, these annotations are given in the MULI corpus.

Entities	Count
Noun phrases	389
Prepositional phrases	355
Conjunctive phrases	85
Intonational boundaries	653
Named entities	57
Utterances	244

Table 4.3: Counts of linguistic entities in MULI

percentage of accented objects given its corresponding information status at that point in the discourse.

It is important to note that the counts in 4.4 represent the percentage of accented objects on a per-word basis. That is, while each word in ‘*Die Pariser Softwaregruppe Cap Gemini Sogeti*’ is delimited as a unique object, some words are deaccented as a matter of prosodic rhythm, as in

(4.1) *Die/O Pariser/ACC Softwaregruppe/O Cap/O Gemini/ACC*

Information status may nevertheless be an important indicator when used in a model where many features are weighed against each other. Another interesting factor in the accenting of discourse referents can be observed in its position within the discourse. We noted, for example, that discourse referents in the beginning of the article tended to always be accented, regardless of information status, whereas discourse referents may be deaccented towards the end of the article. This is most likely due to the genre; in news articles, the most important information is usually contained in the first few sentences<sup>1</sup>.

<sup>1</sup>It was not possible to observe this effect in examples of spontaneous speech due to the lack of prosodically annotated speech over an entire discourse.

Delimitation	Count	Status	Accented	Example
Unique	869	brand-new unused inferrable text-evoked situation-evoked none	52% 57% 54% 56% 47% 43%	‘Frankfurt am Main’
Existential	807	brand-new unused inferrable text-evoked situation-evoked	52% 33% 47% 50% 56%	‘beide Konzerne’
Variable	334	brand-new inferrable text-evoked situation-evoked	59% 51% 63% 57%	‘vereinten Kräften’
None	31	brand-new inferrable none	60% 57% 28%	‘neunzehnhundert’

Table 4.4: Percentage of accented discourse objects in MULI

### 4.1.2 The IViE Corpus

The Intonational Variation in English (IViE) project was originally developed as a means of investigating intonational variation across dialects of English in the British Isles (Grabe *et al.*, 2001b). Speech samples were gathered from a range of English dialects in several speaking styles, and a flexible system for annotating the speech data with prosody was designed for its analysis.

The IViE corpus (Grabe *et al.*, 2001a) contains 36 hours of recorded speech of both male and female adolescent speakers from nine urban dialects: Belfast, Cardiff, Cambridge, Dublin, Leeds, Liverpool, Newcastle, Bradford (British Punjabi English), and London. For each of these dialects, five speaking styles were recorded. Among these include a set of Map task dialogs and a set of conversational dialogs on a given topic (Grabe, 2004). A cross section of this data, constituting about 1 hour of recorded speech, was selected for transcription and prosodic annotations (Grabe & Post, 2002) based on the IViE prosodic labeling system.

The Map task section records a goal-oriented interaction game modeled after Anderson *et al.* (1991). This data consists of 14 dialogs between single sex pairs from seven of the nine dialects: Belfast, Bradford, Cambridge, Dublin, Leeds, London, and Newcastle. Interactions from each region were taken both from female and male speaker pairs, with each dialog lasting about 1 minute. The Map task itself involves two speakers, separated by a screen or other object, who are each given a map of a small town. On the first speaker's map, a route is drawn around a number of buildings, with the name of each building clearly indicated. The second speaker's map shows only buildings, with some of the names of buildings having been changed in the interest of eliciting disagreement and discussion.

The free conversation section records face-to-face discussions on a given topic. This data consists of 14 dialogs with interactions between both male and female speaker pairs from the same seven dialects. The topic of 'smoking' was given in these conversations.

#### The IViE Labeling System

The IViE system (Grabe *et al.*, 1998) for prosodic annotation was modeled after TOBI, but augmented with additional tiers of annotation. The motivation behind the IViE system was to aid the transcription of rhythmic, phonetic, and phonological differences



in both standard and non-standard varieties of English. Prosody, therefore, is transcribed not on a single level, but on three tiers:

**rhythmic structure** The Prominence tier is intended to capture stressed and accented syllables, as well as rhythmic boundaries and hesitations. Prominence is marked with a ‘P’ in the middle of the accented syllable; rhythmic boundaries are marked with ‘%’ at the end of a word following a rhythmic boundary; and hesitations or speech errors are marked with ‘#’ at their locations.

**acoustic-phonetic structure** The Phonetic tier captures the pitch movement around accented syllables in an utterance. Pitch movements are realized within an Implementation Domain (ID) consisting of: (a) the preaccentual syllable; (b) the accented syllable; and (c) post-accentual syllables and the final syllable (if any) up to the next accented syllable. Phonetic transcription is realized from an inventory of six labels: H, M, L represent pitch levels on accented syllables, while h, m, l are used for unstressed syllables surrounding the accented syllable in an ID. Additionally, ‘-’ is used to mark an interpolation between the penultimate label and the final label in an ID. Finally, ‘%’ indicates the end of an ID coinciding with a rhythmic boundary.

**phonological structure** The Phonological tier provides the specification of intonation through pitch accent tones and phrase boundaries. An inventory of tone labels and intonational phrase boundaries are available, although not all labels are used for any given variety of English. The following tables list the possible tone and phrase boundary labels:

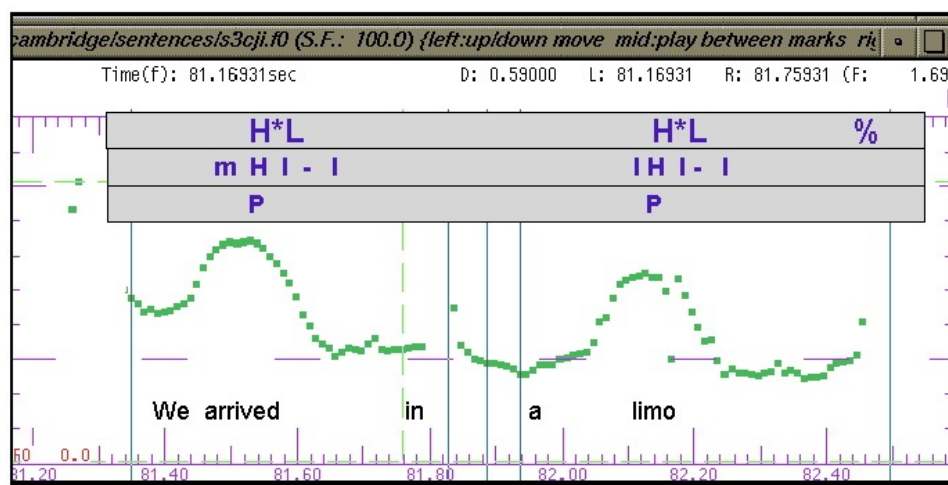
	<b>Tone Contour Description</b>
H*L	High target followed by low target, e.g. H-l, mH-l, mHl-l
H*	High target, e.g. lH-h
L*HL	Low target, followed by high target, low target, e.g. lLh-l
L*H	Low target followed by high target, e.g. mLh-h, mL-h, lL-h
L*	Low target
H*LH	High target on strong syllable, low, high, e.g. mHl-h

Table 4.5: IViE tone labels

Initial	Final	Boundary Specification
%H	H%	High Target
%	%	No Pitch Movement
%L	%L	Low Target

Table 4.6: IViE phrase-initial and phrase-final boundary labels

A complete IViE transcription<sup>1</sup> of the utterance ‘*We arrived in a limo*’ is given below. The blue outlined boxes represent an ID, while the shaded gray boxes illustrate the three tiers of annotation: phonological, phonetic, and rhythmic.

Figure 4.2: The complete IViE transcription of ‘*We arrived in a limo*’

<sup>1</sup>Taken from *The IViE Labelling Guide*. <http://www.phon.ox.ac.uk/IViE/guide.html>.

### The English Data Set

The dataset used in training an automatic pitch detection system for English comprises the subset of the IVIE corpus corresponding to spontaneous speech in dialog format. From this selection, we extracted word level transcriptions along with time-aligned annotations on the phonological tier only. These annotations included pitch accents and intonational phrase boundaries.

Other annotations include the current speaker, as well as syntactic information like POS and phrase chunks. These were obtained from off-the-shelf components and include Penn-style POS tags (Santorini, 1990) from TreeTagger (Schmid, 1994), and phrase boundaries from Chunkie (Skut & Brants, 1998). Named entity information was manually encoded.

The following tables show the distribution of linguistic entities in the selection of the IVIE corpus used for training. Table 4.7 shows the distribution of accented lexical categories. Table 4.8, in contrast, lists the number of linguistic entities such as syntactic phrases and named entities.

Category	Accented	Unaccented
Nouns	63%	37%
Verbs	42%	58%
Function	9%	91%
Other	50%	50%

Table 4.7: Percentage of accented words by lexical category

Other Entities	Count
Noun phrases	911
Prepositional phrases	238
Conjunctive phrases	-
Intonational phrases	844
Named entities	19
Utterances	342

Table 4.8: Counts of various kinds of linguistic entities

## 4.2 Corpora for LSA Training

### 4.2.1 The Wikipedia XML Corpus

Wikipedia<sup>1</sup> is an online collaborative encyclopedia consisting of over 12 million articles across 265 languages. Articles are freely edited by anyone with access to the website using the wiki markup language known as MediaWiki<sup>2</sup>. The Wikipedia XML Corpus (Denoyer & Gallinari, 2006) is a body of Wikipedia articles collected for 8 different languages which has been specially prepared for use in linguistic research. The collections, gathered in 2006, have been converted from the MediaWiki markup to the XML standard for efficient data processing.

	Document Count	Size (MB)	Mean Size (kB)
German	305,099	2.1	6.9
English	659,388	4.6	7.1

Table 4.9: Wikipedia XML Collections

As a web-based encyclopedia, each article in the collection clearly centers around a single subject. Moreover, each collection spans a richly diverse set of subjects that are relevant to modern culture. For this reason, it is an excellent resource for incorporating an element of ‘world knowledge’ in the construction of a semantic space.

A random subset of documents in each language was selected for training, numbering around 30,000 documents apiece. In each set, all xml formatting, hyperlinks, and punctuation were stripped from each document, and all words converted to lowercase for use in training the LSA component.

<sup>1</sup><http://www.wikipedia.org>

<sup>2</sup><http://en.wikipedia.org/wiki/MediaWiki>

## Chapter 5

# Experimental Setup

In order to empirically investigate the importance of discourse and semantic information in accent placement, we implement and train a CRF-based model for pitch accent prediction that incorporates features for approximating the semantic space of a given discourse, as well as its discourse structure, given the shallow textual analysis of an input text. Our experiments were conducted in the domain of speaker-dependent read speech (in German), and speaker-independent spontaneous speech (in English). Further, we compare the performance of automatically extracted discourse and semantic features with the performance of human-identified instances of discourse objects and information status for predicting pitch accent.

In this chapter, we first introduce the prediction task along with the feature sets we intend to investigate. We then describe our method for training and testing a model in order to evaluate its performance. Finally, we present our results for a number of experiments in German and English.

### 5.1 Preliminaries

In the following experiments, we focus on the binary classification task of predicting labels of *accented* or *unaccented*, given an input sequence of words. For this purpose, we collapsed the annotations for pitch accent tones in German and English into a single class specifying simply the presence of an accent.

	<b>Pitch Tone</b>	<b>Accent</b>
<b>German (MULI)</b>	H*, L* H*+L, H+L* L+H*, L*+H O	ACC  O
<b>English (IViE)</b>	H*, L* H*L, L*H L*HL O	ACC  O

Table 5.1: Collapsed accent labels

In addition, we include information for intonational boundaries by default in all of our experiments. In our model, phrase boundaries are incorporated into the input sequence, rather than as features for a given word of the input. We make use of annotations for intonational phrase boundary given by the respective corpus for German and English. For our purposes, we discard all information pertaining to tone of the boundary, preserving only the indication of the presence or absence of a tone.

	<b>Boundary Tone</b>	<b>Boundary</b>
<b>German</b>	H-, L-	%-
	H-%, L-%	%
	H- $\hat{H}$ %, L-H%	
<b>English</b>	%H, %L	-%
	H%, L%	%-
	%	%
	#	#

Table 5.2: Collapsed intonational phrase boundary labels

For the English data, we preserve information for phrase-initial (-%) and phrase-final (%-) boundaries, which may be important in the context of multi-speaker spontaneous speech. Speech errors (#) are likewise retained as a means of distinguishing unexpected breaks in syntax. In read aloud speech, the initial/final distinction is less important, and in fact, does not figure in the German corpus at all. Speech errors are similarly non-existent. This information therefore does not figure in the boundary labels for German.

We do, however, include the distinction between full intonational phrase boundaries (%) from intermediate (%-) boundaries.

Finally, we include information about the current speaker by default in the case of English. The current speaker is incorporated as a global feature on each word of the input sequence.

### Feature Sets

The sets of features we will explore in our experiments consist of, on the one hand, textually-extracted variables derived directly from the input text, and on the other hand, manually-annotated variables representing instances of discourse phenomena as identified in the literature. These feature sets are described below.

Boundary Tone	Boundary
Lexical (LX)	current-word
Syntactic (SYN)	part-of-speech
Discourse Structure (DS)	word-position
	sentence-position
	phrase
Semantic Space (SS)	named-entity
	immediate-concept
	extended-concept

Table 5.3: Feature sets for German and English systems

Table 5.3 outlines the sets of syntactic, semantic, and discourse features we extract for use in both the German and English systems. For the purposes of this work, we include only unigram information (i.e. the current word) in our set of lexical features<sup>1</sup>. In our experiments, the set of syntactic features consists only of part-of-speech.

The set of DS features includes the position of the current word relative to the start of the utterance, the position of the current utterance relative to the start of the discourse, and noun/prepositional phrase structure. The set of SS features include attributes for named entities, as well as immediate and extended concepts (cf. section 3.2.1 for definition of the semantic space).

<sup>1</sup>See (Gregory & Altun, 2004) for experiments using a more extensive set of lexical features.

Feature Set	Feature
Information Status (IS)	object-delimitation object-information-status
Information Status All (IS-ALL)	linguistic-form member object-delimitation object-information-status object-subtype type pointer proposition-subtype referential-link semantic-sort-of-object

Table 5.4: Feature set based on annotated information status for German

Table 5.4 outlines the set of IS features derived from theory-driven annotations of discourse referents and information status provided for German (cf. section 4.1.1 for a detailed description of discourse annotation in German).

### Training the CRF Model

Using the open-source c-based implementation *CRFSuite* (Okazaki, 2007), each model was trained as a first-order Markovian CRF. Parameter estimation was performed using Limited-memory BFGS (Byrd *et al.*, 1994).

	$L^1\sigma$
<b>German</b>	9.0
<b>English</b>	9.0

Table 5.5:  $L^1$  regularization  $\sigma$  for feature variance

To prevent over-fitting, optimization was performed using  $L^1$  regularization with a parameter weight for feature variance. This weight was determined from the maximum overall accuracy of the model when trained on 10 different values (from [1.0, 10.0]).



### Testing the System

For each system, model accuracy was determined using 7-fold cross validation. The dataset was divided into 7 parts across *article* or *dialog* boundaries, respectively. Models were trained on  $n - 1$  parts, with the  $n^{\text{th}}$  part used as held-out test data for measuring the performance of the model. Overall accuracy scores were obtained by averaging the macro-average score of each  $n$  model, and is provided as the final measure of performance for a model given its feature set.

Performance of a model is evaluated in terms of the well-known measures of precision (P), recall (R), and an F1 score.

$$P = \frac{\textit{TruePositives}}{(\textit{TruePositives} + \textit{FalsePositives})} \quad (5.1)$$

$$R = \frac{\textit{TruePositives}}{(\textit{TruePositives} + \textit{FalseNegatives})} \quad (5.2)$$

$$F1 = \frac{2 * P * R}{(P + R)} \quad (5.3)$$

Precision measures the portion of correctly assigned labels, while recall measures the portion of correctly assigned labels across all categories. The F1 score, a combined measure of precision and recall, gives an assessment of the overall performance of the model.

## 5.2 Results

We establish a number of baseline models with which to compare our results. As an absolute baseline, we trained a unigram model for both German and English. For German, the sole features of the model consisted of the input word sequence; in English, the base features were the input word sequence and current speaker. To this, we added a baseline of the unigram model with part-of-speech (POS), as well as a baseline with lexical class (BroadPOS). In the case of German, it is important to point out that POS features are based on gold-standard annotations, and do not reflect the typical error of online systems, as represented here in the case of English. Many systems, in fact, do not make use of the full set of POS features available. Rather, they rely on a reduced set roughly corresponding to lexical class (e.g. Noun, Verb, Function, Other) which can

presumably be predicted with higher accuracy. We therefore include results of both cases for analysis.

Model	Precision	Recall	F1-Score
Unigram	85.51	87.26	86.00
Unigram+POS	88.26	88.50	88.34
Unigram+BroadPOS	87.11	86.96	87.01

Table 5.6: Baseline scores for German

Model	Precision	Recall	F1-Score
Unigram	76.25	76.05	76.07
Unigram+POS	76.59	76.35	76.43
Unigram+BroadPOS	77.24	76.62	76.85

Table 5.7: Baseline scores for English

From Table 5.6, it is clear that POS plays a strong role in predicting accent in German, resulting in a 2.7% improvement over the unigram baseline. Lexical class features in German, however, perform roughly half as well with an 1.2% improvement. Conversely, from Table 5.7, the improvement in English was less marked (at 0.5%) given the full set of POS features, but improved twice as much (at 1%) on the reduced lexical class set. This is due, no doubt, to the characteristic errors of an off-the-shelf tagger.

In our first set of experiments, we take a look at discourse structure features, which include word and sentence position, as well as phrase structure information, in predicting accent.

Discourse structure features contributed to mixed results across data sets. In German, phrase structure and positional features turned out for the most part to degrade performance of the model when compared to the baseline with POS/BroadPOS. Although positional features appeared to improve recall slightly, the results were not statistically significant. This was likely due to the fact that phrase structure and word position provide only redundant information, the former via part-of-speech and the latter via the sequence learning model itself.

In English, the results were somewhat reversed. Phrase structure improved performance over the Baseline+POS, and was statistically significant at 0.90, although

Model	Precision	Recall	F1-Score
Baseline+POS	88.26	88.50	88.34
POS+Position	88.24	88.53	88.35
POS+Phrase	88.26	88.49	88.33
POS+DS	88.10	88.41	88.22
Baseline+BroadPOS	87.11	86.96	87.01
BroadPOS+Position	87.04	86.88	86.93
BroadPOS+Phrase	87.06	86.88	86.94
BroadPOS+DS	87.04	86.95	86.98

Table 5.8: Discourse Structure (DS) scores for German

Model	Precision	Recall	F1-Score
Baseline+POS	76.59	76.35	76.43
POS+Position	76.30	76.02	76.11
POS+Phrase	77.01	76.60	76.84
POS+DS	76.44	76.12	76.23
Baseline+BroadPOS	77.24	76.62	76.85
BroadPOS+Position	76.68	75.87	76.20
BroadPOS+Phrase	77.28	76.63	76.87
BroadPOS+DS	75.71	74.92	75.21

Table 5.9: Discourse Structure (DS) scores for English

only nominally improved results for Baseline+BroadPOS (but was not statistically significant). Overall, phrase structure appeared to add information to the model when part-of-speech itself is error-prone. Conversely, positional features managed to degrade performance of the model across the board. This was surprising given that word position was previously supposed to be an important indicator of accent in English (Gregory & Altun, 2004).

Model	Precision	Recall	F1-Score
Baseline+POS	88.26	88.50	88.34
POS+IMMC	88.49	88.76	88.59
POS+SS	88.57	88.82	88.65
Baseline+BroadPOS	87.11	86.96	87.01
BroadPOS+IMMC	87.12	86.90	86.98
BroadPOS+SS	87.14	86.90	87.00

Table 5.10: Semantic Space (SS) scores for German

Model	Precision	Recall	F1-Score
Baseline+POS	76.59	76.35	76.43
POS+IMMC	77.00	76.63	76.76
POS+SS	76.84	76.58	76.67
Baseline+BroadPOS	77.24	76.62	76.85
BroadPOS+SS	77.34	76.77	77.00
POS+DS+SS	76.30	75.67	75.92

Table 5.11: Semantic Space (SS) scores for English

Semantic Space features performed much better overall across datasets. In German, features on the semantic space resulted in an 0.4% improvement over the baseline with POS (statistically significant at 0.9), but made no difference when combined with BroadPOS). In English, however, there was improvement in both models, with an 0.3% increase in the overall score as compared to the baseline with POS, and an 0.2% increase as compared to the baseline with BroadPOS (both models statistically significant at 0.9).

In a separate experiment, we wanted to compare the performance of shallow, textually-extracted features of discourse and semantics, with features based on gold-standard

manual annotations of discourse object and information status in German.

<b>Model</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>
Baseline+POS	88.26	88.50	88.34
POS+IS	88.60	89.19	88.86
POS+IS-ALL	86.77	86.63	86.66

Table 5.12: Information Status (IS) scores for German

The overall score for POS+IS, which included discourse object and information status features only, was 88.86%, resulting in a clear improvement of 0.6% over the baseline with POS, compared with 88.65% for the semantic space features in POS+SS. Interestingly, turning on all manually-annotated features of discourse, which included member status, semantic sort information, and others (cf. section 4.1.1 for a list of manually-annotated discourse features), the performance of the model degraded substantially (86.66% compared to the baseline of 88.34%).

<b>Model</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>
POS	88.26	88.50	88.34
POS+DS	88.10	88.41	88.22
POS+DS+SS	88.34	88.66	88.46
POS+DS+SS+IS	88.34	88.71	88.49

Table 5.13: Combined scores for German

We next looked at how each set of features affected the overall performance of the model when combined with each other. As before, SS and IS managed to improve performance of the model on their own, as well as combined, while DS degraded performance when compared to scores of those features alone.

In order to methodically weed out the features that negatively affected performance, we implemented a genetic algorithm that would randomly select combinations of features that would maximize the overall performance of the model. For this, the entire pool of discourse, semantic, and syntactic features were made available to the system for finding the optimal set.

In the first experiment of this kind, we began with a simple hill-climbing algorithm that selectively compares models of atomic features. In a first step, the algorithm trains

a model in which all possible features are turned on, and then compares its performance with one of 64 models in which only a single feature is turned on. From this step, the genetic algorithm is then run combining the best 16 solutions with each other, such that the best solutions are more likely to be “mated”, with each parent model having a 50% chance of passing on its feature. If a solution has already been tested, random mutations are added, with up to 200 new combinations tested for each round.

Best Single Features
<i>unigram</i>
<i>adpositional-phrase</i>
<i>coordinated-phrase</i>
<i>noun-modifier</i>
<i>immediate-concept</i>
<i>extended-concept</i>
<i>named-entity</i>
<i>pos</i>

Table 5.14: Best selected single feature set for German

Table 5.14 gives the set of features selected after running for 16 hours. This model outperformed the best models from our previous experiments at 88.95%, with statistical significance at 0.9. Interesting, all of the semantic space features were selected, in addition to several phrase structure variables, while manually-annotated discourse features were not.

Model	Precision	Recall	F1-Score
Best Single Features	88.82	89.17	88.95
Best Combined Features	88.99	88.50	88.34

Table 5.15: Scores for best selected single feature set for German

In a second experiment, we wanted to see whether certain combinations of atomic features might add more to the performance of the model than they could on their own. Since each feature has its own weight in the model, combining an optimal set of atomic features into a single, higher-weighted feature might have more impact on the overall model. These results, however, were not significant. This was most likely due to insufficient time in testing feature combinations.

### 5.3 Discussion

It is clear from our results that part-of-speech on its own is a strong indicator of accent, to the extent that sentence-level discourse structure is simply redundant. Part-of-speech would, after all, capture the presence of a compound noun by indicating it as a string of nouns, without explicit mention of the phrase boundary. Other forms of discourse structure such as cue phrases and referring expressions, while not explored here, might presumably be likewise superfluous given the information available from part-of-speech alone<sup>1</sup>. On the other hand, our results show that such discourse structure can actually boost the overall performance of the model when part-of-speech cannot be reliably obtained. In this case, redundancy of information appears to support the emergence of “correct” information. This is an encouraging outcome for practical applications of accent prediction, which must often rely on many levels of error-prone linguistic analysis.

Nevertheless, part-of-speech would not explain the low performance of passage-level discourse features like sentence position. Rather, the information on where an utterance occurs in the discourse is probably more efficiently represented by the semantic space model. As the semantic space grows dynamically with the unfolding discourse, it will by definition give a higher weighting of givenness to elements towards the end of a discourse than to those at the beginning. Additional indicators of sentence position are then simply redundant.

The semantic space model itself resulted in improving both precision and recall when combined with part-of-speech. All of our semantic space features consistently appeared in the top models in our feature selection investigations, outperforming even gold-standard linguistic annotations of information status and discourse referents. Although the semantic space does not explicitly model semantic relations between elements, which may be a factor in deaccentuation, it nevertheless appears to offer a kind of measure for information content that, importantly, is not founded on a mere distribution of terms in a given corpus. In particular, our findings show that the semantic space model is robust across speaking style (from spontaneous conversation to read speech) as well as across genre (from news articles to intimate dialogs). This makes it an ideal solution for accent prediction in domain-independent uses of TTS.

---

<sup>1</sup>This would not necessarily be the case for discourse relations.

## Chapter 6

# Conclusions

We presented a model for statistical pitch accent prediction that incorporates elements of discourse structure and a semantic space of the spoken conversational context. In particular, we implemented a sequence-learning model based on Conditional Random Fields for the purpose of predicting context-sensitive accent labels given textual input.

We then introduced a model for approximating the semantic space of a discourse by means of a dynamically-updating semantic context. This semantic space was implemented as the set of lexical items that are either explicitly or semantically evoked at a given time-step of discourse. We used Latent Semantic Analysis, which was trained on a subset of Wikipedia, as a means of determining the set of semantically evoked lexical items.

We then designed a set of features based on the semantic space, which were used to decide whether the currently uttered word is “given” by the current discourse context. This approach was unique in that it allowed us to incorporate an element of “world knowledge” into our model that was simultaneously robust against the genre of a specific training corpus.

We also included a set of features designed to capture aspects of the discourse structure that might influence accentuation patterns; these included identifying instances of noun compounds which might induce dropped accents, as well as tracking the position of an utterance in a discourse which might influence de-accentuation.

Among others, we found that while less informative features can degrade performance of the statistical model when paired with strongly informative features, the combination of these with less reliable (i.e. error-prone) features could work well in unison to boost overall performance.



---

We also found that modeling information content as a function of discourse can improve prediction results. This approach moreover proved robust not only across different speaking styles (from speaker-dependent read speech to speaker-independent spontaneous speech), but also across genre (from news articles to informal conversations).

Finally, we showed how information outside the scope of the current input of our sequence learning model can be used to influence decisions. Specifically, we were able to take information from previous discourse, via our semantic space model, and allow it to be weighed against other factors derived from the current input to predict its corresponding accent sequence.

In the future, it would be interesting to test the theory that de-accentuation due to givenness is correlated with specific lexical relations (e.g. hypernymy-hyponymy). One thing we could do is to expand the set of semantic space features with variables indicating lexical relations between the current utterance and elements in the semantic space. This could be achieved using resources such as WordNet (or German variant, GermaNet) to determine such relations.

In addition, given our model for incorporating elements of discourse external to an input sequence, another relevant area of research would be to continue our investigations of discourse context for dialog systems. In particular, just as the semantic space updates in the context of all speakers' discourse, we might want to investigate correlations between discourse *relations* between the utterance of a speaker and the output from TTS, and its influence on accentuation. Other aspects of dialog situations, such as turn-taking practices and grounding conventions might also be interesting to explore.

To conclude, the use of semantic space features as proposed here provides a promising avenue for improving accent prediction, as it does not incur a significant increase in the complexity of the model, and is robust across the domain of speech.

# References

- ALLAN, J. (2002). *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publishers. 55
- ALLAN, J., CARBONELL, J., DODDINGTON, G., YAMRON, J. & YANG, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 194–218. 52, 55
- ALTER, K., MATIASEK, J. & NIKLFELD, G. (1996). Modeling prosody in a german concept-to-speech system. In *Natural Language Processing and Speech Technology, Results of the 3rd KONVENS Conference*, 156–165, Mouton de Gruyter, Hawthorne, NY, USA. 37
- ANDERSON, A., BADER, M., BARD, E., BOYLE, E., DOHERTY, G., GARROD, S., ISARD, S., KOWTKO, J., MCALLISTER, J., MILLER, J., SOTILLO, C., THOMPSON, H. & WEINERT, R. (1991). The HCRC map task corpus. In *Language and Speech*, vol. 34, 351–366. 64
- BATLINER, A., BUCKOW, J., HUBER, R., WARNKE, V., NÖTH, E. & NIEMANN, H. (2001). Boiling down prosody for the classification of boundaries and accents in german and english. In *Proceedings of Eurospeech*, 2781–2784, Aalborg. 13
- BAUMANN, S. (2006). *The Intonation of Givenness: Evidence from German*. Linguistische Arbeiten 508, Niemeyer, Tübingen. 32
- BAUMANN, S., GRICE, M. & BENZMÜLLER, R. (2001). GToBI - a phonological system for the transcription of german intonation. In Puppel, Stanislaw & Demenko, eds., *Prosody 2000. Speech Recognition and Synthesis*, 21–28, Adam Michiewicz University, Faculty of Modern Languages and Literature, Poznan. 21, 59

- BAUMANN, S., BRINCKMANN, C., HANSEN-SCHIRRA, S., KRUIJFF, G.J.M., KRUIJFF-KORBAYOVÁ, I., NEUMANN, S., STEINER, E., TEICH, E. & USZKOREIT, H. (2004a). The MULI project: Annotation and analysis of information structure in german and english. In *Proceedings of the LREC 2004 Conference*, 26–28, Lisbon, Portugal. 57
- BAUMANN, S., BRINCKMANN, C., HANSEN-SCHIRRA, S., KRUIJFF, G.J.M., KRUIJFF-KORBAYOVÁ, I., NEUMANN, S. & TEICH, E. (2004b). Multi-dimensional annotation of linguistics corpora for investigating information structure. In A. Meyers, ed., *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, 39–46, Association for Computational Linguistics, Boston, Massachusetts, USA. 58, 59
- BECKER, S., SCHRÖDER, M. & BARRY, W.J. (2006). Rule-based prosody prediction for german text-to-speech synthesis. In *Speech Prosody 2006*, Dresden, Germany. 31
- BECKMAN, M.E. (1986). *Stress and Non-Stress Accent*. Mouton de Gruyter. 13
- BECKMAN, M.E. & AYERS, G. (1994). *Guidelines for ToBI Labelling, vers. 2.0*. Ohio State University. 20
- BECKMAN, M.E. & HIRSCHBERG, J. (1994). *The ToBI Annotation Conventions*. Ohio State University, ms. and accompanying speech materials. 20
- BECKMAN, M.E. & PIERREHUMBERT, J. (1986). Intonational structure in japanese and english. *Phonology Yearbook*, **3**, 255–309. 19
- BOLINGER, D. (1951). Intonation: Levels versus configurations. *Word*, **14**, 109–149. 12, 14, 15
- BOLINGER, D. (1958). A theory of pitch accent in english. *Word*, **14**, 109–49. 12
- BOLINGER, D. (1972). Accent is predictable (if you're a mind reader). *Language*, **48**, 633–644. 32
- BRANTS, S., DIPPER, S., HANSEN, S., LEZIUS, W. & SMITH, G. (2002). The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol. 58, 61

- 
- BRENIER, J., NENKOVA, A., KOTHARI, A., WHITTON, L., BEAVER, D. & JURAFSKY, D. (2006). The (non)utility of linguistic features for predicting prominence in spontaneous speech. In *IEEE/ACL 2006 Workshop on Spoken Language Technology*. 7
- BROWN, G. (1983). Prosodic structure and the given/new distinction. In D.R. Ladd & A. Cutler, eds., *Prosody: Models and Measurements*, 67–78, Springer Verlag, Berlin. 31
- BROWN, P., PIETRA, V.D. & MERCER, R. (1991). Word sense disambiguation using statistical methods. In *Proceedings of the 29th Meeting of the ACL (ACL-91)*, 264–270, Berkeley, CA. 46
- BUCKERIDGE, A.M. & SUTCLIFFE, R.F.E. (2002). Disambiguating noun compounds with latent semantic indexing. In *International Conference on Computational Linguistics*, 1–7. 47
- BÜRING, D. (1996). On (De)Accenting. Talk presented at the SFB 340 conference in Tübingen. 27
- BÜRING, D. (2006). Focus projection and default prominence. In V. Molnár & S. Winkler, eds., *The Architecture of Focus*, Mouton de Gruyter, Berlin/New York. 27, 28
- BYRD, R.H., NOCEDAL, J. & SCHNABEL, R.B. (1994). Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming*, **63**, 129–156. 43, 72
- CHAFE, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. In Li, ed., *Subject and Topic*, 25–55, Academic Press, New York. 24
- CHAFE, W. (1994). *Discourse, Consciousness, and Time*. University of Chicago Press. 26
- CHEN, K., HASEGAWA-JOHNSON, M. & COHEN, A. (2004). An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model. In *Proceedings of ICASSP*, 509–512. 38

- 
- CHISHOLM, E. & KOLDA, T.G. (1999). New term weighting formulas for the vector space method in information retrieval. Tech. rep., Oak Ridge National Laboratory. 51, 52
- CHOMSKY, N. (1957). *Syntactic Structures*. Mouton. 15
- CHOMSKY, N. & HALLE, M. (1968). *The Sound Pattern of English*. Harper and Row, New York. 15, 26
- CRUTTENDEN, A. (1997). *Intonation*. Cambridge University Press, New York, 2nd edn. 26
- DAHL, Ö. (1976). What is new information? In N.E. Enkvist & V. Kohonen, eds., *Reports on Text Linguistics: Approaches to Word Order*, vol. 8, 37–50, Meddelanden från Stiftelsens för Åbo Akademi Forskningsinstitut, Åbo/Turku. 23
- DENOYER, L. & GALLINARI, P. (2006). The Wikipedia XML Corpus. *SIGIR Forum*. 68
- DUMAIS, S.T., FURNAS, G.W., LANDAUER, T.K. & DEERWESTER, S. (1988). Using latent semantic analysis to improve information retrieval. In *Conference on Human Factors in Computing (CHI'88)*, 281–285, ACM, New York. 47
- FELLBAUM, C., ed. (1998). *WordNet: An Electronic Lexical Database*. MIT Press. 46
- FOLTZ, P.W., LAHAM, D. & LANDAUER, T.K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning. Online Journal*, 1. 47
- FRY, D.B. (1958). Experiments in the perception of stress. *Language and Speech*, 1, 126–152. 12
- GOLDSMITH, J. (1976). An overview of autosegmental phonology. *Linguistic Inquiry*, 2, 23–68. 16
- GOLDSMITH, J. (1996). *The Handbook of Phonological Theory*. Blackwell Publishing. 18

- GRABE, E. (2004). Intonational variation in urban dialects of english spoken in the british isles. *Regional Variation in Intonation*, 9–31. 64
- GRABE, E. & POST, B. (2002). *The transcribed IViE corpus*. University of Oxford Phonetics Laboratory. 64
- GRABE, E., NOLAN, F. & FARRAR, K.J. (1998). Ivie - a comparative transcription system for intonational variation in english. In *Fifth International Conference on Spoken Language Processing*, 0099, ISCA. 64
- GRABE, E., POST, B. & NOLAN, F. (2001a). *The IViE Corpus*. Department of Linguistics, Univerity of Cambridge. 21, 64
- GRABE, E., POST, B. & NOLAN, F. (2001b). Modelling intonational variation in english. the ivie system. In S. Puppel & G. Demenko, eds., *Proceedings of Prosody 2000*, Adam Michiewicz University, Poznan, Poland. 64
- GREGORY, M.L. & ALTUN, Y. (2004). Using conditional random fields to predict pitch accents in conversational speech. In *Proceedings of the 26th Annual Meeting of the ACL (ACL'04)*, 677, Morristown, NJ, USA. 4, 5, 35, 36, 39, 41, 71, 76
- GRICE, M., REYELT, M., BENZMÜLLER, R., MAYER, J. & BATLINER, A. (1996). Consistency in transcription and labelling of german intonation with GToBI. In *Proceedings of the 4th International Conference of Spoken Language Processing*, 1716–1719, Philadelphia. 21
- GROSSMAN, D.A. & FRIEDER, O. (2004). *Information Retrieval: Algorithms and Heuristics*. Springer, 2nd edn. 48
- GROSZ, B.J. & SIDNER, C.L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, **12**, 175–204. 5, 33, 45, 53
- GROSZ, B.J., WEINSTEIN, S. & JOSHI, A. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, **21**, 203–225. 34
- GUNDEL, J.K. (1985). Shared knowledge and topicality. *Journal of Pragmatics*, **9**, 83–107. 29

- 
- GUSSENHOVEN, C. (1983). Focus, mode, and the nucleus. *Journal of Linguistics*, **19**, 377–417. 26
- GUSSENHOVEN, C. (1992). Sentence accents and argument structure. In I. Roca, ed., *Thematic Structure. Its Role in Grammar*, 79–106, Foris, Berlin. 26
- HALLIDAY, M. (1967). Notes on transitivity and theme in english, part II. *Journal of Linguistics*, **3**, 199–244. 24, 25, 26, 28
- HAMMERSLEY, J. & CLIFFORD, P. (1971). Markov fields on finite graphs and lattices, unpublished manuscript. 42
- HAMP, B. & FELDWEG, H. (1997). GermaNet - a lexical-semantic net for german. In *Proceedings of the ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid. 46
- HAYES, B. (1984). The phonology of rhythm in english. *Linguistic Inquiry*, **15**, 33–74. 14
- HIRSCHBERG, J. (1993). Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, **63**, 305–340. 5, 35, 38
- HOEFEL, G. & ELKAN, C. (2008). Learning a two-stage svm/crf sequence classifier. In *CIKM'08: Proceeding of the 17th ACM conference on information and knowledge management*, 271–278, ACM, New York, NY, USA. 41
- ISAČENKO, A.V. & SCHÄDLICH, H.J. (1966). Untersuchungen über die deutsche satzintonation. *Studia Grammatica*, **VII**, 7–64. 12
- JACKENDOFF, R. (1972). *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA, USA. 29
- JORDAN, M.I. (1999). *Learning in graphical models*. MIT Press, Cambridge, MA, USA. 41
- KOCHANSKI, G., GRABE, E., COLEMAN, J. & ROSNER, B. (2005). Loudness predicts prominence; fundamental frequency lends little. *J. Acoustical Society of America*, **11**, 1038–1054. 13

- KUNO, S. (1972). Functional sentence perspective. *Linguistic Inquiry*, **3**, 269–320. 31
- LADD, D.R. (1980). *The Structure of Intonational Meaning: Evidence from English*. Indiana University Press, Bloomington. 25, 28, 29
- LADD, D.R. (1984). English compound stress. In D. Gibbon & H. Richter, eds., *Intonation, Accent and Rhythm*, 253–266, Walter de Gruyter, Berlin. 22
- LADD, D.R. (1996). *Intonational Phonology*. Cambridge University Press. 12, 16, 19, 25, 28
- LAFFERTY, J., MCCALLUM, A. & PEREIRA, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th Intl. Conference on Machine Learning (ICML-2001)*. 41, 42, 43
- LAMBRECHT, K. (1994). *Information Structure and Sentence Form: Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge University Press. 23
- LANDAUER, T.K. & DUMAIS, S.T. (2008). Latent semantic analysis. *Scholarpedia*, **3**, 4356. 47, 48
- LANDAUER, T.K., FOLTZ, P.W. & LAHAM, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, **25**, 259–284. 47
- LEVOW, G.A. (2005). Context in multi-lingual tone and pitch accent prediction. In *Proceedings of Interspeech 2005*, 1809–1812. 38
- LEVOW, G.A. (2008). Automatic prosodic labeling with conditional random fields and rich acoustic features. In *The Third Intl. Joint Conference on Natural Language Processing (IJCNLP-08)*, 217–224. 5, 36, 39, 41
- LIBERMAN, M. (1975). *The Intonational System of English*. Garland, New York. 17
- LIBERMAN, M. & PRINCE, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, **8**, 249–336. 17
- LIBERMAN, M. & SPROAT, R. (1992). The stress and structure of modified noun phrases in english. In *Lexical Matters*, 131–181, Cambridge University Press. 21, 38



- MARCUS, M., KIM, G., MARCINKIEWICZ, M.A., MACINTYRE, R., BIES, A., FERGUSON, M., KATZ, K. & SCHASBERGER, B. (1994). The penn treebank: Annotating predicate argument structure. In *In ARPA Human Language Technology Workshop*, Morgan Kaufmann. 58
- MCCALLUM, A. (2003). Efficiently inducing features of conditional random fields. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*. 41, 42
- NENKOVA, A., BRENIER, J., KOTHARI, A., CALHOUN, S., WHITTON, L., BEAVER, D. & JURAFSKY, D. (2007). To memorize or to predict: Prominence labeling in conversational speech. *NAACL-HLT*. 6
- NOOTEBOOM, S. (1997). The prosody of speech: Melody and rhythm. In W.J. Hardcastle & J. Laver, eds., *The Handbook of Phonetic Sciences*, 640–673, Basil Blackwell Limited, Oxford. 11
- OKAZAKI, N. (2007). Crfsuite: a fast implementation of conditional random fields (CRFs). <http://www.chokkan.org/software/crfsuite/>. 72
- PAN, S. & HIRSCHBERG, J. (2000). Modeling local context for pitch accent prediction. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Hong Kong. 35
- PAN, S. & MCKEOWN, K.R. (1998). Learning intonation rules for concept to speech generation. In *Proceedings of COLING/ACL'98*, 1003–1009. 38
- PAN, S. & MCKEOWN, K.R. (1999). Word informativeness and automatic pitch accent modeling. In *Proceedings of EMNLP/VLC'99*, 148–157. 35, 37, 38
- PIERREHUMBERT, J. (1980). *The Phonetics and Phonology of English Intonation*. Ph.D. thesis, MIT, Bloomington: Indiana University Linguistics Club. 19
- PITRELLI, J., BECKMAN, M.E. & HIRSCHBERG, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the International Conference on Spoken Language Processing*, vol. 1, 123–126, Yokohama, Japan. 21

- PREVOST, S. (1995). *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*. Ph.D. thesis, University of Pennsylvania. 28
- PREVOST, S. & STEEDMAN, M. (1993). Using context to specify intonation in speech synthesis. In *Proceedings of the 3rd European Conference of Speech Communication and Technology (EUROSPEECH)*, 2103–2106. 37
- PREVOST, S. & STEEDMAN, M. (1994). Information based intonation synthesis. In *Proceedings of the ARPA Workshop on Human Language Technology*, 193–198. 37
- PRINCE, E.F. (1981). Toward a taxonomy of given-new information. In P. Cole, ed., *Radical Pragmatics*, 223–256, Academic Press, New York. 22, 24, 29, 31, 61
- PRINCE, E.F. (1992). The ZPG letter: Subjects, definiteness, and information-status. In S. Thomas & W. Mann, eds., *Discourse Description: Diverse Analyses of a Fund Raising Text*, 295–325, John Benjamins. 6
- REN, Y., KIM, S.S., HASEGAWA-JOHNSON, M. & COLE, J. (2004). Speaker-independent automatic detection of pitch accent. In *ICSA International Conference on Speech Prosody*, 521–524. 36
- SANTORINI, B. (1990). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. University of Pennsylvania, 3rd edn. 67
- SCHILLER, A., TEUFEL, S. & STÖCKERT, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Universität Stuttgart. 61
- SCHMERLING, S. (1976). *Aspects of English Sentence Stress*. University of Texas Press, Austin. 26
- SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, vol. 12, Manchester, UK. 67
- SCHWARZSCHILD, R. (1999). Givenness, AvoidF and other constraints on the placement of accent. In *Natural Language Semantics*, vol. 7, 141–177, Springer Netherlands. 28

- SELKIRK, E. (1984). *Phonology and Syntax. The Relation between Sound and Structure*. MIT Press, Cambridge, MA, USA. 17, 19, 26
- SELKIRK, E. (1995). Sentence prosody: Intonation, stress and phrasing. In J. Goldsmith, ed., *The Handbook of Phonological Theory*, 550–569, Oxford, Cambridge, MA, USA. 26
- SGALL, P., HAJIČOVA, E. & BENEŠOVA, E. (1973). *Topic, Focus, and Generative Semantics*. Scriptor, Krongberg/Taunus. 29
- SHA, F. & PEREIRA, F. (2003). Shallow parsing with conditional random fields. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 134–141, Association for Computational Linguistics, Morristown, NJ, USA. 41, 43, 44
- SHATTUCK-HUFNAGEL, S., OSTENDORF, M. & ROSS, K. (1994). Stress shift and early pitch accent placement in lexical items in american english. *Journal of Phonetics*, **22**, 357–388. 14
- SILVERMAN, K., BECKMAN, M.E., PITRELLI, J., OSTENDORF, M., WIGHTMAN, C., PRICE, P., PIERREHUMBERT, J. & HIRSCHBERG, J. (1992). ToBI: a standard for labeling english prosody. In *Proceedings of the 2nd International Conference of Spoken Language Processing*, 867–870, Banff, Canada. 20
- SKUT, W. & BRANTS, T. (1998). Chunk tagger: Statistical recognition of noun phrases. In *ESSLLI-1998 Workshop on Automated Acquisition of Syntax and Parsing*, Saarbrücken. 67
- SRIDHAR, V.K.R., BANGALORE, S. & NARAYANAN, S. (2008a). Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. In *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, 797–811. 6
- SRIDHAR, V.K.R., NENKOVA, A., NARAYANAN, S. & JURAFSKY, D. (2008b). Detecting prominence in conversational speech: Pitch accent, givenness, and focus. In *Proceedings of Speech Prosody*, 380–388, Campinas, Brazil. 38

- 
- STALNAKER, R. (1978). Assertion. In P. Cole, ed., *Syntax and Semantics: Pragmatics*, vol. 9, 315–332, Academic Press, New York. 23
- SUN, X. (2002). Pitch accent prediction using ensemble machine learning. In *Proceedings of ICSLP-2002*, 16–20. 38
- SUTTON, C. & MCCALLUM, A. (2007). An introduction to conditional random fields for relational learning. In L. Getoor & B. Taskar, eds., *Introduction to Statistical Relational Learning*, MIT Press. 42, 43
- 'T HART, J., COLLIER, R. & COHEN, A. (1990). *A Perceptual Study of Intonation: An Experimental-phonetic Approach to Speech Melody*. Cambridge University Press. 10
- TAYLOR, P.A. (1992). *A Phonetic Model of English Intonation*. Ph.D. thesis, University of Edinburgh. 15
- TERKEN, J. & HIRSCHBERG, J. (1994). Deaccentuation of words representing 'given' information: Effects of persistence of grammatical function and surface position. *Language and Speech*, **37**, 125–145. 31
- UHMANN, S. (1991). *Fokusphonologie. Eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie*. Niemeyer, Tübingen. 18, 26
- WAGNER, P. & PAULSON, M. (2006). Stress patterns of complex german cardinal numbers. In *Speech Prosody*. 14
- WENNERSTROM, A. (2001). *The Music of Everyday Speech: Prosody and Discourse Analysis*. Oxford University Press, New York. 24
- ZELIKOVITZ, S. & HIRSH, H. (2001). Improving text classification with LSI using background knowledge. In *IJCAI01 Workshop Notes on Text Learning: Beyond Supervision*, 113–118. 47