Frequency Effects on Pitch Accents: Towards an Exemplar-Theoretic Approach to Intonation

Von der Fakultät für Informatik, Elektrotechnik und Informationstechnik
der Universität Stuttgart
zur Erlangung der Würde eines Doktors der Philosophie (Dr. phil.)
genehmigte Abhandlung

Vorgelegt von

Katrin Schweitzer

aus Heidenheim an der Brenz

Hauptberichter:	Prof. Dr. Grzegorz Dogil
1. Mitberichter:	Prof. Dr. Bernd Möbius
2. Mitberichter:	Prof. Dr. Bettina Braun

Tag der mündlichen Prüfung: 18. Juni 2012
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
2012

Erklärung

Hiermit erkläre ich, dass ich, unter Verwendung der im Literaturverzeichnis aufgeführten Quellen und unter fachlicher Betreuung, diese Dissertation selbständig verfasst habe.

(Katrin Schweitzer)

Publikationen

Aspekte der hier beschriebenen Forschung finden sich in folgenden begutachteten Publikationen:

Schweitzer et al. (2009a) Schweitzer, Katrin, Riester, Arndt; Walsh, Michael; Dogil, Grzegorz; 2009a. *Pitch Accents and Information Status in a German Radio News Corpus*. In: Proceedings of Interspeech 2009. Brighton, UK, pp. 877–880.

Schweitzer et al. (2009b) Schweitzer, Katrin; Walsh, Michael; Möbius, Bernd; Riester, Arndt; Schweitzer, Antje, Schütze, Hinrich; 2009b. *Frequency matters: Pitch accents and information status*. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). Athens, Greece, pp. 728–736.

Schweitzer et al. (2010a) Schweitzer, Katrin; Calhoun, Sasha; Schütze, Hinrich; Schweitzer, Antje, Walsh, Michael; 2010a. *Relative frequency affects pitch accent realisation: evidence for exemplar storage of prosody.* In: Proceedings of the Thirteenth Australasian International Conference on Speech Science and Technology (SST) 2010. Melbourne, Australia, pp. 62–65.

Schweitzer et al. (2010b) Schweitzer, Katrin; Walsh, Michael; Möbius, Bernd; Schütze, Hinrich; 2010b. *Frequency of occurrence effects on pitch accent realisation.* In: Proceedings of Interspeech 2010. Makuhari, Japan, pp. 138–141.

Schweitzer et al. (2011) Schweitzer, Katrin; Walsh, Michael; Calhoun, Sasha; Schütze, Hinrich; 2011. *Prosodic variability in lexical sequences: intonation entrenches too.* In: Proceedings of ICPhS 2011. Hong Kong, pp. 1778–1781.

Danke!

... meinem Doktorvater **Grzegorz Dogil** dafür, dass er mich immer bestärkt und unterstützt hat und für die unzähligen wertvollen Ratschläge und Tipps! Außerdem danke für die offene, moderne, inspirierende und motivierende Athmosphäre am Lehrstuhl.
Und für die sprechende Kaffeemaschine!

... **Bernd Möbius** für die erfolgreiche Kooperation, die Zeit für lange Diskussionen, für lustige Abende auf Konferenzen und für die vernichtendste Trivial-Pursuit Niederlage, die jemals ein Fluggast auf einem Langstreckenflug einstecken musste – sie schmerzt noch immer, war aber sicherlich gut für meinen Charakter.

... **Bettina Braun** für die Brücke zur "Welt der Abstraktionisten" und fürs unglaublich gründliche Lesen meiner Arbeit und die vielen hilfreichen Kommentare!

... **Mike Walsh** für die Kooperation, die Kabbeleien und die Hirnverknotungen bei den verzwicktesten Exemplar-Theorie-Fragen, fürs Korrektur lesen und für die obligatorische "cupán tae" wann immer etwas auswegslos zu sein schien.

... **Antje Schweitzer** dafür, dass sie mir (von Kindesbeinen an) den Kopf gewaschen oder den Rücken gestärkt hat – je nachdem, was grade nötig war. Außerdem hat sie mit Engelsgeduld und größter Präzision und Ausdauer Korrektur gelesen.

... **Hans Kamp** und **Arndt Riester** für ihr unerschöpfliches Semantik-Wissen, das immense Informationsstruktur-Know-How sowie für ihr großes Interesse an Phonetik.

... **Fabienne Cap** für den fantastischsten Kuchen aller Zeiten und die amüsantesten Mittagsrunden, **Nadja Schauffler** fürs geduldige Probehören und Ego-Aufpäppeln, **Sabine Dieterle** für die besten Türrahmenschwätze und die wiederholte Versicherung dass es "anderen Doktoranden genauso geht", **Sabine Mohr** für den unterhaltsamsten interdisziplinären E-Mail-Austausch und der **ganzen Phonetik-Gruppe** dafür, dass ich jeden Tag gern ins Büro gehe.

... **Markus Prechtel** dafür, dass meine Formeln so richtig "mathematisch" aussehen.

... **Judith Lam** und **Stefanie Anstein** fürs fröhlich im gleichen Boot sitzen und fürs Sich-mit-mir-Freuen. Jetzt seid Ihr dran!

... **Mama und Papa** dafür, dass sie (fast) nie gefragt haben, wann ich fertig werde.

... **Steve Jarand** für die schönste Motivation, fertig zu werden.

Diese Dissertation wurde im Rahmen meiner Projektarbeit im Projekt A1 des SFB 732 durch die Deutsche Forschungsgemeinschaft gefördert. Herzlichen Dank auch dafür.

Contents

List of Abbreviations

CSM	Context Sequence Model
F_0	Fundamental frequency
GToBI	German Tones and Break Indices
GToBI(S)	German Tones and Break Indices (Stuttgart version)
HN	Dataset with nuclear H* accents (Exp. 3)
HNbal	Balanced dataset with nuclear H* accents (Exp. 3)
HNequ	Equalised dataset with nuclear H* accents (Exp. 3)
HNmod	Modified dataset with nuclear H* accents (Exp. 3)
LHN	Dataset with nuclear L+H* accents (Exp. 3)
LHNmod	Modified dataset with nuclear L+H* accents (Exp. 3)
IMS	Institut für Maschinelle Sprachverarbeitung (Institute for Natural Language Processing)
PA	Pitch Accent
PaIntE	Parametric Representation of Intonation Events
ToBI	Tones and Break Indices

Abstract

This thesis presents four corpus experiments which aim to bridge the gap between research on exemplar-theoretic phenomena and research on intonation. The main assumptions of these two areas are contradictory in some respects: while exemplar-theoretic, or episodic, approaches to language acquisition assume that acoustic detail is stored, and consequently fundamental frequency should be part of the stored mental representation of linguistic units, the most widespread intonation models, autosegmental-metrical theories of intonation, are based on the assumption that intonation in Germanic languages is assigned post-lexically.

To investigate if and how intonation can be incorporated into existing exemplar-theoretic models, the described experiments examine tonal parameters of German and English speech data with respect to their being subject to frequency of occurrence effects. The studies target three central questions. Firstly, does frequency of occurrence affect pitch accent realisation? Secondly, how do the word and the tonal level interact in exemplar-theoretic selection? And thirdly, what are the essential features of an exemplar model accounting for intonation?

To answer these questions, all four experiments explore the realisation of pitch accent tokens in context.

The first experiment looks at the variability among the tokens of different pitch accent types and relates differences in variability to the frequency of occurrence of these types. To measure the variability of a pitch accent type, the pitch accent tokens are represented as vectors in a 6-dimensional space. The 6 dimensions correspond to parameters which define an accent's shape. These parameters are approximated with a parametric intonation model (PaIntE; Möhler, 2001) and capture the accent's peak height and alignment, as well as the gradients and ranges of its falling and rising slope, respectively.

The similarity between two accent tokens is calculated as the cosine of the angle between the two vectors. For each pitch accent type, each token is compared to every other token of the same type. The resulting distribution of pairwise similarity values for each pitch accent type reveals how variable the type is. The analyses compare different pitch

accent types as well as different information status categories. The data demonstrates that pitch accent realisation is sensitive to the combination of both the frequency of occurrence of the pitch accent type and of the information status category.

The second experiment takes the word bearing the pitch accent token into account by relating the frequency of occurrence of pitch accent+word pairs to details about the shape of the respective pitch accents. Thereby, potential frequency effects of words occurring with a specific tonal contour, i.e. with a specific pitch accent type, can be detected. Such effects of combinations of words and pitch accents are expected if exemplar-theoretic assumptions hold for intonation and exemplars contain segmental as well as suprasegmental properties. The shape of the analysed pitch accents is described by means of the above mentioned PaIntE parameters. The analyses reveal that the frequency of the combination of word and pitch accent type is significantly correlated with changes in pitch accent shape, specifically with increasing pitch accent ranges for more frequent pairs.

The third experiment further investigates the realisation of accent+word pairs. For a given pair, the experiment looks at the effect of exemplars where the same word was accented with another pitch accent, on the tonal characteristics of the pair under investigation. Since Exemplar Theory assumes that existing exemplars contribute to the construction of a production target, such instances of the same word with another pitch accent can be regarded as competing exemplars to a specific production. To investigate how these competitors influence the production of a given exemplar, the relative frequency of a word with a specific pitch accent type is calculated. This value encodes for each pitch accent+word pair, the number of times the word occurs with the given pitch accent type relative to all its pitch accented occurrences. The study examines variability of pitch accent realisation by measuring the Euclidean distance among the tokens of a pitch accent+word pair in the multidimensional space spanned by the PaIntE parameters described above. The analyses demonstrate that pairs that have a higher relative frequency are less variable in the realisation of the pitch accent. If lexical storage of intonation is assumed, such an effect would be expected: words that are stored with little variability in pitch accent contour (that is, a high relative frequency of pitch accent+word) are more likely to be produced with a similar contour again.

The fourth experiment goes a step further and explores the effect of the relative frequency of a word in its lexical context on its tonal properties. That is, the analyses investigate possible dependencies between the word level alone (without being related to a tonal event) and the tonal level. The tonal level is examined in terms of variability

of pitch accent shape (if the word is accented), similar to the methodology employed in the third experiment: variability of pitch accent tokens on a given word is measured by calculating the Euclidean distance in the multidimensional space spanned by the PaIntE parameters. Additionally, the prosodic pattern of the word and its context in terms of pitch accent and boundary placement is examined. The results demonstrate that the higher the probability of a word in its lexical context, the lower the variability of pitch accent realisations on the word (if accented) and the lower the variability of the prosodic pattern around the word. This indicates that intonation, like segmental features (Pierrehumbert, 2001), can undergo entrenchment as a result of frequency of occurrence effects. Lexical entrenchment of intonation argues against a purely post-lexical assignment of intonation patterns.

Taken together the results from the corpus studies demonstrate that tonal contours are subject to frequency of occurrence effects. Specifically, they are influenced by the frequency of the linguistic context. Moreover, they demonstrate that intonation can undergo entrenchment effects and they indicate that the word as well as the tonal level seem to be crucial for exemplar selection.

Since the results of the corpus studies are well explainable in an exemplar-theoretic framework, an exemplar-theoretic model incorporating intonation seems desirable. At the same time, details of the outcomes have implications for such a model. In essence, the model should assume storage of F_0-contours in contexts, including the possibility to store complex units with their tonal contour as one exemplar. The selection of exemplars to construct a production target should consider the word as well as tonal features (such as being prominent). Furthermore, production biases towards certain tonal characteristics such as greater accent ranges for frequent exemplars should be possible, and the model should average over several exemplars to construct a production target so that these biases can be counterbalanced and entrenchment can be modelled.

To summarise, the work presented here constitutes a step towards an integrated exemplar-based model of speech production in which intonation is accounted for. The corpus studies suggest that Exemplar Theory is highly suited to model intonation production and they highlight the impact that frequency of occurrence has on tonal parameters just as it has been shown to be influential on the segmental level. Consequently, if intonation is to be modelled in other frameworks, frequency effects should be taken into account in order to attain a comprehensive picture of the production of intonation.

Deutsche Zusammenfassung

In der vorliegenden Arbeit werden vier Korpusexperimente beschrieben, die zum Ziel haben, Forschung zum Thema Exemplartheorie und Intonationsforschung aneinander anzunähern. Die grundlegenden Annahmen dieser beiden Bereiche sind zum Teil widersprüchlich: während exemplar-theoretische (auch "episodische") Theorien davon ausgehen, dass akustisches Detail im Sprechergedächtnis gespeichert wird und somit konsequenterweise auch die Grundfrequenz, also der Tonhöhenverlauf, einer sprachlichen Einheit mit abgespeichert sein sollte, gehen die meistverbreiteten Intonationsmodelle, die sogenannten autosegmental-metrischen Theorien, davon aus, dass Intonation in Germanischen Sprachen post-lexikalisch von einer separaten Komponente zugewiesen wird.

Um zu untersuchen, ob und wie Intonation in existierende exemplar-theoretische Modelle integriert werden kann, werden in den vorgestellten Experimenten tonale Parameter von deutschen und englischen Sprachdaten auf potentielle Häufigkeitseffekte getestet. Die Experimente haben vor allem zum Ziel, die folgenden Fragen zu beantworten: Erstens, unterliegt die Realisierung von Pitch-Akzenten Häufigkeitseffekten? Zweitens, wie interagieren die Wortebene und die tonale Ebene im exemplar-theoretischen Selektionsprozess? Und drittens, welche grundlegenden Eigenschaften muss ein exemplar-theoretisches Modell aufweisen, um Intonation modellieren zu können?

Alle vier Experimente untersuchen zur Beantwortung dieser Fragen die Realisierung von Pitch-Akzenten im Kontext.

Das erste Experiment beschäftigt sich mit der Variabilität von Pitch-Akzent-Tokens verschiedener Pitchakzent-Typen und setzt diese Variabilität in Beziehung zu ihrer Häufigkeit. Dafür werden die einzelnen Instanzen eines Pitchakzent-Typs als Vektoren im sechs-dimensionalen Raum dargestellt, wobei die Dimensionen dieses Raumes Parametern entsprechen, welche die Form eines Pitch-Akzentes beschreiben. Diese Parameter werden mit Hilfe eines parametrischen Intonationsmodells (PaIntE; Möhler, 2001) ermittelt. Sie beschrieben die Höhe und das zeitliche Alignment des Akzentgipfels sowie die Gradienten und die Amplituden des Anstiegs bzw. Falls der Akzentkurve. Die Ähnlichkeit zweier Akzente wird dann als der Kosinus des Winkels zwischen den zwei

jeweiligen Vektoren berechnet. Jedes Token eines Pitchakzent-Typs wird mit jedem anderen Token des selben Typs verglichen. So ergibt sich eine Verteilung von paarweisen Ähnlichkeitswerten, welche die Variabilität eines Typs widerspiegelt. Die Einzelanalysen des Experiments vergleichen die Verteilungen verschiedener Pitchakzent-Typen sowie verschiedener Informationsstatus-Kategorien. Die Daten veranschaulichen, dass die Realisierung von Pitch-Akzenten sowohl von der Häufigkeit des Pitchakzent-Typs als auch von der der Informationsstatus-Kategorie abhängig ist.

Im zweiten Experiment wird auch die Wortebene mit betrachtet, indem Akzente gemeinsam mit den Wörtern, auf welchen sie realisiert wurden, als Akzent-Wort-Paare analysiert werden. Die Vorkommenshäufigkeit eines solchen Paares wird mit der Form der jeweiligen Pitch-Akzente in Beziehung gesetzt, um potentielle Häufigkeitseffekte von Wörtern mit einer spezifischen tonalen Kontur festzustellen. Wenn die exemplar-theoretischen Grundannahmen sich auf Intonations-Parameter erstrecken und Exemplare sowohl segmentale als auch suprasegmentale Information enthalten, werden solche Häufigkeitseffekte erwartet. Wie im ersten Experiment wird die Form der Pitch-Akzente durch PaIntE-Parameter beschrieben. Die Analysen zeigen, dass die Häufigkeit von Akzent-Wort-Paaren signifikant mit Veränderungen in der Pitchakzent-Form korreliert. Im Speziellen ist hierbei eine Steigerung der Amplitude der Akzentkurve hervorzuheben.

Das dritte Experiment untersucht ebenfalls die Realisierung von Akzent-Wort-Paaren. Für ein gegebenes Paar wird untersucht, wie Exemplare des selben Wortes mit einem anderen Akzent die tonalen Eigenschaften des untersuchten Paares beeinflussen. Da in der exemplar-theoretischen Sprachproduktion davon ausgegangen wird, dass bestehende Exemplare der Bildung eines neuen Produktions-Targets zugrunde liegen, können aus exemplar-theoretischer Sicht für jede Produktion eines Wortes mit einem bestimmten Pitchakzent die Exemplare des selben Wortes mit einem anderen Pitchakzent-Typ als konkurrierende Exemplare zu der jeweiligen Produktion angesehen werden. Um zu untersuchen, wie diese “Mitbewerber” die Produktion eines gegebenen Paares beeinflussen, wird die relative Häufigkeit eines Wortes und eines spezifischen Pitchakzents berechnet. Dieser Wert setzt für jedes Akzent-Wort-Paar die Häufigkeit, mit der das Wort mit dem jeweiligen Akzenttyp realisiert wurde, mit der Häufigkeit, mit der es mit einem beliebigen Akzenttyp realisiert wurde, in Relation. Die Studie untersucht dann die Variabilität von Pitchakzent-Realisierungen, indem die Euklidische Distanz zwischen Vorkommen eines Akzent-Wort-Paares im mehrdimensionalen, von PaIntE-Parametern aufgespannten, Raum gemessen wird. Die Analysen zeigen, dass Paare, die eine höhere relative Häufigkeit aufweisen, weniger variabel in der Realisierung der Akzentkurven sind. Unter

der Annahme von lexikalischer Speicherung von Intonation wird solch ein Effekt erwartet: Wörter, die mit wenig variablen Pitch-Konturen gespeichert sind (also Wörter, die eine hohe relative Häufigkeit mit einem Akzenttyp aufweisen), haben eine höhere Wahrscheinlichkeit, wieder mit einer ähnlichen Kontur produziert zu werden.

Das vierte Experiment geht einen Schritt weiter, indem es die Auswirkungen der relativen Häufigkeit eines Wortes in seinem lexikalischen Kontext auf dessen tonalen Eigenschaften untersucht. Das heißt, das Experiment untersucht potentielle Abhängigkeiten zwischen der Wortebene (unabhängig von einem tonalen Ereignis) und der tonale Ebene. Die tonale Ebene wird auf zwei Arten untersucht: zum Einen wird die Variabilität von Pitchakzent-Konturen untersucht, wenn das Wort Satzakzent trägt. Die Methodologie hierfür ist ähnlich der im dritten Experiment verfolgten: Variabilität wird als Euklidische Distanz im mehrdimensionalen PaIntE-Parameter-Raum gemessen. Zusätzlich wird das prosodische Muster des Wortes und seines Kontexts ermittelt und untersucht. Die Ergebnisse zeigen, dass eine höhere Wahrscheinlichkeit des Wortes in seinem lexikalischen Kontext mit geringerer Variabilität sowohl in der Pitchakztent-Realisierung als auch im prosodischen Muster einhergeht. Dies deutet darauf hin, dass Intonation einem frequenzgesteuerten Prozess des "Entrenchments" unterliegen kann, wie das für segmentale Parameter von gesprochener Sprache der Fall ist (Pierrehumbert, 2001). Lexikalisches Entrenchment von Intonation spricht gegen eine rein postlexikalische Zuweisung von tonaler Struktur.

Die Korpusexperimente zeigen insgesamt, dass tonale Konturen Häufigkeitseffekten unterliegen. Dies betrifft insbesondere die Häufigkeit des linguistischen Kontexts. Darüberhinaus belegen sie, dass Intonation Entrenchment-Effekten unterliegen kann und verdeutlichen, dass sowohl die Wort- als auch die tonale Ebene eine Rolle im exemplartheoretischen Selektionsprozess spielen.

Da alle Ergebnisse der Korpusstudien im Rahmen exemplar-theoretischer Modelle erklärbar sind, erscheint die Ausarbeitung eines Exemplar-Modelles, welches die tonale Ebene berücksichtigt, wünschenswert. Gleichzeitig haben die Ergebnisse Implikationen für ein solches Modell. Im Wesentlichen sollte das Modell die Speicherung tonaler Konturen im Kontext ermöglichen, wobei die Möglichkeit bestehen sollte, komplexe Einheiten mitsamt ihrem tonalen Verlauf als eine Instanz zu speichern. Der Selektionsprozess für die Ermittlung eines Produktionstargets sollte sowohl das Wort als auch seinen tonalen Status (zum Beispiel seine Prominenz) berücksichtigen. In der Produktion sollte eine Tendenz hin zu bestimmten tonalen Realisierungen, wie zum Beispiel erhöhten Akzentamplituden für häufige Exemplare, modellierbar sein, und das Modell sollte bei der

Ermittlung des Produktionstargets über mehrere Exemplare mitteln, so dass Produktionstendenzen ausgeglichen werden und Entrenchment modelliert werden kann.

Alles in allem stellt die hier vorgestellte Arbeit einen Schritt hin zu einem integrierten exemplar-basierten Model der Sprachproduktion dar, in dem Intonation berücksichtigt wird. Die Korpusstudien deuten darauf hin, dass Exemplar-Theorie sehr gut geeignet ist, um Intonationsproduktion zu modellieren, und sie verdeutlichen, dass die Häufigkeit einer Einheit auch einen Effekt auf tonale Parameter hat, analog zu Effekten auf Segmentebene. Deswegen sollten auch nicht-exemplar-basierte Modelle der Sprachproduktion Häufigkeitseffekte berücksichtigen, um ein umfassendes Bild der Intonationsproduktion zu erreichen.

1 Introduction

The frequency with which a linguistic unit occurs can influence how it is realised, categorised, perceived or how it evolves over time, both in the speaker and in the population. These frequency of occurrence effects in language are well documented and occur on a variety of linguistic levels and domains. In the domain of spoken language, most of the documented effects describe effects on phones, syllables, words or phrases with respect to to their durational or articulatory properties (e.g. Cholin et al., 2006; Carreiras and Perea, 2004; Losiewicz, 1992; Bybee and Scheibman, 1999; Jurafsky et al., 2001; Lee et al., 1999). In the domain of prosody there has been only very little research on frequency of occurrence effects (but see Vigário et al., 2006, for effects of frequency on the acquisition of the prosodic word and A. Schweitzer and Möbius, 2004, as well as Walsh et al., 2007, for effects of syllable frequency on durational aspects). However, with the exception of Braun et al. (2006), who demonstrate that random tonal contours gravitate towards frequent tonal contours in an iterative mimicry study, little research on possible effects of frequency of occurrence on tonal parameters has taken place.

Frequency of occurrence effects are generally well explainable in the framework of Exemplar Theory (e.g. Nosofsky, 1986; Johnson, 2006; Goldinger, 1997; Pierrehumbert, 2001; Bybee, 2006). In fact, in this usage-based approach to language production and perception, frequency of occurrence is one of the core aspects: The number of times a speaker of a language is exposed to a linguistic unit is crucial to production and perception and assumed to influence both.

The basic assumption of exemplar-theoretic models is that units are stored in memory as *exemplars*. They are assumed to be highly detailed, that is, they contain fine acoustic detail (see e.g. Goldinger, 1998, 1997; Wade et al., 2010), as well as speaker specific information (Goldinger, 1997), detail about the speaking situation, and any other linguistic or extralinguistic context (Pierrehumbert, 2001). Similar exemplars are, according to Exemplar Theory, grouped together and form an *exemplar cloud* - this accumulation of single instances is the representation of what would be called a *category* in traditional, often generative approaches to language production and perception (Chomsky,

1957). Consequently, in an exemplar-theoretic framework, a category always encodes its frequency. Knowledge about structure or semantic interpretation – the *"grammar"* – is assumed to evolve in the lexicon: frequent exemplars can spread their characteristics by analogy so that unknown exemplars can be decomposed and interpreted (e.g. Bod, 1998, 2009). Thus, grammar is considered to be usage-based and frequency-driven. Highly frequent exemplars are assumed to be produced with less production effort and less variation, a phenomenon called *entrenchment* (Pierrehumbert, 2001).

To follow the basic ideas of Exemplar Theory, one has to assume that intonation, i.e. fundamental frequency, is part of the exemplar and is therefore stored in memory. This might sound unsurprising at first, but in fact intonation in Germanic languages is assumed to be post-lexical according to models of intonation, in particular autosegmental-metrical models of intonation (e.g. Pierrehumbert, 1980). These approaches to prosody production are well-accepted in the research community and have been adapted for many languages. They consider intonation to be assigned *after* retrieving the lexical items. Thus, intonation is linguistically meaningful but not part of the mental lexicon (which only comprises simplex items; longer, complex sequences are not stored but composed or decomposed separately[1]). The assignment of pitch accents and phrase boundaries is assumed to be driven by top-down information (e.g. syntactic or semantic information, see Levelt, 1989; Ladd, 2008). The phonetic realisation is then considered to be dependent on the phonological context and can be specified by rules. This two-step approach to the production of pitch accents is therefore contradictory to the idea of Exemplar Theory, where, in speech production, a speaker is assumed to construct a production target from the previously perceived (and then stored) set of exemplars in one step.

The two conflicting basic assumptions of autosegmental-metrical intonation models on the one hand, and Exemplar Theory on the other hand, might be the reason why to date no exemplar-theoretic model exists that incorporates intonation; nor is there an intonation model that works in a usage-based fashion.[2] Until very recently, the two research branches were mostly separate. However, Calhoun and A. Schweitzer (forthcoming) look at intonation from an exemplar-theoretic perspective and describe a corpus study that argues for lexical storage of intonation, Braun and Johnson (2011) find that Dutch speakers are attentive to pitch differences between non-words in stimuli comparison, indicating that tonal information is preserved, and Braun et al. (2011) show that lexical access is inhibited when an utterance with an unfamiliar intonation contour is perceived,

[1] But see Sprenger et al. (2006) for "superlemmas" to represent idiomatic phrases.

[2] But see Hawkins and Smith (2001) for a theoretical position which argues for a model that allows for the storage of rich phonetic detail with pitch information included in the mental representation.

a finding which again supports a model in which lexical items are represented with fine tonal detail.

The work described in this thesis aims to contribute to bridging the gap between exemplar-theoretic research and research on intonation. The main objective is to examine whether frequency of occurrence effects on pitch accent realisation can be observed and how the possible findings could potentially be integrated into existing exemplar-theoretic models. Specifically, this involves answering the following questions:

Does frequency of occurrence affect pitch accent realisation? Frequency effects of pitch accent type, information status category, word-accent pairs, and word sequences will be examined to obtain a comprehensive analysis of which aspects of the linguistic contexts of pitch accent tokens are subject to exemplar-theoretic effects. Pitch accent realisation will be analysed in terms of pitch accent variability and pitch accent shape. The results have significant implications for theories of intonation which will be discussed.

How do the word and tonal level interact in exemplar-theoretic selection? The relevance of pitch accent labels in exemplar-theoretic selection will be examined by investigating if competitors to a word-accent combination can interfere with a new production.

What are the essential features of an exemplar model accounting for intonation? The results of the corpus experiments will be interpreted from an exemplar-theoretic viewpoint. Starting out from a simple exemplar-theoretic model, some modifications will be outlined that have to be made in order to account for the effects found in the corpus studies.

A description of the corpus studies as well as some aspects of their implications were published in peer-reviewed conference proceedings (Schweitzer et al., 2009a,b, 2010b,a, 2011). The outline of this thesis is as follows: Chapter 2 gives an overview of the concepts and main ideas of Exemplar Theory, including a section on frequency of occurrence effects in speech. Chapter 3 describes models of intonation, focusing on autosegmental-metrical models. Chapter 4 describes literature that connects the two: it gives an overview of exemplar-theoretic effects on other prosodic parameters and it sums up evidence for the lexical storage of intonation. Chapter 5 then describes the corpus-phonetic analyses carried out on German and English speech data. Chapter 6 gives a general discussion, describes the implications of the findings for exemplar-theoretic models and presents ideas for future work, before chapter 7 offers some concluding remarks.

2 Exemplar Theory

Exemplar Theory was introduced in psychology as a model of perception and categorisation (Medin and Schaffer, 1978; Nosofsky, 1986; Nosofsky et al., 1992; Kruschke, 1992). It has been applied to speech perception (e.g. Goldinger, 1997; Johnson, 1997) and production (e.g. Pierrehumbert, 2001; Walsh et al., 2010; Wade et al., 2010). The key idea of exemplar-theoretic perception is that information in the episodic memory is sufficient for the classification of perceptual stimuli and that the procedural memory does not have to be queried. That is, stimuli are assumed to be stored as episodes, rich in detail, not as descriptions.

Applied to linguistic input, the central assumption of Exemplar Theory is that concrete language input is *stored in memory as exemplars in rich detail* (see section 2.1.2 for experimental evidence). These stored exemplars are employed in the categorisation of new exemplars: new input is compared to other exemplars in memory, with newer exemplars being more influential and older exemplars gradually fading away. The new exemplar is associated with similar exemplars – this process can be seen as an exemplar-theoretic categorisation process: an exemplar belongs to the category that contains the most exemplars which are similar to it. Stored exemplars are accessed as production targets. Self-produced new productions are also stored in memory. Hence, the exemplar memory is constantly updated and its content is used as reference for production and perception.

The main difference between such a *usage-based* system and generative models of language processing (essentially models based on Chomsky, 1957) is that there is less abstraction from the fully specified input.[1] Generative models of speech production (e.g. Chomsky and Halle, 1968; Levelt, 1989) assume that lexicon and phonological grammar are two distinct components. The lexicon contains an abstract representation for each unit. In speech production, the lexicon is accessed and the surface form of the utterance is constructed according to rules specified by the grammar. Then, a phonetic component generates the exact speech signal in terms of acoustic and/or articulatory properties.

[1] Some "radical" implementations even assume no abstraction at all (e.g. Wade et al., 2010)

Hence, in such a generative model of speech, neither variance that occurs within a category nor frequency effects can be modelled (Pierrehumbert, 2001).[2] In exemplar models, on the other hand, the realisation of a given unit is directly derived from the previously perceived instances of that unit. The frequency with which a unit occurs is encoded in the lexicon: every time a unit is perceived, it is stored as an exemplar.[3] Therefore frequent units are represented by many exemplars, infrequent units are represented by few exemplars. The variance that occurs within a category is explicitly encoded, as well: all the various realisations of a given unit are stored and will influence new productions. Thereby Exemplar Theory is able to account for variance within a category and effects of frequency of occurrence that have been reported in the literature, whereas traditional models of speech production and perception have no evident means of modelling such effects.[4]

The following chapter first presents experimental findings that motivate Exemplar Theory in section 2.1. Then, several exemplar-theoretic models that have been suggested in the literature are summarised (section 2.2). Finally, section 2.3 outlines a simple exemplar theoretic model of speech production that will be referred to when the results of the corpus studies (see chapter 5) are discussed.

2.1 Motivation

The motivation for the episodic storage of linguistic units given here is three-fold. Firstly, systematic variation within a linguistic category presents a challenge for the assumption that such a category is stored as an abstract unit and that the respective surface representation can be derived by applying a finite set of rules. Evidence for variation occurring systematically in speech is discussed in section 2.1.1.

Secondly, research indicates that fine phonetic detail is stored in a speaker's memory (see section 2.1.2). Such results are also challenging for a generative approach: if the speakers of a language react to fine phonetic detail – be it linguistic or non-linguistic – then the detail must be part of the mental representation.

[2] The concept of *Optimality Theory* (e.g. Prince and Smolensky, 1993) however, adds a way of handling variance in a generative system by introducing constraints that allow variance in a controlled way.

[3] However, if new exemplars differ only very little (i.e. they differ from an already stored one in less then 1 JND (just noticeable difference), and can therefore not be differentiated from it by a speaker), they are often assumed to be stored as if they were identical to the already stored exemplar: the perceptual space is granularised (see e.g. Kruschke, 1992; Pierrehumbert, 2001).

[4] See McQueen (2005) for a discussion of the challenge that variation presents to abstract models, including the suggestion of position-specific prelexical allophones that encode some of the observed variation for a given sound.

The third angle on evidence for exemplar-based storage is frequency of occurrence . If the frequency with which a unit occurs has an effect on the production or perception of that unit then it must be part of the mental representation, as well. While one could argue that generative models could be enriched with auxiliary mechanisms such as counters that keep track of the occurrence of each unit, an exemplar-model of language provides an approach in which frequency effects can be explained naturally since the frequency of each unit is inherently encoded in the lexicon with each perceived unit being stored and similar units being accumulated to represent one category. Since the goal of the experimental work presented here (chapter 5) is on the detection of frequency of occurrence effects on pitch accents, the focus of the current section is on frequency of occurrence effects; they are described in 2.1.3.

2.1.1 Variation in phonetic detail

Phones vary in fine phonetic detail, that is, phones realising the same phoneme can differ considerably in how they are realised acoustically. This variation presents a challenge for generative speech production models, in which a set of implementation rules defines how a particular phoneme is realised acoustically. The phonetic implementation of a phonological form can differ between languages, varieties, sociolects, morphological classes or semantic fields (Pierrehumbert, 1999). For instance, the implementation of stop consonants varies between Canadian and European french (Pierrehumbert, 2001). Pierrehumbert (2001) even states that "it is not possible to point to a single case in which analogous phonemes in two different languages display exactly the same phonetic targets and the same pattern of phonetic variation in different contexts" (p.137). While it could be argued that such differences can be captured by different implementation rules for different languages, there are types of variation in fine phonetic detail which present bigger challenges to generative approaches.

Pierrehumbert (2002) coins the term *word-specific phonetics* which emphasises that the phonetic implementation of the same phonological form can vary across words. Such variation cannot be captured by applying the same phonetic implementation rules to the phonemes which compose the respective word. One example of systematic variation in the realisation of the same phoneme depending on the word in which it is contained, comes from Quebecois French, where a particular semantic class of words (terms that denote authorities, such as organs of the church and the military) resist vowel shifts that other words with the same phonological properties undergo (Yaeger-Dror and Kemp, 1992; as cited by Pierrehumbert, 2002, p.110).

On top of that, changes in phonetic implementation often occur gradually (Bybee, 2000). Therefore Bybee (2000) highlights the advantage of instance-based models of the lexicon: gradual changes cannot be captured by generative, rule-based approaches to phonetic implementation, since units smaller than phonemes are involved. Bybee (2000) argues that "an appropriate model of the lexicon must allow for the representation of ranges of phonetic variation, and these ranges do not necessarily coincide with traditional or generative phonemes" (Bybee, 2000, p.69). Examples of such gradual sound changes within one subject due to socio-linguistic factors are provided in the studies of Harrington and colleagues (Harrington et al., 2000a,b; Harrington, 2006) and Hay et al. (1999). Harrington et al. demonstrate gradual changes in the pronunciation of Queen Elizabeth II's vowels towards a more "mainstream" pronunciation, usually employed by younger speakers and by people from the "middle-class". Hay et al. (1999) analyses Oprah Winfrey's realisation of the diphthong /aY/, which varies between [ay], the General American English variant, and the monophthongised version [a:] which is characteristic for speakers of African American English. Both the frequency of the word as well as the ethnicity of the person that Winfrey was talking about were significant predictors on the realisation of the diphthong..

There are many studies that find different words to be realised differently depending on their frequency of occurrence (see section 2.1.3 for an overview).

Taken together, these findings which document variation in fine phonetic detail of the same phonological form, suggest that there is no universal form of a given phoneme, but rather that the phonetic form is learned during acquisition and constantly updated throughout lifetime – so the actual phonetic implementation is usage-based.

2.1.2 Storage of phonetic detail

Contrary to what generative theories of speech perception (cf. e.g. Green et al., 1991) suggest, there is evidence that speakers of a language do store information in memory that goes beyond the abstract linguistic representation of what they heard. Several studies suggest that fine phonetic detail is stored and that this information can be accessed during subsequent speech production and perception.

The studies show effects in adult speech as well as in first language acquisition. For instance, Jusczyk et al. (1993) demonstrates that in early L1-acquisition phonetic detail of previously perceived words is retained. Specifically, their study shows that infants under one year of age store information about speaker identity: after having heard a story from the same speaker for 10 days, they would listen significantly longer to word

lists from that story if the list was read out by the familiar voice as opposed to a new voice.

As for speech perception in adults, Palmeri et al. (1993) carried out two experiments that examined item recognition (classifying a word as "old", i.e. previously heard, or "new") and voice recognition (classifying the word as "new", "same", i.e. previously heard in the same voice or "different, i.e. previously heard in a different voice). The experiments showed that firstly, item recognition of "old" items was more accurate and faster when the stimulus was repeated in the same voice, and secondly that stimuli that had been heard in the same voice before were recognised as "same" more accurately and faster than stimuli that had been heard in a different voice. By increasing the number of speakers in the set of stimuli and comparing the effects to potential effects of gender the testing procedure ensured that the observed effect was not an effect of the speaker's sex, but really of a familiar voice, whereby Palmeri et al. (1993) corroborate other studies that report a same-speaker advantage (e.g. Schacter and Church, 1992). Other studies found word identification to be facilitated when the words were spoken by a familiar speaker (Nygaard et al., 1994; Nygaard and Pisoni, 1998).

Strand (2000) also shows that phonetic detail about previously perceived stimuli is retained in long-term memory. Her study investigated the effect of gender stereotypes on spoken word processing. In a pre-study stereotypical as well as non-stereotypical male and female speakers were identified. A subsequent word repetition task revealed that listeners could repeat the words produced by stereotypical speakers faster and thus processed the word quicker. Hence, stimuli that are produced by a voice that is similar to many previously perceived instances are processed faster than stimuli produced by unfamiliar voices, suggesting shorter processing times. From an exemplar-theoretic perspective, one could argue that gender performance is encoded in the exemplars serving as reference for the incoming tokens. Johnson (2006) argues that listeners have certain expectations about incoming stimuli. If these expectations are not met, processing is slowed down. The augmented version of his exemplar model is capable of dealing with such effects (Johnson, 2006, cf. section 2.2.2).

Both Johnson (1997) and Goldinger (1997) summarise many experimental results indicating that significant storage of language input exemplars, rich in detail, takes place in memory. All these findings contribute to the idea of a model of language processing in which categories are represented in an instance-based way so that variability within a category is part of the mental representation.

Since the main focus of the work presented here, will be frequency of occurrence effects, the next section will present some experimental results concerning frequency of occurrence in greater detail.

2.1.3 Frequency of occurrence effects

There is a voluminous body of experimental evidence that lexical frequency affects speech production and perception. As outlined before, this presents a challenge for traditional, generative models of language and speech. The following section gives an overview of frequency effects by discussing a number of key studies in various domains. The evidence is presented in two parts: the first and main part of this section is dedicated to production studies which look at processes of reduction, i.e. durational shortening and reduction of phonemes, to the point of their complete deletion (Baker and Bradlow, 2009; Bybee and Scheibman, 1999; Bybee, 2000; Jurafsky et al., 2001; Losiewicz, 1992; Wright, 1997).

The second part gives an overview of experiments that focus mainly on the effect that frequency of occurrence has on human speech processing. The main work in that area looks at response latencies and error rates from psycholinguistic experiments (Bien et al., 2005; Cholin et al., 2006; Luce and Pisoni, 1998) but the section finishes by summarising work by Hay (2001), who carried out a perception study that examines how the relative frequency of an affixed word given its base influences its perceived complexity.

2.1.3.1 Frequency effects on parameters measuring phonetic reduction

There is a considerable body of evidence which demonstrates that the degree to which a linguistic units undergoes processes of reduction is related to frequency of occurrence. All the work presented in the following section investigates reduction effects on different levels. Taken together, the studies cover the investigation of different linguistic units, as well as different ways of calculating frequency of occurrence. For instance, while Bybee (2000) and Jurafsky et al. (2001) examine effects of word frequency, Croot and Rastle (2004) investigate effects on the syllable level and Bybee and Scheibman (1999) analyse collocations of words. Likewise, the methodologies differ, so that taken together the studies document numerous effects of different probabilistic measures, such as absolute frequency (Bybee and Scheibman, 1999; Bybee, 2000; Losiewicz, 1992), various measures of relative frequency, that is probability values (Jurafsky et al., 2001), as well as frequency values in combination with other factors such as the density of the phono-

logical neighbourhood of a word (Wright, 1997) or the number of times it has been mentioned previously in the discourse (Baker and Bradlow, 2009).

Interestingly, as diverse as the methodologies are, all studies have in common the general finding that greater frequency entails a greater degree of reduction.

The remainder of the section describes each study individually, presenting them in chronological order.

Losiewicz (1992) measured the acoustic duration of the English past tense morpheme *-ed* for segmentally similar (i.e. rhyming) high- and low-frequency verbs (e.g. kneaded vs. needed, called vs. mauled). The suffix was realised significantly longer in the low-frequency condition. Furthermore the morphemes were longer than non-morphemic homophones (e.g. *rapped* vs *rapt*). Losiewicz (1992) explains these results by assuming different ways of processing and argues for a dual-access frequency-driven theory of morphology: while simplex forms are stored as one lexical representation (*rapt*), complex forms are combined on-line (*rapped*). High frequency inflected verbs (such as *needed*), however, can be stored in the lexicon as one representation, resulting in an overall shortening of the lexicalised unit as a whole. This hypothesis is corroborated by (Bybee, 2000, see below) who complements the results showing that high frequency regular past tense verbs (e.g. *called*) are more likely to undergo d-deletion than low-frequency verbs (e.g. *mauled*).

In an exemplar model such as the one suggested by Walsh et al. (2010) (see p. 43 for a description of the model) these effects can be explained by different types and numbers of exemplar clouds (one for the word as a unit or several for its constituents) being accessed.

Wright (1997) found a relationship between the production of a word, its lexical frequency and the number of words to which it is phonologically similar. His work was inspired by an earlier study on intelligibility carried out by Luce (1986), who found a word's intelligibility to be affected by not only lexical frequency (as it had been shown in numerous studies, before, see Luce, 1986, p. 5) but also the relative frequency of the lexeme compared to its nearest phonological neighbours and the neighbourhood density.[5] Wright (1997) classified monosyllabic CVC-words according to Luce's intelligibility probabilities as "easy" and "hard", respectively. Easy words were defined as having a high relative frequency compared to their neighbours and a sparse neighbourhood, i.e. there

[5]The number of neighbours was retrieved from an on-line lexicon based on Webster's pocket lexicon, 1967 (see Luce, 1986, p. 6)

are few phonologically similar words. Hard words, on the other hand, come from a dense neighbourhood and their relative frequency is low, compared to their neighbours. All words were of high word familiarity as defined in the Hoosier Mental Lexicon (HML, cf. Nusbaum et al., 1984), which was designed to estimate a native speaker's mental lexicon.[6] Words of high familiarity in the HML were found to be also highly frequent (Large and Pisoni, 1998).

Wright (1997) measured the degree of dispersion of the vowel production by measuring the Euclidean distance of F1 and F2 (in Bark) from the center of the vowel space. F1 and F2 were measured in the middle of the vowel, at the "point of maximal displacement", i.e. where F1 and F2 displayed the most characteristic configuration for the respective vowel. He found greater dispersion for "hard" words, that is they were produced less reduced than "easy" words. Hence, words with a high relative frequency and a low neighbourhood density (easy words), showed more reduction/centralisation.

From an exemplar-theoretic perspective, one could argue that easy words have less "competitors", that is less words that could potentially occur instead of the target word. Hence, the lower the number of competitors, the higher the degree of reduction (i.e. the lower the degree of dispersion). In other words: as relative frequency increases (and neighbourhood density decreases), the words become more reduced.[7]

Bybee and Scheibman (1999) found the degree of *don't*-reduction in collocations (such as *I don't know*) to be related to the frequency of occurrence of that collocation. In their analysed corpus of conversational speech, *don't* is more reduced when occurring with a pronominal subject. Interestingly, those cases are also the most frequent ones. Among them, the most frequent collocation, the collocation *I don't*, displays the greatest degree of reduction. Hence, the frequency of the collocation subject+*don't* correlates with the degree of the phonetic reduction.

Moreover, additional analyses revealed that the reduction of *don't* in the collocation *I don't* is stronger than in any collocation of *don't* and a subsequent verb. The authors argue that this is phonological evidence that *I* and *don't* form a constituent i.e. that the auxiliary is connected more strongly to the subject than to the verb. This analysis differs drastically from a constituent analysis in generative syntax, where the auxiliary

[6] The familiarity rankings are based on rankings of native speakers, see Nusbaum et al. (1984) and Large and Pisoni (1998) for more detail.

[7] In the experimental work presented in chapter 5 here, the number of competitors on the suprasegmental level was also found to be influential on the production of a word, in this case the tonal realisation, see section 5.3, p. 131.

and the verb are analysed as one constituent that combines with the subject. Bybee and Scheibman (1999) claim that the degree of constituency is determined by frequency of usage: the transition probability of *I don't* in their data[8] is very high with *I* being the most frequent item to precede *don't* and *don't* being the second-most frequent item to follow *I*. Hence, according to Bybee and Scheibman (1999), the high relative frequency of the words *I* and *don't* determines their combining to one constituent.

Moreover, the position preceding *don't* is filled by fewer types (that each individually comprise more tokens) than the one following it, which indicates greater collocation strength of *I* and *don't* than the collocations of *don't* with a subsequent verb or adverb, reflecting that those positions differ in how flexibly they are filled in actual language use. With respect to lexical storage, the authors explicitly conclude that "the fewer types and the resultant higher token frequency of individual types creates a stronger syntagmatic relation in the storage unit." (Bybee and Scheibman, 1999, p.592).

The greater collocation strength, determined by the frequency of usage, is reflected in the greater phonetic reduction of *don't* in these collocations.

Such a finding, where frequent contexts display a greater degree of phonetic reduction, corroborates a usage-based account of linguistic structure while traditional generative, rule-based models will encounter problems due to the lack of a mechanism that considers the frequency of a lexical item or collocation when applying a phonetic implementation rule. Reduction effects like the aforementioned one would have to be modelled as (many) exceptions to a rule or a counter would have to be introduced.

Bybee (2000) investigates t/d-deletion in English. This process of deleting a word-final coronal stop has extensively been examined in the literature before, and was attributed to phonetic (the stop is more likely to be deleted if the following words starts with a consonant), grammatical (deletion in past tense verbs is more likely if the stop is part of a regular past tense marker suffix) and social (younger people and members of lower social classes are more likely to delete the stop) factors (Bybee, 2000). It has been claimed that the sound change which causes the final coronal stop to be deleted is lexically gradual, that is, individual items/groups in the lexicon are affected in sequence, not synchronously but phonetically abrupt (Wang and Cheng, 1977; see Bybee, 2000, p.66).

[8]This analysis was carried out on another corpus of conversational speech, due to practical reasons (see Bybee and Scheibman, 1999, p.592).

Bybee (2000) finds the rate of t/d-deletion in 2000 spoken English word tokens[9] to be positively correlated with their text frequency. Token frequency was recorded as listed in the Brown Corpus (Francis and Kučera, 1982). Words with a relative frequency of less then 35 occurrences per million in the corpus were coded as low frequency words, words that occur more than 35 times were coded as high frequency words. In the high frequency group, final t/d-deletion was found in 54.4% of all cases, in the low frequency group in only 34.3%.[10] This difference was statistically significant. Bybee concludes that the sound change seems to be frequency-driven, gradually spreading in the lexicon, according to how often a word has been used.

Such a finding argues in favour of a usage-based model of speech in which units are affected by sound changes every time they are used. Hence, frequent words would display a higher degree of deletion:

> One way of accounting for the continuous nature of such sound changes and the frequency effect they show in lexical diffusion is to assume that sound changes affect words opportunistically each time they are used. The sound change 'rules' apply real time, so that a frequent word has more exposure to the sound change than an infrequent one. If the effects of the sound change are cycled back into the lexicon, as the speaker monitors his/her own speech and the speech of others, the lexical representations for the words gradually adjust to the new productions (Moonwomon, 1992).
> (Bybee, 2000, p.69)

Bybee (2000) proposes that the sound change is also phonetically gradual, not abrupt. She illustrates and supports this proposal by articulatory data from Browman and Goldstein (1990), who found that the lingual gesture for /t/ in the phrase "perfect memory" was present, but overlapped by the adjacent gestures (velar and labial gesture for /k/ and /m/). That is even though the articulatory gesture is present, the phone is perceived as being deleted. Hence deletion could be a gradual process, where the gesture for the coronal stop is slowly shortened until it is perceived as deleted which in turn leads to actual deletion.

Furthermore, Bybee (2000) examines the sound change for past tense words whereby she tests the assumption that the process is grammatically driven and dependent on morphological status. Past tense morphemes have been found to be deleted less often

[9] words that potentially end in a /t/ or a /d/ following a consonant

[10] It has to be noted that the groups differ in the number of tokens, since 80% of the analysed data was coded as having a high frequency, 20% as having a low frequency.

than mono-morphemic words (e.g. Labov, 1972; see Bybee, 2000, p.74). Moreover, irregular past tense verbs have a higher deletion rate than regular ones. These differences have been attributed to the structural differences in word tokens (e.g. Guy, 1991; see Bybee, 2000, pp.74) .

In the analysed data, Bybee (2000) finds a significant change in the rate of t/d-deletion according to token frequency within the set of regular past tense verb forms. Irregular past tense verbs for which the token frequency in the data is generally higher, are also exposed to deletion more often. Moreover, the deletion rate correlates with their individual token frequencies with the most frequent irregular past tense verb (*told*) being subject to deletion most often. Hence, the reported effect in the literature (irregular past tense verbs are subject to deletion more often than regular ones) could also be explained in terms of token frequency. Moreover, the frequency-driven changes in deletion rates within the group's past tense verbs that are equally complex in their morphological structure, cannot be explained with a morphological account of sound change.

Bybee (2000) suggests a model that operates on the word level and assumes lexical storage of even morphologically complex words. It assumes that sound changes apply to the exemplars in real time, and that they are thus spreading gradually from frequent to infrequent words. This usage-based explanation for the phenomenon of t/d-deletion is able to explain all the findings summarised in Bybee (2000), contrary to other models such as a structuralistic account by Guy (1990; cited in Bybee, 2000).

Jurafsky et al. (2001) relate the phonetic reduction (manifested as vowel reduction, final t/d deletion or shorter duration) of a word to various measures of its predictability, viz. measures of relative frequency.

Two datasets of words extracted from a labelled subset (Greenberg et al., 1996) of the Switchboard corpus (Godfrey et al., 1992) were analysed. The first dataset comprises the ten most frequent English function words in the Switchboard Corpus (which are also the ten most frequent words). The second dataset contains all the content words ending with either t or d.

Jurafsky et al. (2001) assess the effect of several measures of probabilistic relations between words, most of which show some effect on reduction. However, in the article, they focus on two measures: *relative frequency* and *conditional probability*. They are calculated as follows:

The *relative frequency* of a word is calculated by

$$P(w_i) = \frac{C(w_i)}{N} \tag{2.1}$$

where $C(w_i)$ is the number of occurrences of the target word, and N is the total number of word tokens in the corpus. .

The *conditional probability* of a target word w_i given a neighbouring word (here the previous word w_{i-1}) is calculated by

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \tag{2.2}$$

where $C(w_{i-1}w_i)$ is the number of times the two words occur together and $C(w_{i-1})$ is the number of times the preceding word occurs. Thus, conditional probability combines several independent probabilistic measures: the joint relative frequency of the two words and the relative frequency of the neighbouring (here preceding) word.

Phonetic reduction of function words was characterised by vowel reduction and shortening of the word. Jurafsky et al. (2001) fitted regression models to predict the word duration and the frequency of vowel reduction with the probabilistic measures as predictors. They controlled for other factors that are known to be influential on vowel reduction such as speech rate, segmental context, syllable type of the word or reduction of the following vowel (by which they aim to control for stress, since the reduction of the following vowel indicates the stress level of the current vowel). Conditional probability was found to be a significant predictor of durational shortening as well as of vowel reduction: the greater the conditional probability, the greater the reduction (higher likelihood of reduced vowels and shorter word durations). This was true for the conditional probability of the word given the preceding word as well as for the probability of the word given the following word. Moreover, the conditional probability of the word given both adjacent words, also yielded a significant effect. Hence, probabilistic relations between a word and its context influence its phonetic realisation.

Phonetic reduction of content words was characterised by final t/d-deletion and shortening of the word. Additional to the factors which they controlled for in the function word analysis, Jurafsky et al. (2001) also included inflectional status, the identity of the underlying segment (t or d) and the number of syllables in the regression model predicting the two measures of phonetic reduction. The words were found to be significantly shorter and more likely to have deleted final t/d with increasing relative frequency. Hence the greater the relative frequency the greater the phonetic reduction. Analogously, words

with a high conditional probability were found to be shorter, that is, more reduced. However, no effect was found for final segment deletion, hence words with a high conditional probability are not more likely to have their final segment deleted.

Overall, the results of Jurafsky et al. (2001) demonstrate that probabilistic factors influence the phonetic realisation of words and present therefore a challenge for generative models of speech production. This is stated explicitly by the authors:

> By showing that probabilistic factors influence lexical production, our results also provide general support for probabilistic models of human language processing (Jurafsky et al., 2001, p. 18)

In such a probabilistic model of speech production, conditional probability can be regarded as a measure of how many competing realisations a target utterance has. If a word occurs relatively often given the context, then there are many perceived instances of exactly this sequence of words in a speaker's memory. Hence, there are few instances in which the word occurs in other contexts. Since perceived and stored instances are supposed to be used as production targets for new productions, few competitors stand for less interference with a new production. For frequent cases, the productions are expected to be entrenched over time and produced with less production effort – hence the reduction.

Baker and Bradlow (2009) investigate how the interplay of probability, speech style and prosody affects word duration. Under "probability" they subsume effects of lexical frequency (derived from the British National Corpus) and what they call *previous mention*, that is, whether the word has already been mentioned in the discourse (and is therefore more likely to occur again). Speech style was differentiated as either *plain* where speakers were instructed to read paragraphs as if they were reading them to a friend, and *clear* where they were instructed to read as if they were reading to a person with hearing loss or to a non-native speaker. Prosody was encoded as the presence or absence of a pitch accent or a prosodic phrase break. Baker and Bradlow (2009) found word duration to be reduced in more probable words, that is words that have a higher frequency and words that had been mentioned before. This result is in line with previous studies, e.g. the analyses of Jurafsky et al. (2001). The shortening of more probable words was found for both speech styles so the hyperarticulation found when reading out very clearly did not override the probabilistic effect. However, only in plain speech was the reduction of second mentions even greater for high frequency words than for low frequency words, which demonstrates that frequency and speech style interact.

When prosody comes into play, first mentions were more likely to be accented than second mentions confirming the often assumed deaccenting of given information (see Brown, 1983, and section 5.1.3.1). However, with different accent status being controlled for, second mentions were reduced, indicating an effect of probability on word duration irrespective of accent status.

Overall the results confirm the frequency effect found in Jurafsky et al. (2001) and they highlight the importance of context for the acoustic realisation of lexical items. In an exemplar-theoretic model the above mentioned effects would occur naturally: they could be simulated by a model that provides labels for all kinds of context (for example speaking style and discourse givenness) and takes those labels into account as selection criteria (see section 2.2.4) for choosing exemplars as production targets or as reference tokens for classification. Another alternative would be a model that explicitly encodes the context of a target unit like the *Context Sequence Model* (Wade et al., 2010, described in section 2.2.3) – there the labels would not even need to be explicitly assigned.

2.1.3.2 Frequency effects on other parameters

While many production studies focus on the effects that frequency of occurrence has on the segmental reduction of words, there is another line of research that aims to identify frequency of occurrence effects on language processing. The work in that area mainly concentrates on psycholinguistic experiments measuring reaction times and logging error rates from tasks that tap into lexical access.

The studies presented in the following differ in terms of the units they investigate. Cholin et al. (2006) look at processing differences between frequent and infrequent syllables, whereas Luce and Pisoni (1998) investigate the interaction of word frequency with the frequency and density of the words' phonological neighbours (cf. Wright, 1997). Finally, the experiment carried out by Bien et al. (2005) examines the effect of morpheme frequency on lexical retrieval.

Essentially, all studies, even though investigating different units, find that processing times are faster for frequent units.

They are presented individually in chronological order, followed by a perception study by Hay (2001) revealing that the perceived complexity of affixed words depends on the relative frequency of the word given its base.

Luce and Pisoni (1998) found results similar to the ones of Wright (1997; cf. p. 26 here) in a series of experiments on lexical access. The degree of phonetic similarity among

about 20,000 word tokens was calculated (according to their phonetic transcription given in *Webster's lexicon*). For each word, its neighbourhood density, that is the relative number of words that are phonetically similar to the given one, was determined in a similar fashion to Wright (1997): many phonologically similar words indicate a dense neighbourhood, whereas a sparse neighbourhood is characterised by few similar words. Moreover, for each word, its frequency in a 1,000,000 token printed corpus (Kučera and Francis, 1967) was determined to account for its frequency of occurrence in American English.

Three experiments on lexical access (perceptual identification of words in noise, auditory lexical decision, and auditory word naming) were carried out, with word length, number of syllables and familiarity of the word according to Nusbaum et al. (1984) being controlled for. The results indicate that effects of word frequency are rather a "function of the neighbours of the stimulus word as well as the frequencies of these neighbours" (Luce and Pisoni, 1998, p. 27). Hence, the three measures word frequency, neighbourhood density and word frequencies of the neighbours interact in lexical access. Specifically, words with many similar lexical neighbours were identified less accurately in lexical identification. Lexical decision and word naming both were faster for high-frequency words, for low density neighbourhoods[11] and for low frequency neighbourhoods. So here, beside the benefit that high-frequency words have in lexical access, the number of competitors (words with a similar pronunciation) seems to have an influence on the perception and processing of the tested stimuli.

Such results could be modelled in an exemplar-theoretic model, where incoming new words are compared to existing exemplars for categorisation: if there are many competitors to a new stimulus, there is more chance for misclassification.

Bien et al. (2005) found the frequency of both head and modifier in Dutch noun-noun compounds to be influential on response latencies in a symbol-position learning-task. Participants were presented with a pair of compounds and learned to associate each word to one of two positions on the screen. After the learning and a subsequent practising phase, they had to produce the appropriate word after being presented with one of the positions on the screen. Under different experimental conditions, the frequency of the head, the modifier or the whole compound was varied while keeping the other factors constant.

[11] However, accuracy was better for high density neighbourhoods, see Luce and Pisoni (1998), p. 21 for an explanation within their model.

In three experiments, Bien et al. classified heads, modifiers and compounds as frequent or infrequent. ANOVAs demonstrated that compounds with either a frequent head or a frequent modifier were retrieved significantly faster (lower response latencies) than compounds without a frequent constituent. The frequency of the compound itself did not influence reaction times significantly. However, there was an unexpected trend towards longer response latencies for frequent compounds.

In a fourth experiment, a regression analysis predicting the response latencies was calculated. Several frequency measures for the constituents together with compound frequency (which was, unlike in the ANOVAs carried out before, found to be significant) predicted the response latencies better than just compound frequency. Therefore, Bien et al. argue for decompositional models of speech production.

Furthermore, the regression analysis sheds light on the (non-significant) inhibitory effect of compound frequency in the ANOVAs: while for the lower range of compound frequencies the naming latencies decrease as compound frequency increases, this effect levels off and turns into increasing naming latencies in the higher compound frequency range. That is, the relationship is not a linear one. However, Bien et al. state that the apparent inhibition might be an artifact of the statistical methodology.

Overall the study of Bien et al. presents evidence that the retrieval of compounds in Dutch depends on frequency of occurrence. Since not only the frequency of the compound influenced naming latencies, but also the frequency of its components, the experiments argue for a decompositional model. The multi-level exemplar model (Walsh et al., 2010, described in 2.2.3) could probably account for both, the frequency effects of the constituents as well as the effects of the frequency of the complex unit.

Cholin et al. (2006) investigated effects of syllable frequency in Dutch in a symbol-position learning task. Participants had to associate auditorily presented nonsense-words consisting of existing Dutch syllables of varying frequency to one of two positions on the screen. After a learning phase in which they heard the correct stimuli for the respective position and a practise phase, in which they had to select the correct position for the stimulus, they were tested by being presented with a position and having to produce the correct stimulus as fast as possible. The syllables were controlled for onsets, offsets, phoneme and bigram frequency as well as transitional probabilities.

Monosyllabic pseudo words were produced faster, when the syllable they consisted of was frequent, than when it was an infrequent syllable. Moreover, participants made less mistakes in the high-frequency condition. Cholin et al. (2006) argue that this supports

the idea of a *Mental Syllabary* (Levelt et al., 1999; Levelt and Wheeldon, 1994),which assumes that frequent syllables are stored as articulatory gestures and can be accessed holistically, while infrequent syllables are produced on-line during production. Cholin et al. follow Levelt and Wheeldon (1994) in that they argue that frequency effects are only expected for units that are stored, and not for units that are computed on-line.[12]

Disyllabic pseudo-words, however, were only sensitive to frequency manipulation of the first syllable (and not to manipulation of the second syllable. Frequent first syllables lead to faster production. The error rates did not show a significant effect. Cholin et al. conclude that articulation can be initiated before the frequency-manipulated syllable was derived from the Mental Syllabary. They follow Meyer et al. (2003) in assuming that articulation cannot start before completion of phonological word encoding, but that it can start after phonetic encoding of the first syllable.

Analogous results (and reasoning) for Spanish can be found in Carreiras and Perea (2004): in a word naming task, they found frequency effects for the first syllable of Spanish pseudo-words, but not for the second one. The data controlled for bigram-frequency as well as for lexical stress.

Croot and Rastle (2004), however, in a syllable naming task, did not find frequency effects on response latencies or durations. Still, their experiment revealed an effect of syllable frequency on spectral properties: Following Whiteside and Varley (1998), they argue that prestored articulatory syllable plans should result in greater coarticulation. Their experiment therefore examined acoustic parameters reflecting different aspects of coarticulation.[13] They found a (non-significant) trend for greater coarticulation for high frequency syllables compared to (phonetically matched) non-existing syllables.

Overall, these findings provide evidence for the explicit storage of frequent syllables in a mental lexicon which is not assumed by traditional generative phonology which would claim that syllables are constructed in a rule-based fashion on-line during production. An exemplar-theoretic model, on the other hand, can explain such differences due to frequency of occurrence since the frequency of a particular stretch of speech is assumed be crucial for the emergence of a category (here: a syllable). If there is no exemplar cloud for a syllable, that is, not a sufficient number of exemplars, then the syllable is

[12] In fact, Cholin et al. (2006) corroborated Levelt and Wheeldon (1994)'s results by replicating them but ruling out a potential effect of segment frequency in Levelt and Wheeldon (1994)'s data that had been criticised (Hendriks and McQueen, 1996).

[13] frequency of fricative noise due to anticipatory coarticulation for fricative-initial syllables; lowering of F3 for vowels following /r,w/ due to carryover coarticulation; lowering of F2 of an open front vowel adjacent to a velar consonant due to more posterior articulation; and raising of F2 of central or back vowels with adjacent (post)alveolar consonants due to increased fronting

assumed to be constructed from its segments (Walsh et al., 2010; A. Schweitzer and Möbius, 2004).

Hay (2001) relates the semantic transparency of affixed words to the relative frequency of the derived form and the base (for instance, *inaudible* is more frequent than its base *audible*). She paired complex words (suffixed as well as prefixed forms) that were more frequent than their bases with counterparts that were less frequent than their base (e.g. *inaudible* vs. *inadequate*). The pairs were matched for junctural phonotactics, stress pattern, number of syllables as well as for the surface frequency of the derived form, using the CELEX database (Celex, 1993). In a perception study, 16 subjects compared the members of the pairs with respect to their perceived complexity and marked the one they considered to be more complex. Forms that were more frequent than the base they contained, were rated significantly less complex than their counterparts. Hay (2001) concludes that the relative frequency of the base form is involved in facilitating decompositionality (Hay, 2001, p.1049).

To get a deeper insight into effects of absolute lexical frequency of derived word forms that have been reported in the literature (e.g. Baayen, 1992, 1994; Bybee, 1995), Hay then analyses the relationship between relative frequency of the derivative and its base and the absolute frequency of the derivative. She finds the two highly correlated in a small corpus experiment (on CELEX data) and concludes that the reported effects might be artifacts of relative frequency.

Finally, Hay examines the semantic transparency of derived word forms using Websters 1913 Unabridged Dictionary. She assumes that a derivative is transparent if the base is used to describe its meaning in the dictionary (e.g. to dis*horn*: to deprive of *horn*s). If a word is not transparent, i.e. if the base is absent in its definition, she assumes that the derived form underwent a semantic drift. She finds relative frequency related to semantic drift, since words where the derived form was more frequent than the base were significantly less likely to mention their base in their definition, i.e. they are less transparent. This held for prefixed as well as for suffixed forms. The absolute frequency did not have a significant effect on transparency in both cases.

The results from Hay (2001) highlight the impact of frequency information on decompositionality and semantic transparency. To model such an effect, it has to be assumed that a word's representation in the mental lexicon can be modified according to the frequency, i.e. the number of times the word is perceived and/or produced.

2.2 Exemplar-Theoretic Models in the Literature

Traditional models of language (e.g. Chomsky and Halle, 1968) assume a symbolic level of representation that abstracts away from the variability and the detailed information that is available in linguistic behaviours. It is assumed that the tokens are recoded into canonical types during perception and production of language and speech. For instance, the phone /t/ is recognised as the phoneme [t], i.e. as the cognitive concept of the phone. The symbolic representation [t] replaces the detailed acoustic input that is perceived.

Contrary to that assumption, there is evidence that demonstrates that people are able to access detailed information about previously perceived stimuli – non-linguistic as well as linguistic ones. For instance, they remember the typeface of visually presented words, the exact phrasing of sentences and the voice in which utterances were presented (see Goldinger, 1997, pp. 36–45 for a detailed overview).

Findings like the latter ones indicate that listeners do not abstract away from the input signal but have access to all its acoustic detail. Exemplar Theory (Bybee, 2006; Goldinger, 1997; Hay and Bresnan, 2006; Johnson, 1997, 2005, 2006; Kruschke, 1992; Nosofsky, 1986; Nosofsky et al., 1992; Pierrehumbert, 2001, 2002) assumes that previously experienced stimuli are stored in memory in detail.

Exemplar-theoretic models have been proposed in different research areas, ranging from the perception of multidimensional stimuli (e.g. Nosofsky, 1986) over speech perception (e.g. Johnson, 1997; Wade and Möbius, 2007) and production (Pierrehumbert, 2001) to syntax and semantics (e.g. Walsh et al., 2007; Bybee, 2006).

In the following sections the origins of Exemplar Theory are presented briefly by outlining a model of human classification of abstract perceptual stimuli (section 2.2.1). Section 2.2.2 then describes an adaption of Exemplar Theory to speech perception. Section 2.2.3 outlines exemplar-theoretic models of speech production. Since the results from the experiments presented here have implications for exemplar-theoretic speech production models, this section is the most detailed one. Finally, section 2.2.4 focuses on the selection process in exemplar-theoretic production models since some of the results from the corpus studies presented here (see chapter 5) have implications for the exemplar-theoretic selection process in the production of tonal events.

2.2.1 Exemplar Theory in (non-linguistic) Perception

The first exemplar-theoretic models simulated human classification of perceptual stimuli (e.g. Nosofsky, 1986; Kruschke, 1992).

The basic assumptions were

1. Perceived stimuli are stored in memory as exemplars in full context.

2. Categorisation takes place by grouping similar exemplars together.

Hence, a perceptual category is not defined as an abstract concept but as a set of single experienced instances of that category that are associated with each other, according to their similarity.

Nosofsky (1986) specified a computational model (Generalised Context Model , GCM) in which exemplars of a category are stored and new stimuli are classified according to their similarity to stored ones. Similarity is measured as the distance between two data points in the perceptual space. When a new stimulus is perceived, it is compared to all remembered instances. Instances that are similar to the new input are activated, i.e. they are taken into account in the process of categorising the new item. To finally determine the category of the new exemplar, for each potential category the activations are added up. The new token is labelled as belonging to the category that displays the highest activation.

Subjects attend selectively to different dimensions of the stimuli, that is some dimensions are more important for the classification than others. To model this, attention weights are assigned to each dimension in the perceptual space which influence the calculation of the distance between two stimuli.

The first version of the model predicts categorisation data from identification data of two subjects and accounts correctly for 96.6% and 93.7%, respectively, of the data. The model is then augmented so that the newly perceived exemplar is categorised and added to the set of exemplars for the respective category. That is, every classified instance is fed back into the model. The augmented model predicts 98.4%/97.1% of the data correctly.

Kruschke (1992) extends the GCM by adding a supervised learning mechanism, in which the attention weights for the perceptual dimensions as well as parameters describing the association strength of a specific exemplar to a specific category, are learned.

A model that assumes highly detailed storage of single instances as opposed to an abstract representation for each category faces the question of how data-overflow can be dealt with. Kruschke's implementation does not store the exemplar as such, but its location in the perceptual space, which is more efficient. Moreover, the perceptual space is supposed to be granularised and exemplars that are very similar to existing ones are not stored separately (see footnote 3 on page 21). Still, the exemplar memory cannot be assumed to be infinitely big. Therefore, in Exemplar Theory it is usually assumed

that exemplars fade over time and the models incorporate a *memory decay* function. The model of Nosofsky et al. (1992) assigns each exemplar a base activation level which is subject to a time function. Thereby, recent exemplars have a higher base activation than older ones. Another option is incorporating an explicit memory decay parameter (e.g. Pierrehumbert, 2001, 2002, described in section 2.2.3).

2.2.2 Exemplar Theory in Speech Perception

Johnson (1997) applied Exemplar Theory to speech sounds, motivated by evidence that the speakers of a language remember the phonetic detail of previously perceived speech tokens (e.g. a facilitating effect of a familiar voice in word naming tasks cf. section 2.1.2). Since in Exemplar Theory instances are assumed to be stored directly, without any abstraction, variability (e.g. speaker-dependent changes in the realisation of the same unit) is preserved in the "lexicon" (the accumulation of perceived instances) and might help later in identifying new tokens.

The model proposed by Johnson (1997) is an adapted version of Nosofsky's model which is capable of dealing with auditory stimuli. Exemplars in this model are associations of auditory features with category labels. The auditory dimensions are assumed to be defined by the auditory system, whereas the category labels reflect classifications which might be relevant to the perceiver. These classifications could be linguistic as well as para- or extra-linguistic. Categorisation by comparison works in the model as follows:

1. The *auditory distance* between two exemplars is calculated as the Euclidean distance between the two exemplars incorporating an attention weight parameter for each auditory dimension that adjusts the importance of the respective dimension. It enables variation on certain dimensions to be ignored. This process can be thought of as shrinking or expanding the parameter space along a certain auditory dimension.

2. The *auditory similarity* between the two exemplars is an exponential function of the auditory distance. An additional parameter limits the influence of distant exemplars, so that the nearest neighbours' similarity is increased even more.

3. To determine the actual *activation of an exemplar* a base activation for each exemplar is assumed that is multiplied with the similarity value. Depending on the effect that is to be modelled, the base activation can be varied. Gaussian noise can be added, optionally to the calculated activation.

4. *Evidence for the membership of the incoming exemplar to a certain category* is the sum of activations of all the exemplars of the respective category.

In such a model, on-line adaption (e.g. to the speaker) is possible because the appropriate exemplars (the ones uttered from the same voice) are activated (Johnson, 2005). The frame of reference for newly incoming items is the set of experienced exemplars, not a vowel space, for instance. This means, it is variable and re-definable during speech perception. Therefore, the model can account for numerous findings from the literature. For instance, it can handle effects of speaker familiarity facilitating lexical access (cf. section 2.1.2) and sociophonetic effects (cf. section 2.1.1).

In Johnson (2006) the model is modified by a *resonance mechanism*. When an exemplar of a certain category is activated, the other exemplars in the category are also activated according to how well they fit into the category. Thus a further step is added to the model:

5. *Resonance within the assigned category* is added by activation of all the exemplars within the category. The strength of the activation depends on the weight between the respective exemplar and the category.

The effect of this is that activation spreads from one exemplar in a category to the others which causes the exemplar cloud to be more centralised. Therefore, perception is more categorical than it was in the former model.

The augmented model successfully simulates the facilitation effect of stereotypical voices in spoken word processing (Strand, 2000, see section 2.1.2) by activating all the exemplars of the recognised category and thus the properties of the voices that have been uttering it.

2.2.3 Exemplar Theory in Speech Production

Pierrehumbert (2001) extends Exemplary Theory to speech production. She outlines a model that is able to deal not only with what she calls *word-specific phonetics* (cf. section 2.1.1), but also with different kinds of *frequency effects* (cf. section 2.1.3). These effects include

Increasing variance with increasing frequency Since mature categories display a certain amount of variation, i.e. for significant acoustic parameters their distribution is expected to be Gaussian-shaped, rather then having a pronounced peak (Pierrehumbert, 2001, p. 144). This variance can arise within an individual speaker as well as across speakers (Goldinger, 1998).

Production biases (Lindblom, 1984), such as higher rate of lenition with increasing frequency (e.g. Bybee, 2000), i.e. target undershoot.

Entrenchment that is, the effect that frequent units display less variation. For instance, during language acquisition, with increasing number of exemplars, the variability of a phonetic category decreases (Lee et al., 1999). Pierrehumbert illustrates the phenomenon with the example of a child learning a string instrument: whereas in the beginning the notes will be highly variable, after years of practise (i.e. when having stored a high number of exemplars) there will be considerably less variance (see Pierrehumbert, 2001, p.147) – and less production effort.

The model of speech production provided in Pierrehumbert (2001) deals effectively with all the phenomena mentioned above. Speech production, within the model, works in the following steps (note that, for the sake of simplification, the model is outlined for one phonetic (e.g. acoustic or articulatory) dimension – even though exemplars are considered to be multi-dimensional):

1. The intention to produce a category activates an *exemplar list*, containing all the exemplars that are associated to that category.

2. Then, a phonetic target is constructed by *selecting* an exemplar or a set of exemplars from the exemplar list i.e. from the exemplar cloud forming the category). Here, Pierrehumbert (2001) suggests two alternative approaches to selection (see 2.2.4 for further details about the selection process):

   a) An exemplar from the list is *selected randomly*. This exemplar forms the base for constructing the phonetic target (as described in the steps below). To account for *memory decay*, the probability for exemplars to be chosen, depends also on the age of the respective exemplars, with newer exemplars being more likely to be selected. That is, the activation for new exemplars is higher than for old ones.

   b) The phonetic target is defined by randomly selecting a *location in the exemplar space* and *averaging over the neighbouring exemplars*, weighted according to their age: older exemplars have a lower activation and contribute less. This causes the distribution for a dimension in the exemplar space to gravitate towards the mean of the distribution. Thereby, variance is decreased as the category grows. Hence, this step enables the model to account for *entrenchment*.

3. To model variation within a category, as the category evolves, *production noise* is added to the respective phonetic dimension of the production target. Thereby variance increases with increasing frequency.

4. To account for production biases, a systematic bias is added to the production target. Hence, with increasing frequency of the category, the value of the respective dimension will slowly move towards the intended direction (indicated by the bias).

Such an exemplar-theoretic model of speech production is able to deal with frequency of occurrence effects. Pierrehumbert (2001) points out, that in general, a strength of exemplar-theoretic models is that

> they provide a foundation for modelling frequency effects, since frequency is built in to the very mechanism by which memories of categories are stored and new exemplars are classified. It is not necessary to posit special frequency counters whose cognitive and neural status are dubious.
> (Pierrehumbert, 2001, p.143)

It is important to note that such a model can also account for variation due to sociophonetic reasons, since it is assumed that exemplars can be indexed with values for e.g. speaker identity, social role or age (Hay and Bresnan, 2006). Generally, categorical attributes of the respective unit are often assumed to be attached to the exemplar as such a *label*. With regard to intonation, it has been argued that pitch accent type is assigned as a label to each exemplar (A. Schweitzer, 2010).

Walsh et al. (2010) specify a model that can deal with both syntactic and phonetic exemplar-theoretic effects. The crucial idea of their *Multi-Level Exemplar Model* is that there are two alternatives to produce a complex linguistic unit such as a syllable or a phrase: for frequent units it is assumed that an exemplar cloud has been shaped and that this cloud can directly be accessed. For infrequent units, the unit can be composed out of its underlying segments such as the phones or the words. The idea of such a *dual route encoding* is based on Levelt et al. (1999) who assume a *mental syllabary* in which frequent syllables are stored as units, while infrequent ones have to be composed on-line in speech production by concatenating the respective segments. Walsh et al.'s model successfully accounts for effects of syllable frequency on the relationship between syllable and segment durations (A. Schweitzer and Möbius, 2004, cf. section 4.1).

Analogously, in perception two alternatives are assumed: either a complex unit is perceived as a complex unit or it is perceived as a concatenation of (unrelated) constituents. Thereby, the model successfully simulated grammaticality judgements.

Walsh et al.'s model works in the following way: For the two processing alternatives, the model employs two databases, to which a unit is compared and in which it can be stored: the *unit database* for complex exemplars and the *constituent database* in for the decomposed constituents.

The model's *parser and composer* parses the unit and breaks it down to its components. The *similarity calculator* determines the similarity of the unit to both the units in the unit database and the respective units in the constituent database, in parallel. The similarity values are returned to the model and define (in the simplest case) the activation α of the respective unit in the unit database and of the individual constituents in the constituent database.

The *decision component* determines whether the unit path or the constituent path is chosen to produce or perceive the current input: If the activation of exemplars from the unit database goes beyond a certain threshold θ then the unit is produced/perceived as a unit. If this is not the case, the parsed constituents are retrieved from the constituent database. In production, this decision decides upon which set of exemplars is used to construct the production target. That is, if the activation of the exemplars from the unit database reaches the threshold, the exemplars of the complex unit are averaged over. Otherwise, the new production is based on the clouds of the respective segments. In perception, the consequence of the unit-set's activation reaching the threshold is that the new input is perceived as a valid complex input token. That is, the way in which the similar exemplars in the unit-database are parsed is taken to be the way in which the new input can be parsed. In this case, the exemplar is stored as a complex token, along with its decomposition into the constituents. If the threshold is not reached, the constituents are stored as unrelated single exemplars.

Figure 2.1 displays the architecture of the model. Only matching exemplars are shown in the unit and in the constituent database. Hence, if a stimulus does not have similar exemplars on the unit level, the exemplar cloud displayed in the figure would be empty.

The strength of the Multi-Level Exemplar Model lies in its capability to deal with different unit sizes. Hay and Bresnan (2006) illustrate that in an exemplar model, both should be assumed: the possibility to store units of varying length as well as their storage along with rich phonetic detail.

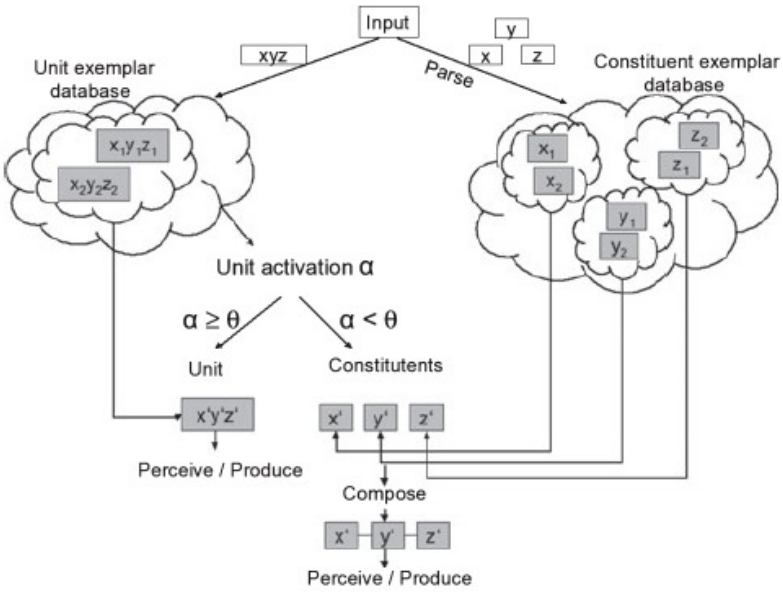


Figure 2.1: The Multi-Level Exemplar Model. New complex input is compared to the complex units in the unit database. These stored exemplars are activated according to their similarity to the new exemplar. If the activation α for the unit reaches a certain threshold θ, the exemplar is processed as a unit. Otherwise it is decomposed into its constituents and processed on the constituent level. In any case, the new exemplar is parsed and stored as both, a unit and as individual segments. (Figure from Walsh et al., 2010, p.547)

For an exemplar-based approach to intonation it is crucial that the model can deal with different unit sizes, especially with units that span multiple words. If one assumes that frequent phrases are stored with their intonation (cf. experiment 4 in section 5.4), then the exemplar-model on which an exemplar-approach to intonation is based, has to be able to account for that.

Wade et al. (2010) suggest the *Context Sequence Model (CSM)*, a context sensitive approach to memory-based speech production. The model is even more data-driven than the models presented above in that it does not assume the existence of any linguistic units. The basic idea is to incorporate the fact that speech unfolds over time and in context. Following the exemplar-theoretic view that linguistic categories can be defined by a description of the context in which they occur without making use of abstract rules Wade et al. (2010) assume that the context is crucial in perceiving or producing a unit. This explicitly includes contexts which are normally considered non-linguistic (such as speaker identity). Such information is useful in interpreting/producing speech signals since it provides information on how to properly interpret certain acoustic cues without using other co-occurring characteristics such as a particular fundamental frequency value.

Exemplars in the CSM are considered to be "stretches of speech much longer than the unit of interests" (Wade et al., 2010, p.228), i.e. exemplars are speech events in context. The model considers the left as well as the right context. For both contexts it is assumed that they can be purely defined acoustically. Hence, the left context can be encoded as the acoustic features of what has been produced/perceived so far, the right context can be encoded as estimations of acoustic features of what will be produced/perceived. However, in the actual implementation described in 2010, the right context has been encoded as the linguistic (i.e. categorical symbolic) information names about what is going to be produced next, for the sake of simplification.

In perception, an exemplar is compared in its context, hence the perceived production is compared to stretches of the same or similar length in memory.

Production is, as suggested by Pierrehumbert (2001), selection from a set of stored instances. The selection of exemplars is determined by the similarity of contexts of the stored tokens to the context of the planned token. The similarity score for the left and the right context are added up to the overall matching value.

The CSM can model and hence explain effects that have previously been used as evidence for hierarchical linguistic units. For instance, it models effects of syllable frequency (Cholin et al., 2006) without assuming a syllable level: segments produced as part of fre-

quent syllables were selected more efficiently and took on context-specific patterns (such as more variability and a higher degree of lenition) compared to the same phones in infrequent syllables. Hence, in the CSM, the importance of the surrounding context is dependent on frequency of occurrence.

With respect to an exemplar-theoretic approach to intonation, the model has the advantage that it can look at longer stretches of speech. Even though Wade et al. (2010) determined 1s of speech (0.5s of context on either side) is sufficient to properly perceive and produce segments in context, it would certainly be possible to adapt the model so that longer stretches of the speech signal are taken into account. Hence, tonal information (usually spanning several syllables or words) could be part of the acoustic information that contributes to select an exemplar. If tonal events can be perceived and produced successfully in such a model, this would question the need for a categorical, phonological label for tonal events.

2.2.4 Selection in Exemplar-Theoretic Models

Exemplar-theoretic models build on the idea that certain exemplars are activated for both production and perception. In perception, activated exemplars contribute to classifying the incoming stimulus: they are employed as references, that is, the similarity between the activated exemplars and the new input is calculated[14] and determines the exemplars with which the new exemplar is associated.

In production, those exemplars on which the construction of the production target is based on, are selected (i.e. activated). Exemplar-theoretic models differ in the way in which exemplars are selected. A description of the different mechanisms is given below. The two processes, viz. activating exemplars to classify incoming input in perception, and selecting single exemplars, sets of exemplars or regions in the exemplar space to construct a production target for the intended new production, are very similar. The following section explicates how the processes are defined in different exemplar-theoretic models.

As described in section 2.2.3, Pierrehumbert (2001) proposes different ways of selecting exemplars. The first way is *random selection* of an exemplar from a given exemplar cloud. In this case exemplars are weighted according to their recency which entails that older exemplars are less likely to be selected – this mechanism can account for memory decay. Then, one of the exemplars in the exemplar cloud matching the intended new

[14]More precisely, the degree of activation of each existing exemplar indicates its similarity to the new input.

production is selected randomly and used as a production target (to which noise and/or production biases are added, cf. section 2.2.3 above). The second way is to select a set of exemplars by choosing a target location and selecting a fixed number of exemplars (the *k-nearest neighbours*[15]) around the target location. A production target is then created by averaging over this set of exemplars.[16]

The models described in Goldinger (1997) and Hintzman (1986) also activate a set of exemplars (here for classification in perception). However, unlike Pierrehumbert's (2001) model, in which a fixed *number* of neighbours around a specific location in the exemplar space is selected, they activate a region around the location with a fixed *size*. As a consequence, in these models rare events are classified on the basis of fewer instances than frequent events because their clouds are less dense. Since these models implement an "echo" (similar to the resonance mechanism in Johnson's (2006) model described above), which influences what is being stored as an exemplar based on the activated neighbourhood, Pierrehumbert (2001) stresses that more episodic information is encoded for rare events leading to a better memory for acoustic detail of rare events, e.g. that "one is more likely to remember that a word was spoken in a particular voice if the word is rare than if it is common" (Pierrehumbert, 2001, p. 149). Though this is a desirable effect in perception, Pierrehumbert (2001) did not implement such an algorithm since it turned out not to be suitable in simulating production: at the beginning of the simulations, when new exemplar clouds are created, the fixed neighbourhood size led to instability in the exemplars, since too few instances contributed to the production of a new exemplar.

Wedel's implementation of an exemplar-theoretic model presents a somewhat hybrid version: first, a random exemplar is picked, then, two more exemplars are chosen within a random Gaussian window around the first one (Wedel, 2006). Consequently, it is likely that exemplars close to the initial exemplar are chosen. To generate a production target, the model then averages over the three exemplars. In this way, exemplars do not split into subgroups and entrenchment can be modelled.

The models of Johnson (2006), Walsh et al. (2010), Wade et al. (2010), Kruschke (1992), Nosofsky (1986) (which is the basis for Kruschke's model) and other instance-based models of attentional learning described in Kruschke (2011) also activate/select a set of exemplars. They have in common that *all* the exemplars of a category, that is, all the exemplars with a certain label, contribute to the classification process. It has to be noted that all these models have sophisticated algorithms to calculate the appropriate

[15] where k is the number of neighbours that are to be selected

[16] In both cases, the likelihood of an exemplar to be selected is dependent on its activation strength, see Pierrehumbert (2001).

degree of activation. For instance, ALCOVE and RASHNL, two models implemented by Kruschke (Kruschke, 1992; Kruschke and Johansen, 1999) activate the exemplars of a certain category according to their similarity to the incoming stimulus. The latter model is able to learn which dimensions are relevant for classification and calculate the activation according to the weight of the respective dimension. Successors of these models are able to calculate this selective attention to certain dimensions separately for the incoming stimuli, that is, the selective attention to certain dimensions is exemplar specific (see Kruschke, 2011, for an overview).

However, for the research questions and the experiments presented here, and also for the implications of the experiments on the selection process within an exemplar-theoretic production model that incorporates intonation (cf. section 6.2 and 6.3) it is important to remember that there are different ways of selecting/activating exemplars: some models select a random exemplar, others a set of exemplars where the number of exemplars is fix (both Pierrehumbert, 2001), yet other models activate a set of exemplars within a certain area in the exemplar space (Hintzman, 1986; Goldinger, 1997; Wedel, 2006) and there is a group of models in which all the exemplars of a particular category, i.e. all exemplars within a particular exemplar cloud, are activated to either classify incoming stimuli or to serve as production targets (Wade et al., 2010; Walsh et al., 2010; Kruschke, 2011).

2.3 An Exemplar-Theoretic Model of Speech Production and Perception

In the following section a simple exemplar-theoretic model of speech production and perception will be outlined, which will be taken as a starting point for the experimental work presented here. The model is based on the Pierrehumbert (2001) model, however, to keep the model very general so that intonation can be incorporated, no segmental production bias is assumed. For the same reason, the selection process will not be specified, yet.

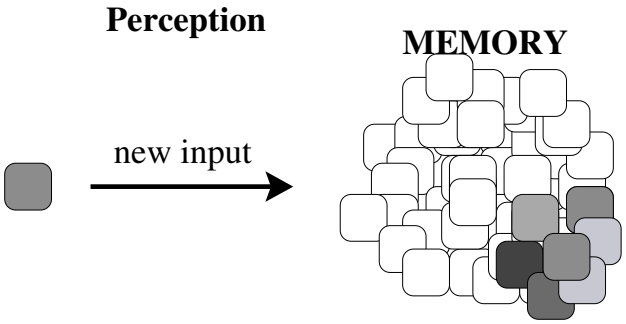
As in all exemplar-theoretic models, the general idea is that language input is stored in memory in the form of explicit episodes. These episodes are specified to a high degree, that is, at the time of storage, there is no abstraction mechanism. If a speaker of a language is exposed to new language input, this input is stored in memory as an exemplar including all its detail – a highly specified episode. Figure 2.2 displays the model graphically. The boxes represent exemplars, that is, concrete language input. The

input is taken to be of different lengths, hence, it can be seen as words, syllables or even phrases – depending on the frequency with which they occur. Like in the Multi-Level Exemplar Model (Walsh et al., 2010, described in section 2.2.3), larger units that occur frequently are assumed to be stored as one exemplar, whereas longer stretches of speech that occur together rarely are assumed to be constructed from their constituents. A specific value for a frequency threshold or a description of the decision component is not provided here, since it is not crucial for the experiments presented here (chapter 5). Similar exemplars are stored close to each other. Since exemplars are highly detailed, their properties open a multi-dimensional exemplar space in which the similarity and hence proximity of exemplars to other exemplars is determined. Thereby exemplars of the same kind form exemplar clouds – clusters of single instances that are stored close together since they are similar, e.g. the same word or phrase spoken by the same speaker.

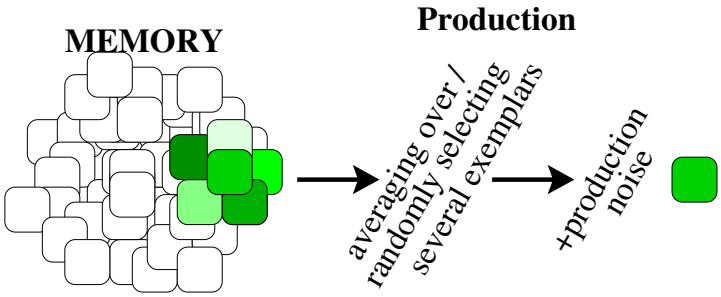
Note that no statement is made regarding the properties which span the multi-dimensional space – it could be auditory perceptual targets as proposed in Dogil and Möbius (2001) and A. Schweitzer (2010). It could also be acoustic properties as in the examples given in Pierrehumbert (2001) or articulatory gestures (e.g. Liberman and Mattingly, 1985). However, the experiments presented here, analyse acoustic properties and therefore, the exemplar properties will be taken to be acoustic properties, henceforth.

Figure 2.2a displays the perception process: when a new exemplar is perceived, it is compared to other exemplars that are already stored in memory. The similarity to these existing exemplars is determined according to proximity in the multi-dimensional exemplar space. Exemplars that are similar to the incoming one are indicated by similar shades of grey in the figure. The process of storing the new exemplar in the right location, hence within an exemplar cloud, is comparable to a classification process. If a value is important for later production, it will be stored with the exemplar.[17] Probably, some continuous values can be transferred to categorical values, e.g. the acoustic values that define the segments and thereby the word that was being said could be translated into a categorical label which is associated with semantic values etc. and can quickly be accessed later. Most exemplar models assume that labels can be assigned to exemplars for later use (e.g. Pierrehumbert, 2001; Goldinger, 1997). A. Schweitzer (2010) assumes that pitch accent labels are assigned to exemplars to mark the pitch accent type of an accented utterance. Some results of the experiments presented here indicate that this is not the case (see sections 6.2).

[17] Redundant information could potentially be stored, as well.



(a) New input is stored in memory. It is associated with similar productions with which it forms an exemplar cloud. This process of association can be regarded as a categorisation process.



(b) In production, a set of exemplars is activated. A production target is constructed form this set by either averaging over the set or randomly selecting one exemplar from the set and then adding production noise.

Figure 2.2: Basic exemplar-theoretic model of speech production and perception

In production, a set of exemplars is selected. Figure 2.2b displays this process. Exemplars are selected according to selection criteria that specify the characteristics of the intended new production. Hence, they are all relatively similar – which is indicated by different shades of green in the figure.

The model presented here assumes that criteria that are relevant in the selection process, for example the word identity, form exemplar clouds or subclouds and that selection criteria can be task-dependent. To illustrate this consider the following example: a speaker has the intention to produce the word “Berlin”. Probably every German speaker has perceived a relatively high number of instances of the word, hence, a high number of exemplars of this word in the respective speaker’s memory can be assumed. These instances probably come from different speakers, that is, they were spoken in different voices. If now, for the construction of a production target for the intended new production of “Berlin”, all instances of “Berlin” are selected (and only instances of

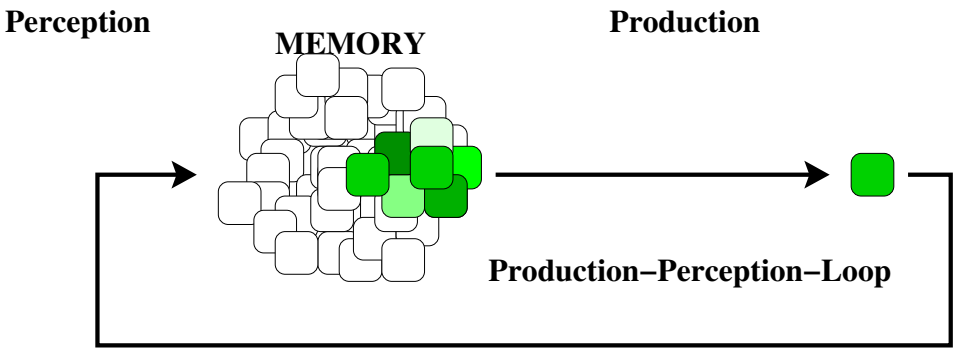


Figure 2.3: Production-perception-loop within the model.

"Berlin"; no exemplars where another word was spoken), one can conclude that firstly, *word identity* is a relevant selection criterion, and secondly *speaker identity* is not.

Note that there might be cases in which subclouds, like the one formed by word identity and speaker identity, are relevant for selection. Consider another task, for example the same hypothetical speaker wanting to impersonate John F. Kennedy and his famous sentence "Ich bin ein Berliner". Here, different criteria might apply. Probably exemplars of the whole phrase are activated. That is, the word "Berlin" or "Berliner" might have many more instances in the speaker's memory but they are not selected since they occur in a different context. Hence, from the exemplar cloud formed by all instances of "Berlin", only those cases where the respective acoustic signal occurred within the phrase "Ich bin ein Berliner", are activated. Moreover, speaker identity might be a relevant criterion since only instances spoken by John F. Kennedy are activated. Hence, the subcloud of instances of "Ich bin ein Berliner" where the speaker was John F. Kennedy, would be selected. Then, for the task of impersonating JFK, *speaker identity* would be a selection criterion. Of course, one could also assume that instances, where other speakers tried to impersonate JFK are also selected. This would imply that a categorical label for "JFK" is assigned to these exemplars and then this label would be a selection criterion for the given task of impersonation.

As for the question which criteria are used to select exemplars in pitch accent production, see section 6.2 which discusses the implications of the experiments presented here on selection.

From the set of selected exemplars, a production target is constructed – either by averaging over the set or by randomly selecting one exemplar as a production target. Production noise is added since it is assumed to be impossible to produce a target with perfect accuracy.

The new production is then of course perceived by the speaker and therefore stored in memory. Consequently, self-produced productions are included in memory and can contribute to the construction of production targets. This exemplar-theoretic production-perception loop is visualised in figure 2.3. As indicated by the colours in the figure, the new production is similar to the exemplars that contributed to the construction of the production target. Thus, certain properties of exemplars in memory are taken over by the new exemplar. Therefore, production biases can arise. If, for instance, a speaker comes from Berlin, where "Berliner" is pronounced as [bE6'li:na][18] instead of [bE6'li:n6], many exemplars will have this property. Therefore, these productions will influence the speaker's new production, which will also be stored. Hence, slowly more and more exemplars will have the final low a-vowel instead of the final central schwa-vowel. See section 5.2.6.1 for results on production biases in pitch accent production that can be explained by the production-perception loop outlined here.

The current chapter presented an overview of Exemplar Theory. First, motivation for exemplar-based accounts of speech perception and production was discussed. Variation in phonetic detail and evidence for storage of phonetic detail demonstrate that the mental representation of linguistic units is detailed and that the degree of abstraction assumed in generative accounts of speech perception seems to be to high. The voluminous body of evidence for frequency of occurrence effects on speech parameters also speaks for exemplar-based models, where frequency effects can be explained naturally, since single instances are stored and therefore the frequency of any unit is encoded in the model. Purely generative accounts on the other hand have no means to efficiently deal with frequency effects - for instance, some models would have to be enriched with counters and additional rules to explain frequency-dependent variation in speech data. Shortlist B is a modern abstract model which is able to account for effects of word frequency (Norris and McQueen, 2008). Note that even there, the underlying assumptions are very "exemplar-theoretic": Norris and McQueen (2008) assume that a listener is able to compute probabilities for abstract categories and that this information is updated throughout the listener's lifetime. Moreover, word frequency is input for the model. Consequently, the model needs knowledge about distributional properties of language to account for frequency effects.

Then, exemplar models of perception and production were described. These models assume that linguistic units are stored in memory as exemplars and that similar exem-

[18] transcriptions are in German SAMPA (Wells, 1997)

plars are grouped together. Moreover, stored exemplars are assumed to be employed in both perception and production: exemplars are chosen as reference to either classify new incoming instances or to construct a production target for an intended new production. The models differ in how the respective reference exemplars are selected – therefore the selection process in different exemplar models has been examined.

Finally, an example exemplar model was described which will be referred to when the experiments presented here (chapter 5) are discussed (chapter 6).

The next chapters throw light on models of intonation (chapter 3) and on the interface of Exemplar Theory and intonation (chapter 4), before the corpus-phonetic experiments which address potential frequency of occurrence effects on pitch accents are described and discussed.

3 Intonation

The following chapter gives an overview of several models of intonation. First, *phonological*[1] models of intonation will be described. These models are the most widespread ones. The focus will be on the most influential model, the approach presented in the dissertation of Pierrehumbert (1980) and the taxonomy that describes German intonation following her approach (Mayer, 1995). Additionally, two models that are subsumed here under the notion of *phonetic* models of intonation will be outlined. They mainly deal with the question of how the F_0-contour of a given utterance can be approximated most closely. The first approach, based on work by Fujisaki (e.g. 1988) is in a way a counterweight to Pierrehumbert (1980): it is also well-accepted and often used in speech technology (e.g. Möbius et al., 1993). However, it differs crucially from Pierrehumbert's approach in the way the tonal contour is described. While Pierrehumbert assumes a linear sequence of tones, the Fujisaki-models assume that different components are superimposed to generate the F_0-contour. The second phonetic model of intonation presented in this chapter is the PaIntE model (Möhler and Conkie, 1998; Möhler, 2001), which parametrises intonation by means of linguistically meaningful parameters. While it is far less widespread than the other ones it has been used successfully in the F_0-generation for German speech synthesis (Festival, 2010) and it is the model used to analyse fine acoustic detail of tonal events in the experiments presented here.

Using the terms *phonetic* and *phonological* to describe the two types of models highlights the fundamental difference between the two approaches. The phonetic models aim to reflect acoustic detail (Fujisaki-models also have a straight-forward articulatory interpretation building on the physiology of fundamental frequency modulation), whereas the models following Pierrehumbert's approach are based on the linguistic description

[1] The term *phonological* for the respective models is not exactly appropriate (cf. Batliner and Möbius, 2005, p.27): Despite the fact that the overall goal of such systems is to find phonologically distinct categories, namely categories that distinguish meaning, there is no ultimate mapping from their tonal descriptions to semantic categories. The relationship is rather a more complex one, with sequences of categories (see Pierrehumbert and Hirschberg, 1990) being related to meaning shifts. However, for instance, the GToBI(S) model (Mayer, 1995) attempts to reduce the set of labels to only those ones that are indeed semantically distinct.

of intonation, abstracting away from acoustics or articulation. However, it is possible to linguistically interpret different components and aspects of Fujisaki-models (see e.g. Möbius, 1995) and the parameters used in the PaIntE-model are linguistically meaningful, as well.

Intonation models in general assume a level which mediates between the actual acoustics and the linguistic function by which they are influenced. Phonological models employ a layer of description that abstracts away from the detailed acoustics of the speech signal – just as on the segment level, where phonemes and distinctive features abstract away from the fine phonetic detail of individual phones. In doing so these models attempt to describe the fundamental frequency contour with a finite inventory of labels – as few as possible but as many as necessary.

Batliner and Möbius (2005) argue that – at least from an application-oriented viewpoint – the intermediate level is unfavourable, since it increases the number of mapping processes that have to be carried out in speech technology applications (between concrete phonetic in- or output and the intermediate level as well as between the intermediate level and the linguistic function, instead of a direct mapping between linguistic function and phonetic realisation). Therefore, Batliner and Möbius propose a more *functional* approach to intonation where no intermediate level is necessary, that is, where linguistic function is directly mapped to phonetic form. Their ideas and another functional model, PENTA (Xu, 2005), are presented in the last section of this chapter.

3.1 Phonological models of intonation

3.1.1 Pierrehumbert (1980) – autosegmental metrical phonology

Pierrehumbert (1980) introduces a model of English intonation where an abstract representation characterises the intonation of an utterance. This model is probably the most commonly accepted phonological intonation model and it was highly influential as it has formed the basis for many similar models for various languages. These models, also denoted *tone sequence models*, assume that the intonation of an utterance can be described with a discrete sequence of categorically different *tones*. There is a "grammar of allowable phrasal tunes" (Pierrehumbert, 1980, p.10) that defines which tone sequences are well-formed. Secondly, there is a component that represents the metrical structure of a given text with a metrical grid (Liberman, 1975; Liberman and Prince, 1977). The metrical grid describes the stress pattern of a text in terms of the stress pattern of the

words which in turn is described by the feet the word is composed of. Hence, it describes which syllables are stressed and unstressed and how they are hierarchically organised relative to each other. In doing so, it also marks the strongest accent in a phrase.

Tone sequence models are also referred to as *autosegmental-metrical* models of intonation (Ladd, 1996) because they have both a *metrical* and *autosegmental* aspect: they assign relative prominence to elements within the utterance following Liberman (1975), and they represent tones and tunes autonomously from text.

Besides the tonal and the metrical component, there is a third component specifying rules how to align the tonal contour with the metrical structure of the text and how to derive F_0-contours from the abstract representation.

Inventory: categorically different tones. The tones describe the pitch contour by means of two levels, low (L) and high (H) representing a region low or high in the speaker's register. That is, a H describes a high local target, a peak, and L indicates a valley in the contour. The inventory of tone labels comprises labels for pitch accents, for phrase accents and for boundary tones:

Pitch accents are bitonal or monotonal. Monotonal accents can either be L* or H*, where H and L indicate the F_0-target on the accented syllable. Bitonal accents are described by two tonal levels, the starred tone signals the level on the accented syllable, the unstarred one denotes pitch height on the syllables adjacent to the accented one. If the unstarred tone describes the level on the preceding syllable, it is called a *leading* tone, if it follows the accented syllable, it is called a *trailing* tone. The labels for bitonal accents are H*+L, H+L*, H*+H, L*+H and L+H*. Because the starred tone describes the tonal level on the accented syllable, the accent labels encode the alignment of the contour with the text. Accordingly, while for example H*+L and H+L* both describe a contour that falls from a high target to a low target, they sound different since in the first case the fall is on and after the accented syllable, whereas in the second case the fall is before the accented syllable.

Boundary tones occur at the edge of the tonal phrase. Phrase-finally, they are marked by either H% or L% indicating a high or a low phrase boundary, respectively. They can also occur phrase initially, then they are marked by %H or %L.

Phrase accents are responsible for the tonal movement between the nuclear (usually the last) pitch accent of the phrase and the boundary tone. Hence, together with the

boundary tone, they describe the tonal characteristics of the end of an intonation phrase. Phrase accents can be H- or L-, indicating a movement to a high or low region, respectively.

Grammar of well formed tunes. The model introduces a hierarchy of these three types of tones: each intonational phrase consists of at least one pitch accent (but potentially more) one phrase accent and one boundary tone, characterising the tonal movement at the end of the phrase. It has to be noted that in successors of the model, such as the modified version presented in Beckman and Pierrehumbert (1986) or the ToBI standard for English (**To**nes and **B**reak **I**ndices; Silverman et al., 1992) or the German ToBI models (Baumann et al., 2001; Mayer, 1995), phrase accents can also occur without a boundary tone - then they mark a more subtle phrase break: an *intermediate* phrase. Each intonation phrase then can consist of one or more intermediate phrases, each of them being labelled with a phrase accent. The well-formedness of tonal phrases can be formalised with a finite state grammar (cf. figure 3.1). Pierrehumbert (1980) states that in English there are no other rules restricting the well-formedness of tonal sequences, but later versions prohibit some sequences (e.g. Mayer, 1995).

Association rules control how the tonal layer is aligned with the text. Boundary tones occur at the end of phrases, making their alignment straightforward: they are aligned with the right edge of the intonation phrase. Phrase accents are found near the end of the word bearing the nuclear accent. Since Pierrehumbert assumes that their exact placement is not linguistically relevant, there is no further rule with respect to their alignment (Pierrehumbert, 1980, p.32). Pitch accents are assigned according to the metrical structure, following rules from metrical phonology such as the Nuclear Stress Rule[2] or the Relative Prominence Projection Rule[3]. Pierrehumbert's framework then requires that

> a) if a foot has a pitch accent, any foot of equal or stronger metrical strength in the phrase also has a pitch accent, except that
> b) there are no pitch accents after the nuclear stress of the phrase.
>
> (Pierrehumbert, 1980, p.37)

[2]Nuclear Stress Rule: The rightmost constituent is the most prominent one (Chomsky and Halle, 1968; Liberman and Prince, 1977, as cited in Pierrehumbert, 1980, p. 33).

[3]Relative Prominence Projection Rule: The strong element in strong constituents is stronger than the strong element in weak constituents (Liberman and Prince, 1977, as cited in Pierrehumbert, 1980, p. 36).

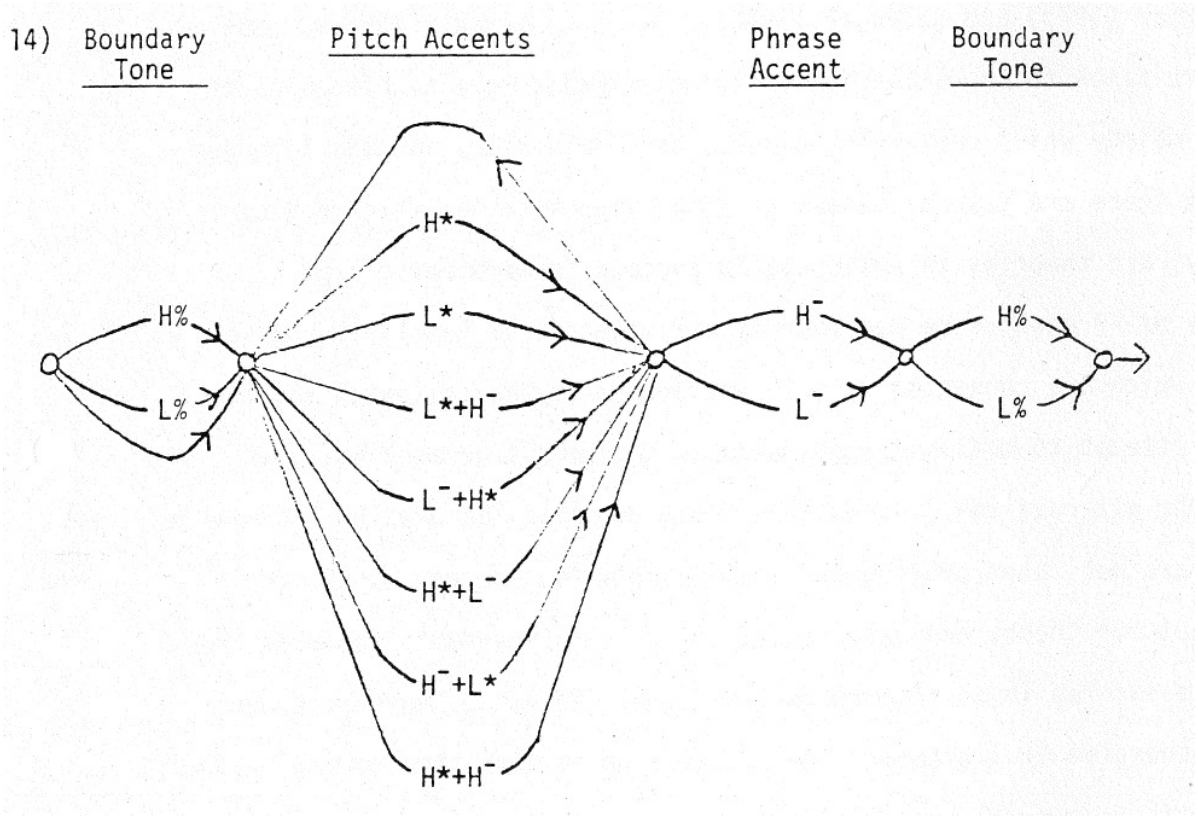


Figure 3.1: In Pierrehumbert's tone sequence model (Pierrehumbert, 1980) well formed sequences of tones can be formalised with a finite state grammar. Since Pierrehumbert assumes that the phrase accent and the unstarred tone in bitonal accents group together as a unit that is aligned with neither the accented syllable nor the phrase boundary, in this first version of the system, the hyphen marking phrase accents is also used in the labels for the bitonal accents, to mark the unstarred tone (H*+L-, H-+L*, etc).
(Figure from Pierrehumbert, 1980, p.29)

The abstract tonal representations created by the tonal and the metrical component are then converted into F_0-contours by two kinds of rules: first, the tones are evaluated phonetically[4] and then the F_0-contour between one target and its successor is constructed. The contour between the targets is interpolated as soon as the next target is available.

Local or global? The model works predominantly sequentially, so that unlike in superposition models (e.g. Fujisaki, 1988, cf. section 3.2.1), there is no phrase level defining the global trend of an utterance. In fact, it is explicitly claimed that intonation is exclusively determined on a local level (e.g. Pierrehumbert, 1980, p.2). However, there are aspects of the model that are non-local. For example, the metrical grid generates the global stress pattern, consequently hierarchically organised stress rules are applied. Moreover, to model declination the target values for pitch accents are lowered throughout the utterance (see footnote 4) – that is, declination is inherent in the model and not produced sequentially. Therefore the approach of Pierrehumbert (1980) has been classified as having elements of a superpositional approach, especially the version of Pierrehumbert and Beckman (1988) where tone scaling on the phrasal level is treated formally by modifying the pitch range parameters of the tones according to their position in the phrase. However, Pierrehumbert (2000) argues that there are crucial differences between the two types of models: firstly, she claims that the pitch range parameters are single numbers as opposed to time functions in superpositional models, and secondly she highlights that those parameters can be established for a whole dataset as opposed to having to be defined for each phrase separately. Moreover, she argues that they can affect different tones differentially – hence they are more flexible than superposition models where the functions being added always combine in the same way (Pierrehumbert, 2000, p.30).

3.1.2 Taxonomies based on Pierrehumbert (1980)

Pierrehumbert's approach was modified several times (Liberman and Pierrehumbert, 1984; Beckman and Pierrehumbert, 1986; Pierrehumbert and Beckman, 1988) and led to the formalisation of the *Tones and Break Indices (ToBI)* standard for describing the

[4] F_0-generation follows the assumption that each speaker has an individual baseline, which declines throughout the utterance and that F_0-peaks are scaled with reference to the baseline and can therefore be measured in peak-to-baseline distances divided by the baseline value at the location of the peak: $\hat{P} = \frac{P-B}{B}$ (see Pierrehumbert, 1980, p.49; and p.124 for experimental evidence). These values for the peaks, measured in "baseline units above the baseline" (p.49) are lowered by a factor k throughout the utterance to account for declination.

prosody of American English (Silverman et al., 1992). The inventory for labelling tonal events within ToBI is slightly modified compared to Pierrehumbert's (1980) approach. H*+L has been deleted from the inventory of pitch accents and a diacritic for *downstep* has been introduced: any high target which is considered to be lower than a preceding high target is marked with a leading "!" – it is downstepped.[5] Moreover, there is no initial low boundary tone %L, but only the high version %H. There are two levels of phrasing: intermediate phrases are marked by phrase accents, intonation phrases by boundary tones.

ToBI systems have been established for various languages. In German, there are different "varieties": GToBI (German ToBI, see Baumann et al., 2001) and GToBI(S) (German ToBI, Stuttgart System, see Mayer, 1995), which is the Stuttgart variety of the German ToBI model. The German data in the experiments presented here is labelled with the GToBI(S) standard, therefore it is described in greater detail below.

GToBI(S) The Stuttgart system reduces the original label inventory to a set of labels which are supposed to differ in their discourse meaning. In doing so, it integrates the ToBI systems with the approach from Féry (1993). The standard accents in this system are L*H and H*L as well as their monotonal variants H* and L*. The latter two are assumed to have their trailing tone realised later in the tonal phrase, just before the next accent is realised. This phenomenon is called *linking* and the trailing tone on the last syllable before the next accent is marked with either ..L or ..H. Moreover, the system provides additional labels for pitch accent types that deviate from Pierrehumbert's (1980) approach in two aspects: firstly, there are two tritonal accents, L*HL and HH*L. The former one is characterised by a rise on the accented syllable and a fall on the post-accented one, the latter one is realised with a peak on the pre-accented syllable, followed by a high target and a fall on the accented one. Secondly, there is an accent that makes use of a middle-range tonal level, marked by an M: H*M is characterised as having a peak on the accented syllable followed by a fall that only reaches the middle of the speaker's register. It has to be noted, though, that those three accent types occur very infrequently.

For phrase boundaries the model assumes that in German, the tonal contour at the end of the phrase is controlled by the nuclear accent and the edge tone. Therefore, phrase accents are merely used to mark rhythmical structure – they indicate intermediate

[5] Pierrehumbert (1980) also assumes downstep but regards it as being implicitly triggered by the left-hand context.

phrases. No tonal level is assigned to them, consequently there is only one label for phrase accents (-). Intonation phrase boundaries can be high (H%), low (L%) or tonally unspecified (%). In the unspecified (default) case, the contour between the nuclear accent and the phrase boundary is assumed to be an extrapolation of the tonal movement defined by the nuclear accent.

3.1.3 Phonological models and exemplar-theoretic effects

Generally the assumption of autosegmental-metrical theories of intonation is that the tonal contour of an utterance is added to the segmental realisation by a separate component and that the placement and the realisation of pitch accents is solely rule-based. The separation of the lexical and the tonal level in English is explicitly stated in Pierrehumbert (2000; p.20):

> ... pitch accents are not underlying properties of words. Instead, they are independent pragmatic morphemes which are co-produced with words.

Their being correlated with the lexical stress (derived from the metrical grid) is, according to Pierrehumbert, due to entrainment in motor control – the process of two independent organisms becoming synchronised. However, no evidence is given for that assumption. An approach to speech production that assumes the storage of explicit percepts in memory (cf. chapter 2.2) would not have to assume such a synchronisation process – pitch accents could be associated to words and particularly word sequences, just in the same fashion as lexical word accents. In fact, there would not even be the need for an explicit association mechanism since all information available about any incoming percept is part of the stored representation and thus available as targets for further productions or as reference in perception.

If, on the other hand, pitch-accenting is assumed to be post-lexical and solely rule-based, frequency of occurrence effects in intonation, as they are found in experiments 1-4 presented here, are harder to explain. The degree of abstraction assumed in autosegmental theories is problematic when it comes to explaining such a result. Since there the F_0-contour is calculated for each pitch accent in the same way (with regard to its position in the phrase), the system cannot deal with differences resulting from frequency of occurrence. Counting mechanisms and extra rules would have to be added (Pierrehumbert, 2001), making the system more complex and less intuitive.

Moreover, word- or word-sequence-specific prosody as found in experiment 4 (section 5.4) presents a challenge for such a system in that all information that cannot be retrieved

from the metrical grid (i.e. all information that goes beyond weak/strong stress, such as tonal movements or boundary placement) cannot be learned for words or word sequences individually.

Consequently, to account for the findings presented in chapter 5, the phonetic realisation rules would have to be made more detailed and more variable, potentially taking into account frequency of occurrence or word-specific random variation of certain parameters.

3.2 Phonetic models of intonation

3.2.1 Modelling the physiology of intonation

Superposition models, most of them being based on the probably best-known example, the Fujisaki-model 1988, regard the F_0-contour as being composed of components operating locally and of components operating globally. The local components are related to F_0-movements on accented syllables, while the global components are related to F_0-trends resulting from larger linguistic units such as phrases or sentences. The two components are considered to be independent of each other (Möbius, 1995). They are additively superimposed and added to a base F_0-value to approximate the F_0-contour. The two components are based on the physiology of the human larynx, with each component representing a different way of how fundamental frequency can be varied: rotation of the thyroid corresponds to accent commands, and back-forward movement of the thyroid is modelled by the phrase commands (Fujisaki, 1988; Möbius, 1995).

Figure 3.2 displays the two components and how they are added. Each component is modelled by a function term, the parameters of which are fitted to match the actual F_0-contour best. The upper left function illustrates the phrase component. The phrase commands are realised as impulse commands with varying amplitude (the amplitude is formalised as parameter *Ap* for each command in the function term, cf. the figure). The temporal alignment of the phrase commands is captured by one parameter defining the timing of each command. The phrase commands can be chosen so that the approximated F_0-contour initially rises and then falls towards the end (cf. figure 3.2). Linguistically, the phrase commands mark sentence mode, tonal phrasing (which is related to the syntactic phrasing of the utterance) and declination. To model final lowering in German, for instance, a negative amplitude can be chosen for the last phrase command (Möbius, 1995). Analogously, positive high values can be used to model a final rise in questions.

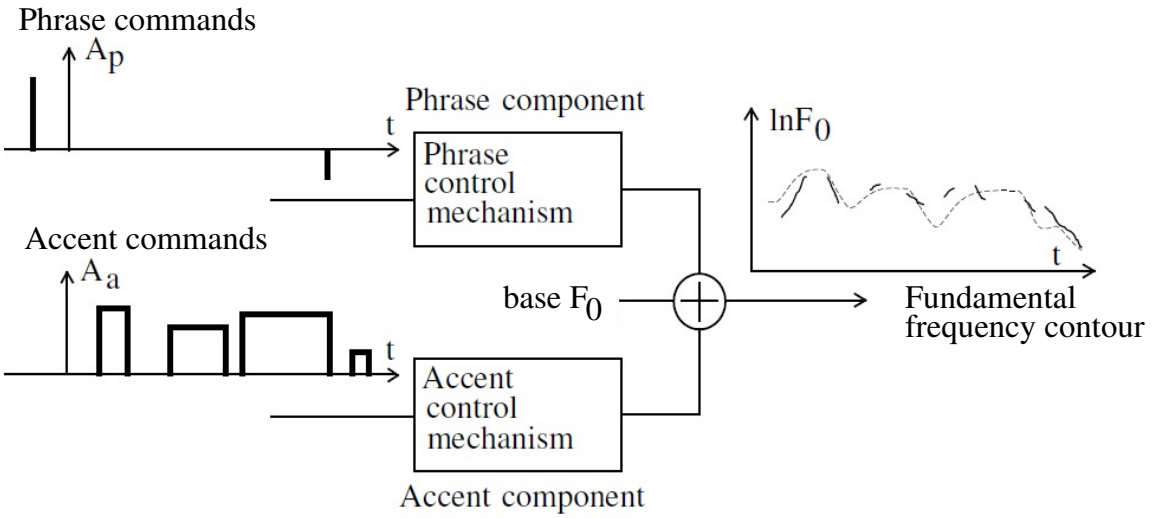


Figure 3.2: In the model of Fujisaki, three components are additively superimposed to model the F_0-contour. (Figure adapted from Möbius, 1995, p.110)

The accent commands are added to the contour generated by the phrase command (and the base (logarithmic) F_0-value). Accent commands occur at accented syllables (or morae) and can be concatenated to form plateau-like patterns (Fujisaki and Ohno, 1995, cf. figure 3.2). The accent component is displayed in the lower left of figure 3.2. As can be seen, accent commands vary in their amplitude *Aa* as well as in their duration. Hence, there are two parameters defining temporal aspects of the accent commands: one for the onset and one for the offset of each command. Linguistically the accent component corresponds to pitch accents in Germanic languages. Möbius (1995) in his implementation of the model for German intonation uses the concept of *accent groups*: accented syllables and the following unaccented syllables are part of the same accent group which is sensitive to syntactic boundaries (but not word boundaries).

The estimation of the parameters that fit the function can be automatic or manual. Möbius (1995) uses an automatic method for German. The estimated parameters are then analysed linguistically and are the basis for rules that adjust the function's parameters for F_0-synthesis. See Möbius (1995) for the full description of the parameter values and their linguistic interpretation.

The tonal patterns of various – sometimes crucially different (see Fujisaki and Ohno, 1995) – languages have been modelled successfully with superpositional models, among them tone languages as well as pitch accent languages (Fujisaki et al., 1998, and many more).

3.2.2 Modelling the acoustic detail of intonation

The PaIntE-model (Möhler and Conkie, 1998; Möhler, 2001) is also a data-based approach to intonation modelling, originally implemented for F_0-generation in speech synthesis.

The model approximates stretches of F_0 and interpolates between them. PaIntE is short for "Parametric Representation of Intonation Events" (Möhler, 2001). The name indicates the fashion in which the approximation works: it is implemented as a (linguistically motivated) mathematical function with 6 free parameters defining the tonal contour within an analysis window spanning the syllable marked with a tonal event and potentially the neighbouring syllables. These 6 parameters are linguistically meaningful (see below). The PaIntE-function is a function $f(x)$ of time which is composed of two sigmoids. The sigmoids are subtracted from a basic value giving the function's maximum value within the analysis window. Thereby the upper bound for the function is defined. The two sigmoids are defined each by 3 free parameters (a,b, and c, where a and c are sigmoid-specific and hence indexed according to their belonging to the first or the second sigmoid e.g. as $a1$ or $a2$, respectively) and a constant alignment parameter γ. The function term is given in equation (3.1).

$$f(x) = d - \frac{c1}{1 + exp(-a1(b - x) + \gamma)} - \frac{c2}{1 + exp(-a2(x - b) + \gamma)} \tag{3.1}$$

The approximation takes place over a winow of varying size, depending on the prosodic context. For instance, the window never extends prosodic boundaries (A. Schweitzer, 2010). If no prosodic boundary interferes, the window spans either 2 or 3 syllables: in the original PaIntE-model (Möhler and Conkie, 1998; Möhler, 2001) three pitch accent types are defined as "early accents": the GToBI accents H+!H*and L+H* (Grice and Benzmüller, 1995) and the GToBI(S) (cf. Mayer, 1995, and section 3.1) accent HH*L. For these accents the analysis window starts at the preaccented syllable, i.e. it spans 3 syllables. Note that A. Schweitzer (2010) modified PaIntE with respect to the length of the analysis window and that this augmented PaIntE-model (see section 3.2.2.2) was used in the experiments presented here.

The analysis window is normalised for syllable length. Thereby the syllable duration is mapped so that the beginning of the preaccented syllable is at time -1, the accented syllable starts at 0, and the post-accented syllable is mapped to the interval between 1 and 2.

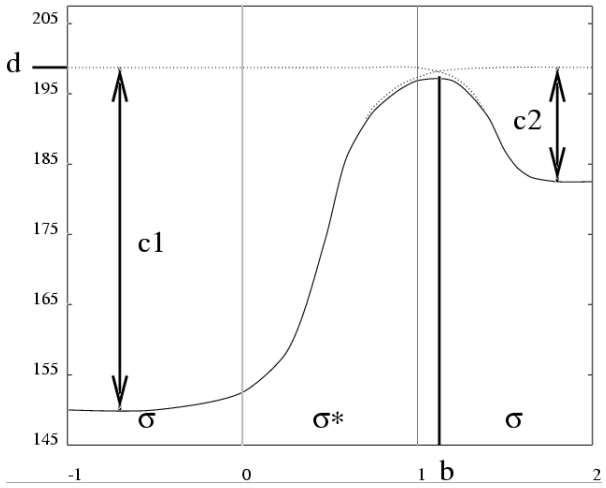


Figure 3.3: The PaIntE model function over a 3-syllable window, normalised for syllable length. The starred syllable $\sigma*$ is the one bearing the pitch accent. Four of the 6 free parameters of the PaIntE function given in equation (3.1) are marked in the figure: while $c1$ and $c2$ denote the amplitude of the accents rise and fall (in Hz), parameter d and b define the location (b) and height (d) of the peak. Parameters a1 and a2 are not displayed: they give the amplitude-normalised steepness of the rising and falling slope, respectively. (Figure from Möhler, 2001, p.2)

The function results either in a peak-like curve[6], as is the case for pitch accents, or it results in just a rising or a falling contour as is expected for boundary tones, and accents, the duration of which is longer than 3 syllables.

An example resulting contour is given in figure 3.3. Here, the function returns a curve that is shaped like a rising accent – like L*H in the GToBI(S) taxonomy (cf. section 3.1). The starred syllable is the accented one. The figure illustrates the linguistic interpretation of the parameters in equation (3.1): parameter b marks the alignment of the highest point in the curve, namely the alignment of the accent's peak, with the syllables. In the example contour, the high target is reached early in the post-accented syllable – therefore the hypothetical accent would probably have been labelled as L*H.[7] Parameter d determines the height of this peak, here it is between 195 and 200 Hz, thus the hypothetical accent would probably be spoken by a male speaker. Parameters $c1$ and $c2$ model the amplitude of the rising and the falling sigmoid, respectively. Linguistically, they give the range of the accent. Here, of course, for a rising accent (like the one in

[6]Note that PaIntE only models high targets. However, a modified PaIntE-version for Italian offers the possibility to model low targets, i.e. valleys in the contour (Cosi et al., 2002).

[7]However, the peak in L*H accents is typically at the end of the accented syllable (in word-final syllables) or later in the post-accented syllable (in word-internal syllables, see A. Schweitzer, 2010, p.74 for a detailed analysis).

the figure), the range of the rising slope is more significant, whereas for a falling accent (H*L), where the tonal movement on the accented syllable is the fall, the range of the falling sigmoid is more crucial. Parameters $a1$ and $a2$ estimate the steepness of the rise and the fall, respectively. They are normalised for the amplitude of the respective slope, i.e. the actual values of the gradients are divided by $c1$ or $c2$, respectively (Möhler, 2001).

3.2.2.1 Example parameter settings

Different accent types are expected to have different parameter settings. However, the two canonical accents L*H and H*L for instance can be described with very similar parameters, since they are basically mirror images of each other: An L*H accent is expected to have a pronounced rise on the accented syllable, followed by a peak in the post-accented syllable. After the peak, the fall is expected to be smaller than the rise caused by the F_0-contour slowly being interpolated to the next tonal event. H*L, on the other hand, would be expected to have a peak on the accented syllable, followed by a pronounced fall. The rise preceding the peak, however, is not expected to be as pronounced as the fall since the contour is interpolated between the preceding tonal event and the accent.

Figure 3.4 displays the values for a canonical L*H and a canonical H*L accent. The values were selected so that they depict the expected shape of an accent as described above. The gradient of the rise and fall were kept constant so that the influence of the other parameters can be seen. The values for L*H are $a1 = 10, a2 = 10, b = 1.5, c1 = 40, c2 = 10, d = 200$ for L*H and $a1 = 10, a2 = 10, b = 0.5, c1 = 10, c2 = 40, d = 200$ for H*L.

3.2.2.2 The augmented PaIntE model

As mentioned above, the original version of the PaIntE model (Möhler and Conkie, 1998) as well as its successor (Möhler, 2001) were implemented in such a way that the analysis window only spans 3 syllables if the analysed accent is one of H+!H*, L+H* or HH*L. A. Schweitzer (2010) modified the model so that the GToBI (Stuttgart) accents H* and H*L as well as their downstepped variants !H* and !H*L are analysed within a 3-syllable window, too. Moreover, she implemented a "long-window" option with which the analysis can be forced to always operate on a window that spans both the preaccented syllable and the accented one. In this case, the default analysis window size is 3 syllables, but the window is shortened to two syllables if the tonal phrase ends after the accented one, i.e. if a tonal boundary is realised on the accented syllable. For all the experiments

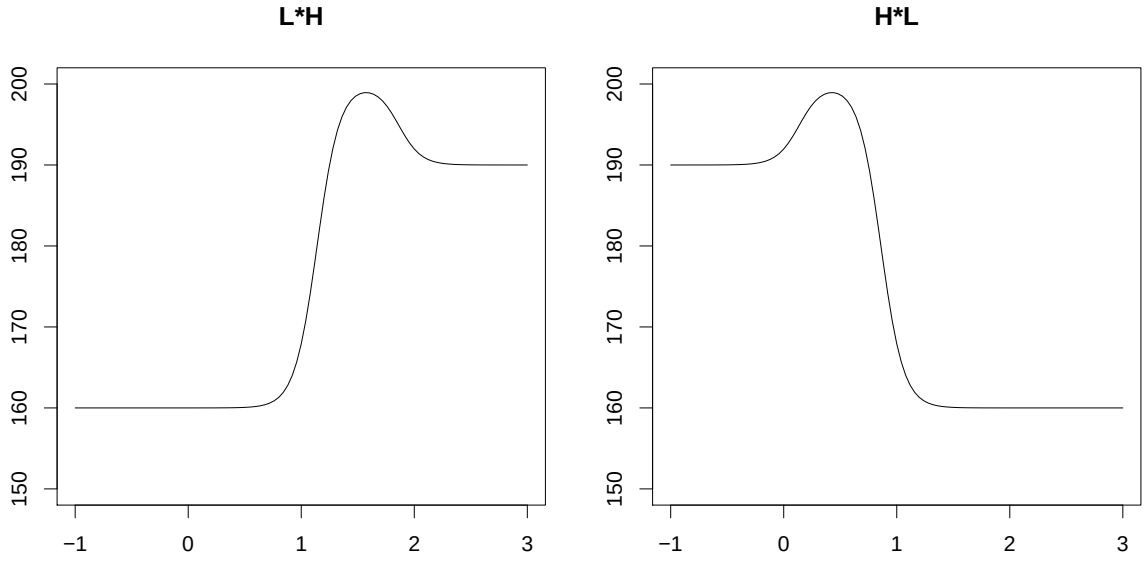


Figure 3.4: Canonical PaIntE functions for a falling L*H and a rising H*L accent. The vertical axes display the fundamental frequency values in Hertz, the horizontal axes mark the time, normalised with syllable length. The parameter settings are $a1 = 10, a2 = 10, b = 1.5, c1 = 40, c2 = 10, d = 200$ for L*H and $a1 = 10, a2 = 10, b = 0.5, c1 = 10, c2 = 40, d = 200$ for H*L.

presented here, the augmented PaIntE model was used. In experiments 1–3, where the pitch accent type was known, the approximation window size varied according to accent type as described above. For experiment 4, where only accent placement was labelled, the analysis was carried out with the long-window option; consequently, the analysis window was 3 syllables long if the accent was not followed by a phrase boundary.

3.2.2.3 Approximation variants and their consequences for parameter interpretation

There are three different ways in which the F_0-approximation takes place within the PaIntE model. They are illustrated in detail in A. Schweitzer (2010). For the work presented here, it is important to know that there is a standard case, referred to as the *two-sigmoid case*, henceforth. In the two-sigmoid case, the contour is approximated with two sigmoids that are composed according to equation (3.1). This case applies if the PaIntE algorithm finds a peak within the analysis window, thus, if it finds an F_0-maximum and two minima, one to the left and one to the right of the maximum that

are reasonably well apart from each other.[8] In this case all PaIntE parameters can be interpreted linguistically as explained above.

If the algorithm detects an F_0-maximum, but the minimum on either the left or the right of the maximum cannot be determined, then there is no peak within the analysis window, and the contour is either only a rise or only a fall. In this case, the approximation function uses only one sigmoid, i.e. only one of two complex terms in equation (3.1), to model the F_0-contour. This case is referred to as the *one-sigmoid case* hereafter. The sigmoid-specific parameters, that is, parameters $a1$ and $a2$, as well as $c1$ and $c2$, for the remaining sigmoid are determined according to the function term; for the omitted sigmoid they are set to $c = 0$ and $a = -1$. The setting of these parameters has some implications for their being meaningful: if an accent has for instance no rising part, it is reasonable to say that the range of the rise is 0, since there is no rise. However, the somewhat arbitrary value -1 for a is less meaningful, since the value cannot be interpreted on a numerical scale. A non-existent rise does not have a rising gradient of -1, but the value rather marks a categorical difference between accents that do actually have a rise, and the ones that do not. This has some consequences for the calculation of similarity in experiment 4, see section 5.4.4.

3.2.3 Phonetic models and exemplar-theoretic effects

The phonetic models presented here are application-oriented. No assumptions are made about whether intonation is organised separately from the segmental structure in human speech production. With the interplay of several continuous parameters defining the timing and the amplitude of tonal movements there is less loss of acoustic detail due to abstraction as is the case in autosegmental-metrical models.

In principle, if one assumes storage of intonation, as is indicated by the experiments presented here (cf. sections 5.1–5.4), the values for the acoustic parameters of the phonetic models could be stored as part of the exemplar representation.[9] Then, the Fujisaki- or PaIntE-parameters would be the tonal dimensions of the multidimensional exemplar (cf. A. Schweitzer, 2010, for PaIntE).

[8] "Reasonably well" is defined as at least 5 frames within the analysed window – see A. Schweitzer (2010) for more detail.

[9] For example, the parameter values of the accent component in Fujisaki-models could be stored along with the segmental representation of words, and those of the phrase component could be stored along with longer phrases.

3.3 Functional models of intonation

The following sections outline two functional approaches to intonation. While Batliner and Möbius (2005) argue for a model where linguistic functions are mapped directly to acoustics, Xu (2005) assumes that the biophysics of the articulatory system restrict possible mappings and should therefore be considered when specifying a functional model of intonation.

3.3.1 Batliner and Möbius

As mentioned in the introduction to this chapter, Batliner et al. (2001) and Batliner and Möbius (2005) propose an approach towards a functional model of prosody, where no level mediates between linguistic function and phonetic surface. Traditional (and well accepted) intonation models like autosegmental metrical models (see Pierrehumbert, 1980) employ an intermediate level that abstracts away from phonetic detail.

Batliner and Möbius's angle is the perspective of speech technology. While in automatic speech recognition prosodic models have not played a major role, so far, in speech synthesis, different models, e.g. tone sequence models such as ToBI (Silverman et al., 1992) or superposition models such as the model of Fujisaki (1988) have been used. To map categorical descriptions onto acoustics, a number of different approaches has been applied, as well: e.g. acoustic stylisation with PaIntE (Möhler and Conkie, 1998) or the application of templates and alignment parameters as in Van Santen and Möbius (2000).

According to Batliner and Möbius (2005), for speech technology, the benefit of autosegmental models of intonation is controversial. For classification in automatic speech recognition, the intermediate level introduces classification errors. Generally, the more information made available for an automatic classifier, the better the classification (Siepmann, 2001) – and any abstraction entails reduction of detailed information. With respect to speech synthesis, Syrdal et al. (1998) provide evidence that generated intonation is perceived as more acceptable if ToBI accent *type* labels are provided instead of labels for accent *position*, however, the authors also argue for an alternative version of ToBI with only two different accents types and one boundary type. Batliner and Möbius (2005) highlight that the finding that ToBI pitch accent *type* labels improve intonation generation, does not necessarily indicate that ToBI is the most appropriate intonation model – it could simply demonstrate that a greater amount of variation in the data is beneficial for intonation generation.

Consequently, Batliner and Möbius argue that a functional approach to prosody would be desirable for both speech recognition and synthesis. In such an approach, linguistic functions could be directly mapped onto acoustic parameters.

The authors criticise phonological models of intonation with respect to the original ideas of the Prague school concerning phonology:

> The classical phonological concept of the Prague school has been abandoned in contemporary intonation models, namely that phonemes – be they segmental or suprasegmental – should only be assumed if these units make a difference in meaning. This functional point of view has given way to more formal criteria such as economy of description.
> (Batliner and Möbius, 2005, p.27)

However, it has to be noted, that the GToBI(S) model is an attempt to reduce the inventory of labels to only those ones that differentiate between discourse meanings.

Batliner and Möbius (2005) consider the granularity provided by ToBI models to be not abstract enough to really represent linguistic functions, nor detailed enough to really represent the natural variation in speech. To actually train models, that is, algorithms predicting or understanding prosody, a large database annotated with categories for the required linguistic functions should be constructed. Consequently, a functional model of prosody needs an inventory of linguistic functions that are relevant for prosody production and perception. Batliner and Möbius provide an example catalogue of such functions. Their proposed inventory includes labels for accent and phrase boundary positions, discourse structure, paralinguistic functions such as emotions, user states and speaker specifics, prosodically relevant segmental properties, such as phoneme, syllable or word boundaries (derived from the acoustic signal), lexical features of the word such as word accent position and part of speech tag, syntactic and semantic features, for instance sentence mood, syntactic boundaries as well as prosodically relevant acoustic parameters of all kinds (not only tonal ones), since these parameters co-occur and interact.

3.3.2 Xu (PENTA)

Xu (2005) argues that both approaches to intonation modelling, i.e. the direct mapping of acoustic correlates to linguistic functions and the mapping of linguistic functions to phonological structures, have weaknesses. He claims that communicative prosodic functions should be mainly defined with respect to their communicative meaning. With

respect to this, his main critique about the phonological models of intonation is – similar to Batliner and Möbius's criticism – that the phonological units

> [. . .] are considered to be eventually linked to communicative meanings. However, the links are formulated separately from the definition of the accents, and are not treated as part of the core of intonation.
> (Xu, 2005, p.221)

However, he also argues that it is not possible to map directly between communicative function and acoustics, because of several reasons. Firstly, the acoustic surface is variant and dependant on the biophysics of the articulatory system. For example, the speed with which the fundamental frequency in speech can be changed, is dependent on how quickly the tension in the vocal folds can be changed resulting in different velocities for rising and falling movements – simply due to articulatory restrictions.

Secondly, the target assignment of pitch-targets is language-specific and, moreover, often not a one-to-one mapping, for instance because of other factors of the context. However, in a model as proposed by Batliner and Möbius (2005), the context could be encoded – the information is there, the question of how much context is analysed only depends on how long the analysis window is.

The third reason, according to Xu (2005), why acoustics cannot be directly mapped onto communicative function is that several communicative functions can operate in parallel, influencing the F_0 contour in different ways.

Therefore, in his functional approach to speech melody (Xu, 2004, 2005), the *parallel encoding and target approximation (PENTA)* model, he incorporates *encoding schemes* that each encode a specific communicative function and its effect on a set of melodic parameters (pitch target, pitch range, strength and duration). These parameters are assigned categorical values (such as "high", "low", "long" or "short") which in turn control parameters for a target approximation component that simulates articulatory implementation of the pitch targets. This component is also referred to as the TA-model (Xu, 2004).

3.3.3 Functional models and exemplar-theoretic effects

From an exemplar theoretic perspective, a functional approach to intonation is straightforward. Exemplars can be indexed with categorical information (e.g. Pierrehumbert, 2001). Thus, for the communicative functions a label could be assigned to the exemplars. Since all kinds of detail, acoustic, kinetic or perceptual is assumed to be potentially stored

with the exemplar, the mapping from linguistic function to acoustic or articulatory correlates would implicitly be there. An assignment of pitch-accent labels, as assumed by A. Schweitzer (2010), would not be necessary. In fact, results from experiment 3 indicate that the tonal contour is stored with the exemplar and not an abstract label that describes the contour (e.g. a ToBI accent).

The linguistic functions would probably be selection criteria, that is, they would narrow down the set of exemplars which is taken into account for the construction of a production target. Experiment 2 and 3 indicate that prominence is a communicative function that is a criterion in the selection process.

Generally, all experiments presented in chapter 5 are compatible with a functional perspective on intonation. Therefore, it seems desirable to identify the linguistic functions that are relevant for intonation and incorporate them into an exemplar-theoretic model of speech production and perception.

Before turning to the description of the experiments, however, chapter 4 will given an overview of the literature that connects Exemplar Theory and intonation. This includes studies demonstrating exemplar-theoretic effects on prosodic parameters as well as work indicating that the lexical storage of intonation in Germanic languages is possible.

4 Exemplar Theory and Intonation

To date, no exemplar-theoretic model exists that incorporates intonation, nor is there an intonation model that works in a usage-based fashion.[1] Therefore, this chapter focuses firstly on exemplar-theoretic effects that have been found for *other* prosodic parameters, and secondly on evidence for *storage* of intonation. Both lines of research are described in the following.

4.1 Exemplar Theory and Prosody

Several of the studies described in detail in section 2.1.3.1 demonstrate frequency effects on the prosodic parameter duration: Losiewicz (1992) finds the duration of the past tense morpheme -ed in English to be shorter for frequent verbs, Jurafsky et al. (2001) confirm the hypothesis that words with a higher relative frequency and a higher probability given their context are shortened. Baker and Bradlow (2009) report that frequent words are shortened, as well as words that are more likely in the context because they have been mentioned before.

Apart from these effects, there is additional evidence for usage-based prosody production: A. Schweitzer and Möbius (2004) directly address the question of whether an exemplar-based account can explain frequency effects in predicting syllable durations, and their corpus data has been modelled with the *Multi-Level Exemplar Model* (Walsh et al., 2010, described in section 2.2.3). Vigário et al. (2006) posit that the acquisition of prosodic words in European Portuguese is frequency-driven. Both studies are described in the remainder of this section.

A. Schweitzer and Möbius (2004) look at exemplar-based production of prosody. They investigate syllable and segment durations in a speech corpus that was originally designed for unit selection speech synthesis. Following Levelt and Wheeldon (1994) they

[1] But see Hawkins and Smith (2001) for a theoretical position which argues for a model that allows for storage of rich phonetic detail with pitch information included in the mental representation.

assume that there is a mental syllabary in which syllables are represented as units, and, following exemplar-theoretic considerations, they expect that the frequency of occurrence of a syllable determines whether this syllable is stored as a unit that can be accessed in speech production or whether it has to be composed online from its underlying segments during production. They claim that consequently, for very infrequent syllables, there should be a strong relationship between their duration and the duration of the underlying segments, whereas this relationship should be weaker for very frequent syllables that are directly accessed as production targets (and that are therefore less dependent on their constituents).

To test this relationship, they fitted two linear regression models that predict the syllable duration from the mean duration of the involved segments. They use z-scores of durations to account for different standard deviations of segment durations, that is, for the different degrees of elasticity of different phonemes. Syllable frequency classification was based on syllable probabilities from multivariate clustering (Müller, 2002) and not derived from their corpus, since a unit selection corpus does not represent typical frequency distributions. The duration of frequent syllables is predicted less accurately than the duration of infrequent syllables, demonstrating a weaker relationship between frequent syllables and their constituents. A. Schweitzer and Möbius conclude that frequent syllables are not composed of their segments during production but can be accessed as a unit.

The Multi-Level Exemplar Model (Walsh et al., 2010) has successfully modelled this effect of syllable frequency (Walsh et al., 2007) and thereby demonstrated that the result can be explained if prosody-production is assumed to be exemplar-based.

Vigário et al. (2006) examine frequency distributions of prosodic words in European Portuguese. They compare the frequency patterns of child speech with those of adult speech and of child-directed speech. To acquire child and child-directed speech data, two children about 1 year old were recorded in three sessions, with two and one and a half month intervals in between the sessions. Adult speech was taken from a spontaneous speech corpus for European Portuguese. The frequency distributions for prosodic words were compared in terms of percentage of monosyllabic (here, they differentiated between CV syllables ending on an oral vowel and other syllables), disyllabic, trisyllabic and more complex prosodic words. The main difference in the frequency patterns between adult and child-directed speech lay in a much higher percentage of monosyllabic prosodic words ending on an oral vowel in child-directed speech, accompanied by

a smaller percentage of prosodic words that were trisyllabic or had even more syllables. Interestingly, child speech was significantly correlated with the adult speech but not with child-directed speech. So, even though the child-directed speech comprised less complex patterns than adult speech, the children's speech showed complex patterns and was therefore more similar to the frequency distribution of European Portuguese, in general. This result demonstrates that the acquisition of the prosodic word in European Portuguese is frequency-driven.[2] Therefore, it could be modelled naturally with a usage-based account of language acquisition.

4.2 Storage of Intonation

While there is a growing body of evidence that word prosody is stored in the mental lexicon, sentence-level intonation is usually assumed to be autonomous from the lexicon. However, there is evidence that indicates possible storage of intonation along with lexical items. In the following sections a brief overview of the storage of word prosody (section 4.2.1) is given. This is followed by studies that indicate that storage of sentence-level intonation is possible (section 4.2.2).

4.2.1 Lexicalisation of word prosody

Several studies demonstrate that word prosody is stored in the mental lexicon (see Sekiguchi, 2006, for an overview). For instance, Lindfield et al. (1999) conducted a spoken word recognition task for English where subjects hear only the onset of a word they are asked to identify. The onset time that was needed for the correct identification of a word decreased significantly when the prosodic information of the word (number of syllables and syllabic stress) was made available to the subjects, both in comparison to only hearing the onset and to hearing the word duration without prosodic information. Soto-Faraco et al. (2001) showed in a cross-modal priming study on Spanish that primes which prosodically match the target words facilitated the lexical decision made on the target. Mismatching primes on the other hand inhibited lexical decision. Cutler and Donselaar (2001) examined the processing of minimal stress pairs (words differing only in the position of the lexical stress) in Dutch. In lexical decision, the members of the minimal stress pairs did not prime the other member but only the identical word, i.e. the segmentally identical pairs were disambiguated suprasegmentally. In the same study,

[2]However, the authors emphasise that they assume that possible grammar effects will strengthen those frequency-based effect on prosodic word acquisition (Vigário et al., 2006, p.201).

word spotting of words embedded in nonsense strings turned out to be harder when the nonsense string itself formed the beginning of a competing word, but only if the competing word had the same suprasegmental structure. They conclude that Dutch listeners effectively exploit suprasegmental cues in recognising spoken words.

Taken together all these studies demonstrate that word prosody is part of the mental representation and is thus stored along with the segmental representation. The next section investigates research that indicates that there might be cases where sentential intonation is stored, too.

4.2.2 Lexicalisation of Intonation

There is little research from the exemplar-theoretic domain that is concerned with intonation. However, Goldinger (1997) reports a pilot study in which subjects in a shadowing experiment adapted their pitch to the pitch of the speakers who recorded the stimuli, which can be interpreted as evidence that fundamental frequency is an inherent part of word storage.

Looking specifically at the question whether tonal contours might be lexicalised, Calhoun and A. Schweitzer (forthcoming) carried out a corpus study and a follow-up perception experiment, which demonstrated that words often combine with the same tonal contour and that tonal contours can have different discourse meanings depending on the words they occur with. The findings indicate that the lexicalisation of tonal contours is possible, and even common.

Apart from this exemplar-theoretically oriented work, which specifically tackles the question of whether intonation can be stored, there are other studies with outcomes that indicate a possible storage of sentence prosody.

These studies are firstly situated in the domain of psycholinguistics, where several experiments demonstrate that the familiarity or frequency of prosodic parameters influence speech processing, perception and production (Braun et al., 2006; Braun and Johnson, 2011; Braun et al., 2011; Mandel et al., 1994; Van Lancker and Canter, 1981; Van Lancker et al., 1981). A second research area which provides evidence for lexicalised storage of intonation is the area of machine learning, where various studies showed that word identity helps in predicting the tonal features (Pan and McKeown, 1999; Pan and Hirschberg, 2000; Brenier et al., 2006; Nenkova et al., 2007), and where instance-based learning of prosody outperforms other types of learning (Marsi et al., 2003). All relevant results are discussed below individually. Section 4.2.2.1 presents the phonetic studies, i.e the production data from corpus and lab experiments (Calhoun and A. Schweitzer,

forthcoming; Braun et al., 2006), followed by perception studies (Braun et al., 2011; Braun and Johnson, 2011; Van Lancker et al., 1981; Van Lancker and Canter, 1981; Mandel et al., 1994). Section 4.2.2.2 is dedicated to machine learning studies.

4.2.2.1 Evidence from phonetics: corpus data, production and perception experiments

Calhoun and A. Schweitzer (forthcoming) identified and examined what they call *intonational collocations*, viz. combinations of lexical items (words or short phrases) and F_0-contours, which frequently occur together.

Following exemplar-theoretic considerations, they assume that utterances are stored with their intonation. They argue that consequently, if intonational collocations are found to be widespread, this indicates storage, i.e. the lexicalisation of the contour.

Furthermore, they expect the stored exemplars to have specific discourse meanings distinct from proposed compositional ones (e.g. Pierrehumbert and Hirschberg, 1990; Ladd, 1980). Compositional approaches to intonational meaning assume that tonal contours themselves have a specific meaning (which can be derived from the meaning of the individual tonal events). Calhoun and A. Schweitzer on the other hand expect the meaning of an intonational contour to be dependent on the lexical item it combines with.

Following Bybee and Eddington (2006) they argue that discourse meanings are expected to spread from frequent instances to less frequent ones by analogy.[3] That is, a frequent collocation (of an intonation contour and a word or a short phrase) that has a specific discourse meaning, would be expected to be able to spread the discourse meaning on to collocations of semantically related words or phrases with the same intonation contour, so that these collocations have the same discourse meaning.

For their corpus analysis, Calhoun and A. Schweitzer use a large corpus of conversational speech (Switchboard, Godfrey et al., 1992), which is not annotated for accent type, but only for accent placement. To identify similar intonation contours, they extracted PaIntE parameters which describe the shape of a F_0-movement (Möhler and Conkie, 1998, and chapter 3.2.2), and durational information for nuclear pitch accents. Then they used clustering techniques to group similar contours together. These clusters

[3] Bybee and Eddington (2006), in a corpus study and a follow-up perception experiment, examine the four Spanish verbs for "become" and their usage in combination with adjectives. Not all four versions of the verb occur with all types of adjectives. Speakers rate high frequency combinations of verb+adjective as most acceptable, and semantically related adjectives in the combination with the same verb as more acceptable than the combination of the same verb with a semantically unrelated adjective. Bybee and Eddington (2006) argue that the novel instances of verb+adjective are based on analogies to previous (more frequent) experiences.

represent typical intonation contours and can be regarded as an approximation of pitch accent types. However, there were more clusters than traditional ToBI accent types (Silverman et al., 1992), i.e. the analysis based on accent-clusters is more fine grained than it would be with ToBI-categories. Intonational collocations were identified by statistically testing the likelihood of co-occurrence of a word or a short phrase together with a particular accent-cluster.

Firstly, a quantitative analysis revealed that the proportion of intonational collocations in the corpus is relatively high: 34% of all tokens were found to be in an intonational collocation. Among the frequent lexical types, the proportion was even higher covering 76% of the types. That is, frequent lexemes tend to occur often with a particular tonal contour, which indicates possible storage of this contour together with the lexical item. Calhoun and A. Schweitzer argue that the high coverage of the data even suggests that this might be the norm.

Secondly, in a qualitative analysis, Calhoun and A. Schweitzer (forthcoming) identified specific discourse meanings for the collocations. For example, two different collocations of the word "sure"+accent-cluster were analysed as having different discourse meanings: while "sure" with one accent-cluster expresses agreement with and sympathy to the conversational partner, it was analysed as expressing acknowledgement without any implicature about the speakers opinion when it occurred with another accent-cluster. The authors claim that the discourse meanings for the collocations are much more specific than traditional accounts of the discourse meaning of pitch accents alone (e.g. Pierrehumbert and Hirschberg, 1990; Ladd, 1980) and argue that this is evidence for storage of the discourse meaning. For this assumption, it is important to note, that in Calhoun and A. Schweitzer's study, the accent-cluster alone is not sufficient to determine the meaning. The same accent-cluster on different lexemes can express different discourse meanings .

Moreover, they observe patterns that hint that the discourse meaning of a collocation might indeed be able to spread over to lexically related collocations with the same pitch contour. For instance, the discourse meaning of the "sure"-collocation expressing sympathy and agreement could result from an analogy to the more frequent collocation of "yes" with the same accent-cluster. That is, the semantic similarity of "sure" to the lexeme "yes" would be the reason why the discourse meaning of the collocation with "yes" (i.e. the discourse meaning *sympathy and agreement*) spreads to the collocation with "sure".

To corroborate this hypothesis they carried out a perception experiment (similar to the one carried out by Bybee and Eddington, 2006) in which subjects had to rate the ac-

ceptability of intonational collocations. The collocations varied in lexical frequency. For low frequency collocations, the authors tracked whether the collocation was semantically related to a high frequency collocation or not. Subjects rated high frequency collocations, and low frequency collocations related to high frequency ones, significantly better than low frequency collocations that were *unrelated* to a high frequency one. Contrary to Calhoun and A. Schweitzer's expectations, high frequency collocations, and low frequency collocations related to a high frequency one, were not rated differently. However, they point out that the difference in token frequency between these two cases is only 2 – this might be too little a difference to demonstrate such an effect. Collocations classified as having a low frequency might in fact be frequent enough to display the same effects as high frequency collocations.

Overall, the study described by Calhoun and A. Schweitzer can be summed up as follows: firstly, intonational collocations are widespread, indicating that storage of tonal contours with words or phrases might be common. Secondly, the same tonal contours can have different discourse meanings, depending on the words they occur with. And thirdly, their study provided some evidence that novel combinations of word and tonal contour can receive a specific discourse meaning because of analogy to a more frequent collocation. All these findings are in line with exemplar-theoretic predictions, while traditional autosegmental theories of intonation would have difficulties explaining them.

The experiments presented in chapter 5 corroborate Calhoun and A. Schweitzer's findings by demonstrating firstly, that the frequency of occurrence of combinations of pitch accents and words has an effect on the realisation of the accents, and secondly, that collocations of words are realised with less prosodic variation than unrelated word strings.

Braun et al. (2006) examined the production of intonation contours in a mimicry task. Ten English subjects were presented with 100 randomly generated intonation contours which they had to imitate. In subsequent sessions, they were presented with their own contours from the preceding session and had to imitate them again. Over the sessions, the contours became less variable (indicated by decreasing standard deviations) and converged towards a limited set of tonal contours that are known to be patterns of English. However, the contours did not immediately collapse to contours that correspond to tonal categories as predicted by phonological intonation models (section 3.1). Hence, the subjects remembered and reproduced fine detail of intonation, which argues for a richer mental representation of intonation (see Braun et al., 2006, p.4013).

This is compatible with an exemplar-theoretic idea of speech production and perception: exemplars are assumed to be highly detailed and consequently, the mental representation of pitch would be expected to be detailed.

Moreover, from an exemplar-theoretic angle the results could be interpreted as entrenchment of intonation: the contours become less variable and converge over time to frequent tonal patterns because previously perceived instances contribute to the construction of the production targets for new instances. Indeed, Braun et al. explicitly state that episodic storage is one possible explanation for their findings (Braun et al., 2006, p.4014).

Braun et al. (2011) describe 3 experiments demonstrating that unfamiliar intonation contours slow down lexical access. Dutch subjects performed a) a word monitoring task in which a visually presented target word had to be identified in an auditory stimulus, b) a cross-modal priming experiment in which they had to perform a lexical decision to test lexical access and c) another cross-modal priming experiment in which the semantic category of the target word had to be determined. Since semantic classification takes longer than lexical access, the last experiment aimed to test whether any effects on lexical access are robust and long-lasting. In all three tasks, subjects were presented with familiar and unfamiliar tonal contours.[4] On the segment level the unfamiliar stimuli were not changed, hence the stimuli only differed in intonation. The target words were controlled for part-of-speech tag and relative frequency, and the authors kept track of when the respective stimulus was presented in the experimental session to control for effects of either tiredness or habituation of the subjects throughout the task. In all three experiments the subjects were slower if they were presented with an unfamiliar contour.

One possible explanation for the finding is that intonation is directly involved when the lexical item is accessed. Such a finding is difficult to model in generative models that assume abstract mental representations (see Braun et al., 2011, p.367) and advocates a model that assumes that lexical items are represented with fine phonetic detail as exemplar models do. However, Braun et al. argue that the hindering effect of unfamiliar contours might also arise from difficulties in interpreting the contour which generally leads to a greater cognitive demand and therefore less capacity to perform the respective task.

[4] In the first experiment, the unfamiliar contours were re-synthesised, in experiments 2 and 3, all stimuli were produced by a female Dutch speaker.

Braun and Johnson (2011) investigated pitch processing as a function of language experience and linguistic function. Dutch and Chinese (Mandarin speakers) subjects performed a speeded ABX match to sample task: a stimulus X had to be compared to two reference stimuli A and B and classified in terms of its similarity to them. All stimuli were non-words that could differ segmentally or suprasegmentally. Three kinds of pitch contrasts were examined: non-linguistic contrasts, post-lexical contrasts and potential lexical ones. Post-lexical contrasts were contrasts that correspond to linguistically meaningful differences in pitch contours for the Dutch speakers: the contours were shapes conveying sentence mode (declarative sentence vs. echo question). Lexical contrasts were contrasts that could differentiate between lexical items for the Mandarin speakers. The results demonstrated that the subjects were attentive to pitch in all tasks. This was indicated by longer response times if the stimuli differed in pitch, that is, all subjects were slower at determining similarity between non-words when the pitch contour did not match. The response time was even longer if the mismatch corresponded to a contrast that was linguistically meaningful: Dutch subjects' reaction times were longer when the contrast was on a position where it could convey different sentence types, i.e. post-lexical information, than when the pitch contrast was linguistically non-meaningful. Mandarin subjects' response times, however, were even longer in the case of a mismatch that could potentially signal lexical contrast. Braun and Johnson conclude that "linguistic function guides how pitch information is processed" (p.592) indicating a high compatibility of the results to ideas of functional models of intonation (e.g. Batliner and Möbius, 2005, cf. section 3.3). In addition, they posit a scale of how relevant pitch information is (lexical>post-lexical>non-linguistic) which corresponds to how strongly pitch is taken into account when performing the task.

From an exemplar-theoretic viewpoint, it is important to note that mismatches in pitch increased reaction times in all tasks. This could indicate that pitch is part of what is being stored and is considered when exemplars are compared. The results could be explained in an exemplar-theoretic model: since the words were all non-words, no exact match can be assumed to be stored in memory before the task. If, during the task, a non-word is perceived and the target stimulus matches it exactly (on the segmental and on the suprasegmental level), the decision can be performed quicker than if there is no exact match. If the mismatch is on a dimension that has been shown to be extra important in existing exemplars, the decision is slowed down even more. This could be modelled with different attention weights for different dimensions (e.g. Nosofsky, 1986, cf. p.39 here).

Van Lancker and Canter (1981) examined the perception of English expressions that have both a figurative and a literal reading such as "The coast is clear" or "I hit the sack". Native English listeners were able to distinguish reliably between the two readings when the speaker was instructed to convey one or the other. Consequently, prosodic cues were sufficient for native English speakers to distinguish between the literal and the figurative meaning of English expressions. Van Lancker et al. (1981) investigated the prosodic cues that signalled the intended reading and found that pausing, fundamental frequency height and contour, and duration significantly differed for the two readings. As for the results on fundamental frequency, literal expressions were marked by a higher number of tonal events (rapid rises and falls in the F_0-contour) and a higher mean F_0. Van Lancker and Canter (1981) posit that the idiom is stored as a unit while the literal reading is constructed from its constituents. Consequently, the F_0-contours would have to be stored with the lexical expression to produce the described results. Then the result could be modelled in an exemplar-theoretic model where expressions are indexed with semantic values.

Mandel et al. (1994) investigated how well infants (2 months old) remember segmental properties of words. In two experiments they modified either one or two phonemes in a given test sentence. For instance, the sentence "The rat chased white mice." was modified to "The cat chased white mice." and "The cat raced white mice.". The experimental setup measured the sucking rate of the infants which indicates their awareness of a change. In the first experiment, two types of stimuli were created: firstly, stimuli that were read out as one prosodic unit, hence, the speaker read the sentence as given above. Secondly, the sentences were concatenated from single words, where the speaker read out lists which contained the target words. Three groups of infants were tested: a group that was presented with two phonetic changes, a group that was presented with one phonetic change and a control group that heard the same sentence. Significant differences in the sucking rate between the control group and the groups that were presented with phonetic changes were only found in the sentence condition. Consequently, when the target words were organised as a prosodic unit they were remembered better than when they were prosodically fragmented.

In the second experiment, Mandel et al. (1994) compared the memorability of stimuli where the information is organised in one prosodic unit and stimuli that were uttered as part of two prosodic phrases. For instance, the words "Cats like park benches" were read as one sentence that was realised in one prosodic phrase, for the one-phrase condition.

For the two-phrase condition, the words were read as part of two successive sentences, like "Brigid really knows what *cats like. Park benches* are their favourite things to climb on." Only in the one-phrase condition infants did discriminate between the phonetically different stimuli.

The authors argue that both experiments indicate that prosody acts as "a kind of "perceptual glue" which keeps spoken information together as a unit" (e.g. Mandel et al., 1994, p.175). From an exemplar-theoretic point of view, the finding that prosodic information influences the memorability of sequences of words is interesting since it might indicate that those word sequences are stored along with their prosody as exemplars.

4.2.2.2 Evidence from machine-learning

Marsi et al. (2003) compared two algorithms predicting prosody. Interestingly, one of the algorithms was a memory-based classifier. This classifier works in an exemplar-theoretic fashion: a set of training instances is treated as points in a multidimensional feature space and is stored without any abstraction in memory. New instances are classified by similarity to stored instances. For the determination of similarity, a distance measure is employed (k-nearest distance classifier, see Marsi et al., 2003, p.492). The other classifier is a classification and regression tree (CART) as implemented in Taylor et al. (1999). This algorithm operates in a highly distinct fashion: it aims to find the smallest possible decision tree, i.e. the most efficient way of generalising over the data; individual instances are not remembered. The memory-based account outperformed the CART in a number of tasks (predicting accent placement, predicting prosodic boundaries and a combination of the two).

It is important to note that the classification was based only on features that could be retrieved from text. These were lexical or syntactical features (e.g. word form, POS-tag, position in a noun or verb chunk), features that can be retrieved from orthography (orthographical indication of emphasis, punctuation, distance to sentence boundary), features capturing the information content of the word (e.g. information content of the word and the bigram, relevance of the word in the text), as well as a feature that can be seen a very coarse approximation of givenness: the distance to the previous occurrence of the word. The gold standard, i.e. the prosodic annotation performed by human annotators, was also carried out merely on text. That is, the annotators marked the read text for expected locations of pitch accents and boundary tones. They were provided with feedback in the form of synthesised speech to verify their assumptions (see Marsi et al., 2003, p.491).

From an exemplar-theoretic angle, this result can be seen as indicating storage of contexts. New instances are classified on the basis of similarity to previously experienced contexts. Similar contexts are grouped together and are activated together in production. Consequently, the new instance being produced is likely to have similar prosodic properties: if previously experienced contexts received a pitch accent or a phrase boundary, new contexts similar to the stored ones, are likely to do so, too.

Several other machine learning studies provide further evidence for the storage of intonation contours: Pan and McKeown (1999) used a classification-based rule-induction system (RIPPER, see Cohen, 1995) and hidden Markov models (Rabiner and Juang, 1986) to predict accent placement. Unigram probability (i.e. the probability of a word to occur) was a better predictor of accent placement than part-of-speech tag. That is, word identity reveals more information about the likelihood of being accented than word class does: generalising over individual words to POS-tags entails loss of information that is relevant for pitch accenting. This indicates an intrinsic connection between the word and its tonal contour.

In a follow-up study Pan and Hirschberg (2000) trained RIPPER models to predict whether a noun is accented or not. In order to investigate the influence of word context, they use different measures of word collocation, one of them bigram probability, as predictive features. Bigram probability measures the likelihood of a word to occur given the occurrence of the previous word. Compared to a classifier that uses unigram probability, the classifier using bigram probability is significantly better. This demonstrates that the lexical context of a word is predictive of the word's accent status, which again indicates a strong coupling between words and pitch contour.

Further studies on the prediction of accent placement employ a feature called *accent-ratio* (e.g. Brenier et al., 2006; Nenkova et al., 2007). Accent ratio captures the probability of a word being accented. Brenier et al. (2006) as well as Nenkova et al. (2007) train decision trees (in WEKA, cf. Witten and Frank, 2005) to predict accent placement in a subset (12 conversations) of the Switchboard Corpus (Godfrey et al., 1992). Besides accent ratio, they employ several predictors: complex linguistic features like information status or contrast, that have been manually annotated in the corpus (Calhoun et al., 2005) as well as shallow features that can be determined automatically, for instance part-of-speech tag and uni or bigram probability. Both Nenkova et al. (2007) and

Brenier et al. (2006) calculate accent ratio on the basis of a larger proportion of the Switchboard corpus.[5]

In feature-wise comparison, accent ratio is the single-most powerful predictor, in both studies. When a combination of features is used as predictors, all combinations that yield highest accuracy contain accent ratio.

These results demonstrate that the relative frequency with which a word was accented in a large corpus is more predictive of the accent status of a new instance of this word than any other features that have been reported to correlate with prominence (for instance, information status or part-of-speech tag). This indicates that the connection between a word and its intonation contour is stronger than autosegmental theories of intonation assume.

Furthermore, Brenier et al. (2006) demonstrate that it is beneficial to incorporate context information: the classifier that uses not only the features computed for the target word, but also the features of the adjacent words performs best. This can be seen as an indication of the storage of contexts. Similar contexts are then again likely to render similar intonation.

The studies described above indicate that exemplar-theoretic storage of intonation might be possible. The following chapter describes 4 corpus experiments that were specifically carried out to examine potential exemplar-theoretic effects on pitch accent realisation.

[5]In Nenkova et al. (2007) accent ratio is calculated on the basis of 60 conversations, in Brenier et al. (2006), 50 conversations were included

5 Corpus studies: frequency effects on pitch accent realisation

The core idea of Exemplar Theory is that language is stored in memory in rich detail (chapter 2.2). It therefore seems logical to assume that intonation is part of the information being stored, especially since word prosody can be stored in Germanic languages (see section 4.2.1) and since there is evidence that intonation can be lexicalised as well, i.e. stored along with the word information (see section 4.2.2). The idea that intonation is part of the episodic lexicon in which language input is stored conflicts with traditional autosegmental models of intonation (Pierrehumbert, 1980, cf. chapter 3), which claim that intonation in Germanic languages is post-lexical, that is, not stored in the lexicon, but assigned after retrieving the lexical item.

Assuming that intonation is stored and that the production of tonal events happens in an exemplar-theoretic fashion (Pierrehumbert, 2001), one would also expect that the contexts (e.g. prosodic, discourse, phonological context) in which an exemplar occurs, influence the production of a word or a phrase and its tonal structure. The idea behind this is that exemplars from similar contexts are associated with each other. When a new instance is produced, associated exemplars are expected to be employed in the selection of a production target (see section 2.2.3). These exemplars would thus influence the production of the intended instance. Exemplar Theory predicts that differences in the number of exemplars that are considered in the selection of a production target can give rise to frequency of occurrence effects (section 2.1.3). Consequently, if intonation is indeed part of the exemplar representation, frequency effects of different contexts are expected for intonation contours, as well.

This prediction is explicitly targeted in the experiments presented in sections 5.1–5.4. All experiments examine frequency effects on pitch accent realisation. Experiment 1 (Schweitzer et al., 2009b,a) demonstrates that pitch accent realisation in general is sensitive to the frequency of discourse and prosodic context, irrespective of the word on which the accent is realised. Experiment 2 (Schweitzer et al., 2010b) then presents a

more fine-grained analysis that shows how pitch accent realisation is influenced by the frequency of the word+accent pair (the combination of a word and the accent it occurs with). Experiment 3 (Schweitzer et al., 2010a) then looks at the influence of competitors to such pairs in production. It shows that the proportion of competitors to a word+accent pair significantly influences the production of the respective pitch accent tokens. Finally, experiment 4 (Schweitzer et al., 2011) investigates the relationship between the relative frequency of word sequences and their prosodic variability, demonstrating that word sequences that occur together relatively often display less prosodic variability.

Table 5.1 gives an overview of the experiments. As described above, in all experiments, a tonal parameter is related to a frequency measure to examine frequency of occurrence effects on intonation. To get a comprehensive picture of the nature and validity of frequency of occurrence effects at various levels, the experiments vary in several aspects: The unit for which the frequency is computed varies, as does the way the frequency measure is calculated. Moreover, two different Germanic languages, American English and German as well as different speech genres were examined.

Exp	Unit	frequency measure	tonal parameter	language	speech genre
1	PA	absolute	PA shape variability	German	radio
2	PA + word	absolute	PA shape	German	radio
3	PA + word	relative	PA shape variability	AE	radio
4	word + context	relative	PA shape variability prosodic context variability	AE	spontaneous

Table 5.1: Overview of the experiments. In all experiments, a tonal parameter is related to a frequency measure to examine frequency of occurrence effects on intonation. PA stands for pitch accent.

5.1 Experiment 1: Absolute Frequency of Pitch Accents and Information Status Categories

This experiment seeks to investigate whether frequency of occurrence effects (as described in section 2.1.3) can be found in intonation. More specifically, it examines whether or not the frequency of a tonal event, or the frequency of its broader linguistic context, influence the production of that event. If this is the case and tonal events are indeed subject to frequency of occurrence effects, this gives rise to the question

how these effects can be explained. Autosegmental-metrical phonology (chap. 3) is silent on frequency effects. Exemplar Theory (chap. 2), on the other hand, provides a way of explaining them by using the following assumptions:

1. Language emerges from the storage of single exemplars
2. These exemplars are rich in phonetic detail.
3. Stored percepts are employed as production targets.

Following these assumptions one can expect that fundamental frequency can be part of the information being stored. If this is the case, frequency of occurrence effects can be explained naturally, since different numbers of exemplars lead to different numbers of potential production targets and can therefore account for frequency effects.

The study presented here compares the tonal realisation of different pitch accent types which vary in frequency of occurrence, in a radio news database (Rapp, 1998).

To take the broader linguistic context of pitch accent tokens into account, which might also influence intonation, the information status of the tokens is analysed, as well.

Information status is one aspect of information structure and classifies expressions according to whether they are already given in the discourse, or *accessible* in some sense, or whether they are new to the discourse. Hence, it describes properties of expressions that reach over phrase and sentence boundaries. Information status has been shown to influence tonal realisation (e.g. Pierrehumbert and Hirschberg, 1990; Baumann and Grice, 2006), and is thus expected to influence intonation in the analysed radio news data, as well. To verify this, a pre-study will test if information status has an influence on intonation in this data. If this is the case, the frequency of occurrence of information status categories will be incorporated into the analysis of pitch accent variability as a representation of frequency effects of a broader linguistic context – the discourse context.

Before describing the pre-study, the information status annotation scheme and the data will be outlined in sections 5.1.1 and 5.1.2.

5.1.1 Information Status Annotation

The data used in the pre-study and in experiment 1 (see section 5.1.2) has been manually labelled with respect to the information status of noun phrases. Information status describes whether expressions have already occurred in a discourse or whether they are new. Hence, determining the information status of noun phrases in texts involves classifying in which way they are *given* or *new*. The term *information status* replaces

and specifies more clearly the often vaguely used term *givenness*. Different classification systems have been proposed in the literature (see Riester, 2008a, for a comparison of several information status labelling schemes).

The annotation of information status in the IMS Radio News corpus is based on a semantically motivated taxonomy descending from presupposition theory (see Riester, 2008b, for more detail). The main theoretic assumption is that information status categories (for definite descriptions) should group expressions according to the *contextual resources* in which their antecedents are found.[1] For example, an expression like "next Sunday" can be resolved in the situative context, while expressions like "Tom Waits" are resolved in the encyclopaedic context – identifying the referent is general knowledge. Based on differences like these, Riester (2008a) assumes several information status categories. Those that are relevant for the data used here, are described below. Since the information status categories descend from semantic principles which fundamentally differ from each other, they will be grouped here according to the underlying semantic process. These processes are a) anaphora resolution, b) accommodation and c) the introduction of a new referent; they will be described, along with the labels they are subsuming, in the following.

Anaphora Resolution Anaphoric expressions refer to another expression. This referent (the antecedent) is found either in the discourse context (i.e. it has been mentioned in the current discourse) or in the situative context (i.e. it is part of the speaker environment) or it is an entity that the recipient knows, which can therefore be resolved in the encyclopaedic context.

The labels in this group are:

d-given-*: The group of *d-given* (i.e. discourse-given) categories is defined on the basis of co-reference. Expressions that refer to entities that are given in the discourse context fall under that criterion. The scheme offers a more fine-grained distinction between different types of d-givenness: *d-given-repeated* for exact repetitions, *d-given-epithet* for expressions that use new lexical material to refer to an entity that has already been mentioned in the discourse, *d-given-pronoun* for pronominals and *d-given-short* for short forms of expressions that have already been mentioned. Different classes of *d-givenness* can be found in the following example:

[1] This stands in contrast to several other annotation schemes following Chafe (1994) in which the categories are mainly grouped according to cognitive activation.

(5.1) [Tom Waits]$_{\text{ACCESSIBLE-GENERAL}}$ *(first mention)* ...
[Tom Waits]$_{\text{D-GIV-REPEATED}}$ *(second mention)* ...
[Waits]$_{\text{D-GIV-SHORT}}$ ... [he]$_{\text{D-GIV-PRONOUN}}$ ...
[the American singer-songwriter]$_{\text{D-GIV-EPITHET}}$

accessible-general: Expressions that are not present in the previous discourse but refer to entities known to the intended recipient and can thus be resolved in the encyclopaedic (also: knowledge) context (cf. example (5.1), first expression).

situative: Expressions referring to antecedents in the situative context (i.e. the speaker environment). Typically these are the discourse participants, demonstratives referring to objects in the speaker situation or deictic expressions, as below:

(5.2) this setback needs to be counteracted [now]$_{\text{SITUATIVE}}$

Accommodation Accommodation occurs when a recipient assumes that a presupposition is true. Riester (2008b) links the classical presupposition theory to information structure (cf. Riester, 2008b, p.87). He points out that definite descriptions can refer to an antecedent that neither has been mentioned in the discourse nor is generally known by the recipient. Being confronted with such an expression, the recipient adapts his or her discourse environment so that the referring definite description can be resolved, i.e. the expression is accommodated. This is an issue that has not been addressed in earlier annotation schemes, in which these expressions would have most likely been subsumed under the concept of *general accessibility*. Riester (2008b) however provides an extra information status label for definite descriptions that have to be accommodated:

accessible-description Expressions that are not resolvable in any of the contexts (they have not been mentioned, nor does the audience know or perceive them) and hence, have to be accommodated, as in example (5.3).

(5.3) [the leadership crisis lasting for months among the Hamburg Social Democrats]$_{\text{ACC-DESC}}$

Introduction of new discourse referents When an expression is new to the recipient and adds information to his or her discourse context, a new discourse referent is introduced. The label *new* marks such cases:

new Indefinite expressions that are unrelated to any of the contexts as in the following example:

(5.4) the peace talks have been continued in spite of [a deep crisis]$_{\text{NEW}}$

By classifying expressions with the help of the semantic processes outlined above and especially with the contextual resources that are used in resolving referring expressions, the annotation scheme of Riester (2008b) defines information status categories strictly by semantic means. The main differences to other information status taxonomies are the existence of the category *accessible-description*, the definition of givenness based on co-reference (in contrast to literal previous mention, as it is defined e.g. in Baumann (2006; following Chafe, 1994 and Lambrecht, 1994), the fine-grained distinction into several different classes of givenness, and a relatively tight definition of novelty, which, unlike other taxonomies, does not include definites.

Another characteristic of the scheme is that Riester (2008b) assumes that textual information alone (i.e. no prosodic or speech related information) is not always sufficient to fully determine the information status associated with a particular phrase. Hence, there are cases where phrases have multiple annotations, reflecting textual underspecification of information status.

5.1.2 Data: The IMS Radio News Corpus

The pre-study and experiment 1 analyse speech data from the IMS Radio News Corpus (Rapp, 1998). This corpus was automatically segmented and manually labelled according to GToBI(S) (cf. Mayer, 1995, and chapter 3) and consists of approximately 1 hour of speech. It contains data from three speakers, two female and one male, with the male speaker having produced 72.9% of the data.

The orthographic transcript of the speech signals was manually labelled with respect to information status according to Riester's (2008b) scheme (see above) by two trained independent annotators and subsequently compared by a third trained annotator who – in cases of disagreement – made a casting decision.

As mentioned above, the annotation scheme is based solely on the written text and allows for multiple information status labels on the same noun constituent. To draw clear-cut conclusions on pitch accent preference for the different information status categories in the pre-study as well as to analyse the influence of the information structural context on the tonal realisation in the main study, a one-to-one relationship between information status category and tonal properties is required. Therefore, only a subset of the corpus is used in both of the following studies: it includes only expressions where the information status category was unambiguous.

5.1.3 Pre-study: Testing the influence of information status on intonation in the IMS Radio News Corpus

To test whether information status has any influence on tonal events in the IMS Radio News corpus, a pre-study was carried out. Specifically, the study tests whether there is an influence of information status on the choice of pitch accent. If this is the case, information status categories will be regarded as being indicative of how broader linguistic context can influence tonal events.

The description of the study is structured as follows: section 5.1.3.1 presents literature on the relationship between information status and pitch accents (commonly regarded as the main correlate of information status, e.g. Halliday, 1967). Then, section 5.1.3.2 describes how the relevant tokens for the study were extracted from the corpus. Section 5.1.3.3 gives an account of the statistical analyses that were carried out to test for dependencies between information status and intonation in the data. The presentation of results that follows, will be two-fold: the influence of the different information status categories will be presented in section 5.1.3.4, while section 5.1.3.5 describes how the underlying semantic processes interact with the tonal realisation. Section 5.1.3.6 then discusses the results.

5.1.3.1 Background: pitch accents and information status

There is a solid body of literature stating that pitch accents mark information status. Generally, accenting is said to signal *novelty* while deaccenting signals *given* information (e.g. Brown, 1983); yet there is counter evidence: various studies note accented information being *given* (Nooteboom and Kruyt, 1987; Yule, 1980; Bard and Aylett, 1999) and *new* information was found to be deaccented (Terken and Hirschberg, 1994).

As for pitch accent type (mostly in terms of ToBI categories, cf. chapter 3.1) H∗ (an accent with a high target in the accented syllable) was found to be the standard *novelty* accent for English (Pierrehumbert and Hirschberg, 1990) and German (Baumann, 2006). *Given* information, on the other hand, if accented at all, carries the L∗-accent (low target in the accented syllable) in English (Pierrehumbert and Hirschberg, 1990).

Baumann et al. (2006) states that deaccentuation is the most preferred tonal realisation of *given* tokens in German in controlled lab experiments, however, he also found them realised with H+L∗ (a high target followed by a low target in the accented syllable) in a corpus study. With regard to *accessibility*, Baumann (2006) found H+L∗ and deaccenting to be judged as acceptable in a perception experiment. His corpus study con-

firmed this: H+L∗ and deaccenting were the most frequent realisations of *accessibility*, followed by L∗.

5.1.3.2 Data extraction

To explore possible dependencies between information status and pitch accent type in a straight-forward way, the data was extracted in two steps.

The first step dealt with the underspecification of written text and the resulting ambiguity with respect to information status labels. As mentioned above, for experiments 0 and 1, only unambiguously labelled data (i.e. constituents that were only annotated with one information status label, cf. section 5.1.1) were extracted from the corpus.

The second step was concerned with the problem of where information status manifests itself in a noun phrase. The annotation scheme assigns information status labels to full phrases rather than to the nouns themselves. These constituents can be very long and can therefore not only include several potentially accentable syllables, but also several pitch accents. It is thus not trivial to decide which syllable(s) of a noun constituent should be included in the statistical analysis.

Three options were considered: firstly, all potentially accentable syllables (that is those that carry word stress) could be included. However, this would result in an overproportion of non-accented syllables whereby the results would be blurred. A second option would be to only include nuclear accents, since they are assumed to be the most significant ones. The disadvantage of such an analysis is that it cannot capture deaccentuation. Therefore, a third alternative, which is in a way a compromise between the two, was chosen: the analysis includes only those syllables that carry word stress (and are thus potentially pitch-accentable) and that are part of the last word in the annotated noun constituent. Thereby the word stressed syllable of the constituent's head is selected in most of the cases.

These restrictions outlined above apply to 608 syllables in the corpus.[2]

5.1.3.3 Statistical analysis: χ^2-test for independence and sparse data problems

To test for potential dependency of information status category and pitch accent type, χ^2-tests for independence of the two variables were applied. Table 5.2 shows the contin-

[2]In those cases where the syllables are accented (382), nuclear (last accents in a phrase) as well as non-nuclear accents are analysed. Hence, differences in the position within a tonal phrase were disregarded for the sake of having an adequate number of tokens. However, because the data was restricted to the last items in noun phrases, nuclear accents occur more than twice as often (273) as non-nuclear ones (109).

gency table reflecting the relationship between information status categories and pitch accent types. The χ^2-test resulted in a significant p-value of $p \ll 0.001$ indicating a dependency between the two factors. However, it has to be noted that the χ^2 approximation might be incorrect because of the lack of data in several combinations of information status and pitch accent type.

	H*	HH*L	!H*L	H*L	L*HL	L*H	L*!H	NONE
acc-descr	3	1	4	10	1	13	1	47
acc-gen	1	2	2	18	2	52	2	34
situative	4	0	2	10	0	43	4	33
d-giv-rep	0	0	0	2	0	8	0	7
d-giv-short	1	0	2	1	1	9	1	6
d-giv-epi.	3	1	1	14	1	39	2	32
d-giv-pro	2	0	0	0	0	10	0	29
new	4	1	12	40	5	37	10	38

Table 5.2: Frequency of pitch accent types by information status categories

Therefore the pitch accent types were subsumed under three more coarse-grained categories: *falling* for the falling accents H∗L, HH∗L, H∗, and L∗HL[3], *rising* for the rising accent L∗H and its down-stepped version L∗!H, and *NONE* for unaccented syllables. With these three categories all cells of the contingency table are filled and in only two cases the frequency is less than 5 tokens (see table 5.3). The χ^2-test was repeated and still resulted in a significant p-value of $p \ll 0.001$.

In the following, all tests that are reported as significant were applied to both the fine-grained pitch accent classification as well as the coarse-grained one, that avoids sparse data problems. In all cases, significance holds across both tests.

5.1.3.4 Results: Pitch Accents and Information Status Categories

Table 5.4 displays the relative frequency of each pitch accent type within an information status category in percent.

It clearly displays the difference between the two information status categories *accessible-description* and *accessible-general*. While the former is mainly realised without an accent (58.75%), the latter shows a clear preference for the rising accents (L∗H and L∗!H), which add up to 47.79%. The difference in pitch accent choice between these two information

[3]L∗HL was classified as falling since it is perceived as a falling tonal movement by human annotators.

	falling	rising	NONE
acc-descr.	19	14	47
acc-gen	25	54	34
situative	16	47	33
d-giv-rep	2	8	7
d-giv-short	5	10	6
d-giv-epi.	20	41	32
d-giv-pro	2	10	29
new	62	47	38

Table 5.3: Frequency of coarse-grained pitch accent categories by information status categories

status categories is statistically significant ($p < 0.005$ in a χ^2-test on the subset of the data only comprising these two categories).

Situatives are mainly realised with a rising accent (L∗H and L∗!H add up to 48.96%). The second most preferred realisation is deaccenting which occurs in 34.38% of all cases.

Among the subcategories of *d-givenness*, pitch accent type is significantly dependent on information status ($p < 0.05$ in a χ^2-test on the subset). A closer look gives insight into the nature of the differences: while *d-given-repeated*, *d-given-short*, *d-given-epithet* and do not differ significantly in pitch accent choice, the category for pronouns, *d-given-pronoun*, differs significantly from the others ($p < 0.005$). This is certainly due to the fact that pronouns are deaccented in the vast majority of cases (70.73%) while the other categories of *d-givenness* prefer rising accents (L*H or L*!H): their relative frequency ranges from 44.09% for *d-given-epithet* to 47.62% for *d-given-short* (numbers for both rising accents added). Deaccentuation is also very common for all three categories (41.18% for *d-given-repetition*, 28.57% for *d-given-short*, and 34.41% for *d-given-epithet*). However, it should be noted that for the categories *d-given-short* and *d-given-repeated* the number of tokens is relatively small (21 and 17 tokens, respectively).

The selectivity of *new* is not as pronounced as for some of the other categories. The distribution of accents differs significantly from the other information status categories ($p < 0.001$). The main accent types L∗H, H∗L and deaccenting are almost evenly distributed, which means that, compared to the other categories, there is an over-proportion of H*L accents. Falling accents in general (summing up to 42.18%) are clearly preferred over rising accents (31.97%) or deaccenting (25.85%).

	falling					rising			preferred accent
	H*	HH*L	!H*L	H*L	L*HL	L*H	L*!H	NONE	
acc-descr	3.75	1.25	5.00	12.50	1.25	16.25	1.25	58.75	NONE
acc-gen	0.88	1.77	1.77	15.93	1.77	46.02	1.77	30.09	L*H
situative	4.17	0.00	2.08	10.42	0.00	44.79	4.17	34.38	
d-giv-rep	0.00	0.00	0.00	11.76	0.00	47.06	0.00	41.18	
d-giv-short	4.76	0.00	9.52	4.76	4.76	42.86	4.76	28.57	L*H
d-giv-epi.	3.23	1.08	1.08	15.05	1.08	41.94	2.15	34.41	
d-giv-pro	4.88	0.00	0.00	0.00	0.00	24.39	0.00	70.73	NONE
new	2.72	0.68	8.16	27.21	3.40	25.17	6.80	25.85	fall (rise, NONE)

Table 5.4: Relative frequency of pitch accent type for each of the analysed information status categories in percent. The preferred realisations for each information status category are highlighted in light grey and summed up in the last column. Realisations given in brackets indicate that they are common, as well.

5.1.3.5 Results: Pitch Accents and Semantic Processes

While the relationship between information status category and pitch accent type does not appear to be a clear-cut one (at least not in terms of a one-to-one relationship), the data still demonstrates that the semantic context influences tonal patterns. As mentioned in section 5.1.1, for the definition of the information status categories according to Riester (2008a), three underlying semantic processes can be differentiated: *anaphora resolution*, *accommodation* and *introduction of a new discourse referent*. The experiment demonstrates, that these semantic processes are realised with different pitch accent types in the analysed data.

Table 5.5 displays the preferred pitch accents for the different processes in the IMS Radio News Corpus. While accommodation is mainly realised without a pitch accent, and the introduction of a new discourse referent is signalled by a falling accent in the majority of the cases, anaphora resolution is typically realised with the rising L*H accent. The cases in which the anaphora is a pronoun (*d-given-pro*) form an exception since they are mainly realised without a pitch accent.

semantic process	information status category	preferred accent
accommodation	acc-descr	NONE
anaphora resolution	gen.-acc situative d-giv-rep d-giv-short d-giv-epi. d-giv-pro	L*H NONE
introduction of a discourse referent	new	fall (rise, NONE)

Table 5.5: Accent preferences for different semantic processes.

5.1.3.6 Discussion

The pre-study described above was designed to test whether information status is a good example for a broader linguistic context which has an influence on pitch accent choice in the data.

Although there is not a one-to-one relationship with respect to Riester's (2008b) different information status categories, the results still demonstrate a significant influence of information status on pitch accent choice within the data. Moreover, the underlying semantic processes display a clear pitch accent preference.

Therefore the information status categories are taken to be suitable for an analysis of how the frequency of occurrence of a particular linguistic context, which can span sentence boundaries, influences pitch accent realisation.

5.1.4 Experiment 1: Absolute frequency of pitch accents and information status categories

Having established that information status categories are suitable representatives to model how linguistic context in general influences pitch accents, experiment 1 was conducted. The experiment examined the tonal realisation of different pitch accent types in relation to frequency of occurrence. The frequency of two types of linguistic information was explored:

1. the pitch accent type itself
2. the information status category (as one aspect of the linguistic context in which pitch accents can occur)

For both types of information, the frequency of occurrence was determined and related to the tonal realisation of pitch accent tokens. The PaIntE model (see Möhler and Conkie, 1998, and section 3.2.2) provides a way of capturing the tonal realisation of a pitch accent by 6 parameters defining its shape. Of course, different pitch accent types differ inherently in their shape. For instance, the temporal alignment of an accent's peak (measured by the PaIntE-parameter b) will be consistently earlier for a canonical H*L accent than for a canonical L*H accent: The peak of an H*L is expected in the accented syllable, while an L*H's peak is expected in the post-accented syllable. Therefore, the parameters were not compared directly. Instead the *variability* of the pitch accent types was determined and compared. The examination of variability is interesting from an exemplar-theoretic viewpoint, since Exemplar Theory assumes variability and frequency of occurrence to be interwoven (e.g. Pierrehumbert, 2001; Bybee, 2006). There are two main aspects to how variability changes with changing frequency of occurrence. On the one hand, there is the notion of entrenchment, which describes the effect that frequent units display less variation (cf. section 2.2.3). On the other hand, in the very early stages when a category evolves, the realisations of that category have to display increasing variability with increasing frequency.

While pitch contours, to the author's knowledge, have not been examined with respect to a potential frequency of occurrence effect, another prosodic parameter, *duration*, was shown to be related to frequency of occurrence: A. Schweitzer and Möbius (2004) found the variability of syllable duration to be affected by frequency of occurrence of the syllable (cf. section 4.1). By investigating the relationship between pitch accent variability and frequency of occurrence, this experiment answers the question whether *pitch*, being another prosodic parameter, is sensitive to frequency of occurrence, too.

Experiment 1 presents a first step in investigating possible frequency effects on pitch accents. It examines the frequency of pitch accent tokens regardless of the lexical material, that is of the words carrying those accents. Experiments 2 and 3 (sections 5.2 and 5.3) will provide more fine-grained analyses by looking at the frequency of a pitch accent type in combination with a particular word.

The experiment will be presented as follows: section 5.1.4.1 explains how variability is measured; the methodology for this involves the calculation of similarity between two pitch accent tokens. The statistical methodology is then described in section 5.1.4.2.

The experiment comprises two analyses: The first one (section 5.1.4.3) looks at within-type variability of different pitch accent types. It aims to answer the question whether three different pitch accent types, varying in frequency of occurrence, display a different

degree of similarity or variability among the acoustic implementation of their tokens. If this is the case, the analysis will examine whether there is a relationship between the frequency of occurrence of a type and its within-type variability and what the nature of the potential relationship is.

The second analysis (section 5.1.4.4) examines the influence of discourse context on within-type variability. Here, the influence of two different information status categories (*d-given* and *new*, cf. section 5.1.1) will be explored and the result will be related to their frequency of occurrence.

5.1.4.1 Calculation of pitch accent variability

To investigate the variability of the tokens of a pitch accent type, the shape of each token was captured by calculating the PaIntE-parameters (cf. section 3.2.2) from the smoothed F_0-contours.[4] Thereby the shape of each pitch accent was determined by the gradient of its rise and fall (parameters a1 and a2), the temporal alignment of its peak (parameter b), the range of its rise and fall (parameters c1 and c2) and the height of its peak (parameter d).

In exemplar theoretic-terms, these parameters can be thought of as the dimensions that span the exemplar space. Each pitch accent token can be understood as a vector in this six-dimensional space, i.e. the six values define a point in the six-dimensional vector space. This point is the endpoint of a vector with respect to a given Cartesian coordinate system. The vector's origin is identical to the origin of the coordinate system. Figure 5.1 illustrates this for a two-dimensional vector space. The values and dimensions are hypothetical, they could for instance correspond to peak height and peak alignment. The value for the first dimension (displayed on the horizontal axis) is 4, the value for the second one (displayed on the vertical axis) is 5. The resulting vector $v = [4, 5]$ can be visualised with the help of a Cartesian coordinate plane in which the origin is the starting point of the vector.

To measure the variability of a pitch accent type, the tokens of the respective type were compared pairwise and the similarity of each pair was calculated, employing a standard similarity measure, the cosine similarity: it measures the cosine of the angle between two vectors.

Figure 5.2 illustrates this in two-dimensional space. The angle between two vectors indicates their relative difference.

[4] F_0-contours were estimated from the speech signal with the ESPS get_f0 algorithm; smoothing was done using smooth_f0 from Taylor et al. (1999)

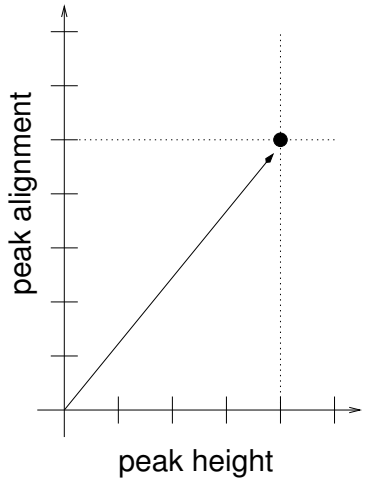


Figure 5.1: Illustration of a vector in a two-dimensional space. The two dimensions could represent e.g. peak height and peak alignment of each accent token. The value for the dimension displayed on the horizontal axis is 4, the value for the dimension displayed on the vertical axis is 5. These values mark the endpoint of the vector originating in the origin of the coordinate system.

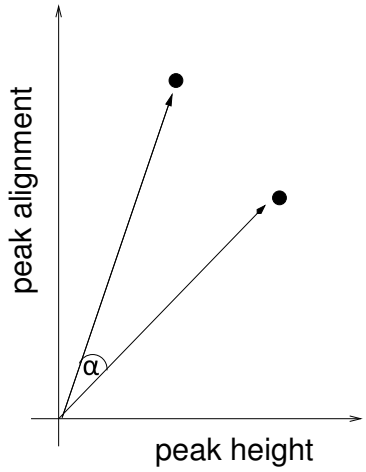


Figure 5.2: Illustration of two vectors in a two-dimensional space. If the vectors are parallel, the cosine is 1, indicating maximum similarity. If the angle is bigger, the cosine value is smaller, indicating less similarity.

Thus, two vectors that are parallel are perfectly similar (not necessarily identical, though): The cosine of the angle between such two vectors is +1, representing the highest possible similarity. If for a certain type many pairwise token comparisons result in a high similarity value, it can be concluded that the type displays low variability. On the other hand, if for a type many cosine comparisons result in low values (with -1 representing the lowest possible similarity), then the type is highly variable.

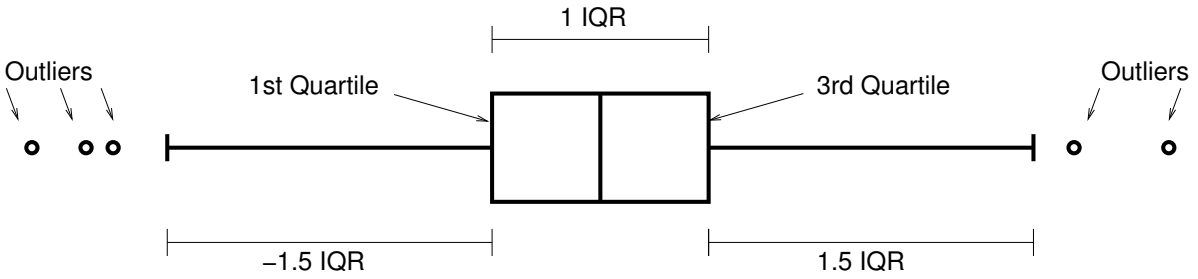


Figure 5.3: Example boxplot illustrating the definition of outliers. Data-points that are located outside the whiskers in a boxplot, were considered as outliers. The box spans 50% of the data; this range is called inter quartile range (IQR). Whiskers mark the area that are left and right of the box. In the definition of whisker length used here, a whisker is 1.5 IQR long.

To determine the variability of a pitch accent type, outlying tokens were removed from the data, the data was normalised, and similarity values for each token of a type and each other token of the same type, were calculated. These steps are described below.

Outlier removal Each token for which one of the PaIntE values was identified as an outlier was excluded from the data. Outliers were identified using R Development Core Team's (2006) boxplot function: data-points located outside the whiskers of a boxplot of the values were classified as outliers. The box in a boxplot spans 50% of the data, marked by the first quartile and the third quartile. Quartiles are the 3 points that divide the data into 4 equally sized groups. The distance between the first and the third quartile (which corresponds to the box size) is called interquartile range (IQR). The whiskers in a boxplot mark the data-points which are no more than 1.5 IQR[5] away from the quartiles (Abebe et al., 2001). More precisely, data-points that are neither more than -1.5 IQR away from the 1st quartile, nor more than +1.5 IQR away from the 3rd quartile, fall inside the whiskers (cf. figure 5.3).

This led to a reduced number of pitch accent tokens: while the original dataset comprised 1233 L*H accents, 704 H*L accents and 162 H* accents, the reduced set consisted of 1021 L∗H accents, 571 H∗L accents and 134 H∗ accents. It is noteworthy that there is a continuum of frequency of occurrence, high to low, from L∗H to H∗.

Normalisation The dimensions of the vectors were z-scored to account for the different ranges of the different PaIntE-parameters. The z-score value represents how many stan-

[5]There are other definitions, but here, the default definition as implemented in R (R Development Core Team, 2006) was used.

dard deviations the value is away from the mean value for that dimension and allows comparison of values from different normal distributions.

For each six-dimensional pitch accent category token the z-score value for each dimension $dim \in \{a1, a2, b, c1, c2, d\}$ was calculated. The formula to calculate the z-score z is given by:

$$\mathbf{z_{dim}} = \frac{x_{dim} - \mu_{dim}}{\sigma_{dim}} \tag{5.5}$$

where x_{dim} is the raw value to be normalised, i.e. one of the PaIntE-values, μ_{dim} is the mean of the population (the mean of all values of the respective PaIntE-parameter of the pitch accent type under investigation) and σ_{dim} is its standard deviation.

Hence, each of the z-scored values indicates if the PaIntE parameter value of the accent in question is increased or decreased with respect to the typical values for this dimension. For instance, a z-score value of 1 means that the parameter value is one standard deviation higher than the average value of the parameter for the accent in question.

After this procedure, each pitch accent was represented by a six-dimensional vector where each dimension value is a z-score.

Calculation of similarity values Each pitch accent token was compared to every other token of the same accent type.[6] Since pitch accent tokens were represented as (z-scored) vectors, the similarity between two tokens could be determined by calculating the cosine of the angle between the two vectors.

The cosine of the angle between two vectors $a = (a_1, a_2, ..., a_n)$ and $b = (b_1, b_2, ..., b_n)$ is given by:

$$\mathbf{cos}(\mathbf{a}, \mathbf{b}) = \frac{a \cdot b}{\| a \| \| b \|} \tag{5.6}$$

where a and b are vectors of the same pitch accent category, $\cdot$ represents the dot product (see equation (5.7) and $\|a\|$ is the magnitude of vector a (equation (5.8)).

The dot product of two vectors $a = (a_1, a_2, ..., a_n)$ and $b = (b_1, b_2, ..., b_n)$ is

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i \tag{5.7}$$

[6]Note that in the second analysis (section 5.1.4.4), the types being compared were not pitch accent types, but combined types of pitch accent and information status, e.g. one type would comprise all new H*L tokens.

where n is the dimension of the vector space, i.e. in the present study $n = 6$, since each PaIntE parameter represents one dimension.

The magnitude of a vector $a = (a_1, a_2, \ldots a_n)$ is the square root of the dot product of the vector with itself:

$$\|\mathbf{a}\| = \sqrt{a \cdot a} = \sqrt{a_1^2 + \cdots + a_n^2} \tag{5.8}$$

Pairwise comparisons With this similarity measure, each of the pitch accent tokens of a certain type was compared to every other token of the same type. Each comparison between vectors yields a cosine value that represents a similarity score in the range [-1,1], where -1 represents high dissimilarity and 1 represents high similarity.

Distributions of similarity values Using the methodology described above, for each of the analysed types the distribution of similarity values can be calculated by comparing each of the tokens with every other token of the same type. To compare two pitch accent types, their distribution of similarity values can be compared. The more similar the tokens of a particular type, the less variable is the type. This will be indicated by the distribution of similarity-values being shifted towards +1 (maximal similarity).

5.1.4.2 Statistical Analysis

To compare two distributions of similarity values, the Kolmogorov-Smirnov test was applied to test for equality of the two distributions.

When comparing two distributions (i.e. performing one test), the significance level was set to $\alpha = 0.05$. In those cases where two tests were carried out, the level of significance was adjusted according to the Bonferroni correction to $\alpha = 0.017$.[7]

The Bonferroni correction is discussed controversially. The main criticism of the adjustment concerns the increased likelihood of type II errors which lead to non-significance of actually significant findings (Pernegger, 1998). Although this conservative adjustment was applied, the statistical tests in this study resulted in significant p-values indicating the robustness of the findings.

[7]The Bonferroni correction adjusts significance levels according to the equation $\alpha = 1 - (1 - \alpha_1)^{\frac{1}{n}}$ where α_1 represents the target significance level (set to 0.05) and n represents the number of tests being performed, hence in this case $n = 2$.

5.1.4.3 Analysis 1: Pitch accent variability and frequency of occurrence of pitch accent types

In essence, this analysis examines how similar pitch accents of the same type are realised, that is, how variable the type is with respect to the tonal properties of its tokens. The analysis seeks to determine whether the variability of pitch accent types is related to their frequency (and if so, how they are related).

Three pitch accent types were analysed. As mentioned before, their frequency of occurrence differs: After removing the outliers, the most frequent type (L∗H) comprised 1021 tokens, mid-range frequency H∗L comprised 571 tokens and least frequent H∗ had 134 tokens.

Following the methodology set out in section 5.1.4.1, the PaIntE parameters of the extracted pitch accent tokens were all z-scored, stored as multi-dimensional vectors and, within each type, every token was compared to every other token of the same type by calculating the cosine of the angle between their vectors. Thus, for each type, a distribution of similarity values was obtained.

Pitch accent variability by pitch accent type Figure 5.4 depicts the density plots of the cosine similarity distributions for each of the analysed pitch accent types. The horizontal axis shows the cosine values. The vertical axis can be regarded as displaying the normalised frequency of occurrence of these values. Because of this normalisation, the three different types can be compared graphically in spite of the differences in frequency of occurrence and the resulting differences in the number of pairwise within-type comparisons.

As described above, the cosine values range from -1 to 1, with a value of 1 expressing maximal similarity (however not necessarily identity) and a value of -1 expressing minimal similarity. Hence, a distribution of cosine comparison values that is centred towards the left hand side of the graph (where the low cosine values are displayed) expresses high variability of the type: a lot of pitch accent tokens are dissimilar to each other. Distributions that are shifted towards the right hand side of the graph indicate that the analysed type is relatively invariable: the compared pitch accent tokens resemble each other and therefore the comparison results in higher similarity values.

An initial observation is that L∗H tokens tend to be realised fairly variably; the distribution is centred around zero. Tokens of H∗L tend to be produced more similarly i.e. the distribution is skewed towards higher similarity values, and tokens of H∗ more similarly again. These three distributions were tested against each other for significance

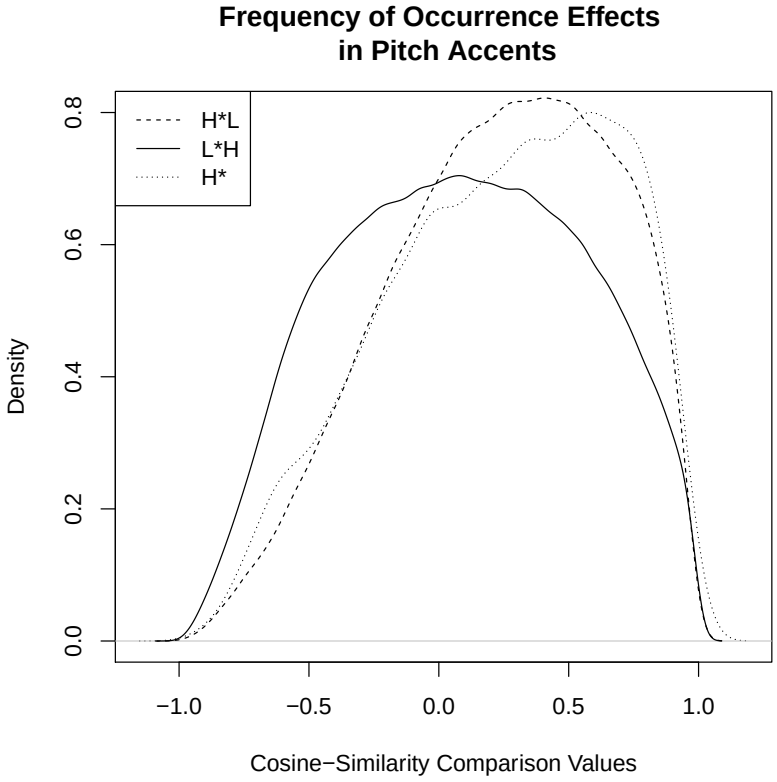


Figure 5.4: Density plots for similarity within pitch accent types (pitch accents were realised at any position within the tonal phrase). All distributions differ significantly from each other. There is a trend towards greater similarity from high-frequency L∗H to low-frequency H∗.

using the Kolmogorov-Smirnov test ($\alpha = 0.017$), yielding p-values of $p \ll 0.001$ which indicate significant differences between the three distributions.

It is noteworthy, that an increase in frequency of occurrence from one type to another co-occurs with an increase in variability among the corresponding tokens. In figure 5.4, most frequent L∗H is the leftmost distribution, hence, it is more variable than mid-frequency H∗L, which in turn is more variable than least frequent H∗ (with the distribution for H∗L being left of the one for H∗). Least frequent H∗ shows the lowest variability with the distribution being the rightmost one in the graph.

That is, an increase in the frequency of occurrence of the three pitch accent types co-occurs with an increase of the variability among their tokens. This effect seems to be subtle, given that the differences between the graphed distributions do not appear to be highly marked. However, they are statistically significant. Moreover, they are in keeping with exemplar-theoretic expectations, see sections 5.1.4.5 and 5.1.5 for a discussion.

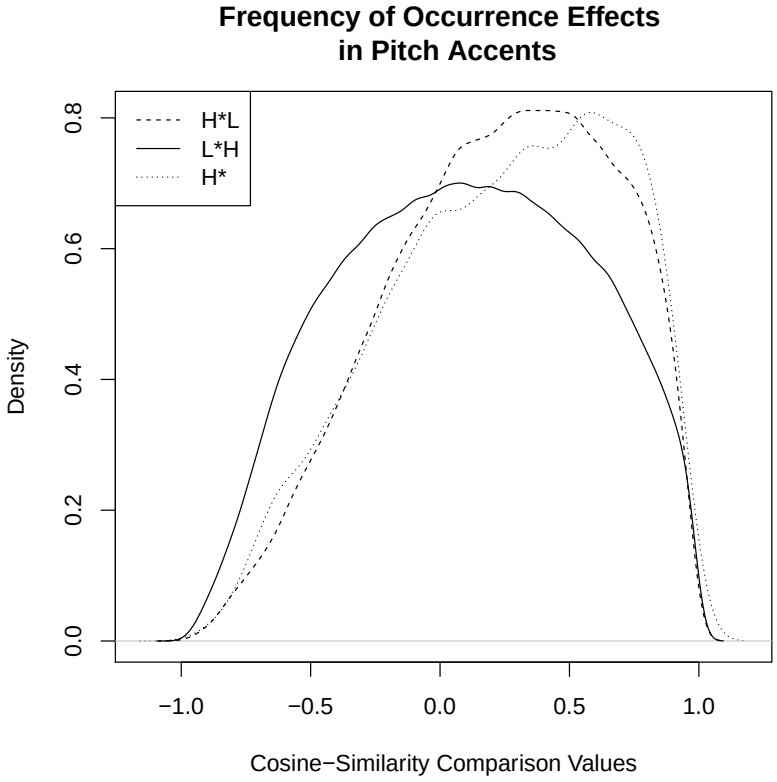


Figure 5.5: Density plots for similarity within pitch accent types in prenuclear position. All distributions differ significantly from each other. There is a trend towards greater similarity from high-frequency L∗H to low-frequency H∗.

At this point, it is also important to note that Walsh et al. (2008) reported significant differences between these distributions as well. However, there did not appear to be a frequency of occurrence effect. The differences between their findings and the results obtained in the present study, can be ascribed to the augmented PaIntE model (section 3.2.2), which was used in the experiments presented here.

Prenuclear vs. nuclear accents The extracted pitch accent tokens occur in different tonal positions, that is, nuclear and prenuclear accents are grouped together to avoid data-sparseness (which was an even bigger problem for analysis 2, where the tokens were reduced to accents with a certain information status). Moreover, the difference between nuclear and prenuclear accents is not considered to be a categorical one in autosegmental-metrical models of intonation on which the prosody annotation scheme is based (GToBI(S), see Mayer, 1995, and section 3.1.2). However, there is evidence that there are differences between the two groups with respect to the temporal alignment of

the accents (Silverman and Pierrehumbert, 1990; Mücke et al., 2006) and other models of intonation posit a categorical difference between these two groups (e.g. Crystal, 1969).

To attend to possible differences in the acoustic properties of nuclear vs. prenuclear pitch accents, the analysis described above was repeated with only prenuclear accents. The dataset consisting only of prenuclear accents comprised 121 H* accents, 225 H*L accents and 480 L*H accents.

Figure 5.5 displays the result for only prenuclear accents. The findings and interpretation described above held for the reduced dataset: the distributions for L*H, H*L and H* differed significantly ($p \ll 0.001$), and their trend towards greater similarity co-occurred with their decreasing frequencies. This indicates that the frequency effect occurs regardless of the accent's position in the tonal phrase. Therefore, nuclear and non nuclear accents were not separated in the following analyses.

5.1.4.4 Analysis 2: Pitch Accent Variability and Discourse Context

The analysis described here examines the influence of discourse contexts on the variability of the tonal realisation of pitch accents. Specifically, it aims to establish whether the discourse context has an influence on the variability of pitch accent tokens and if so, whether an effect of frequency of occurrence can be observed.

Having established that information status is a contextual property that influences intonation in section 5.1.3, two information status categories, *d-given* (hence expressions that were annotated with any of the labels that mark co-reference within the given discourse, cf. example 5.1 in section 5.1.1) and *new*, and their potential influence on the variability of pitch accents are analysed.

This analysis was carried out in the same manner as the first analysis, but only pitch accent tokens that were realised on either *given* or *new* expressions were examined. This resulted in a drastically reduced dataset with 87 *new* H∗L tokens, 44 *given* H∗L tokens, 102 *new* L∗H tokens , 114 *given* L∗H tokens, 21 *new* H* tokens and 10 *given* H* tokens.

To separate the influence of discourse context from the effect that frequency of occurrence has on pitch accent tokens of a certain type, pitch accent type was kept constant while information status varies. That is, the pitch accents L*H and H*L are analysed separately. Since the token counts for H* were extremely small, it was not examined further.

Variability of H*L by information status Figure 5.6 depicts the results for H∗L accents. As can be seen the two distributions of similarity values are clearly distinct. The

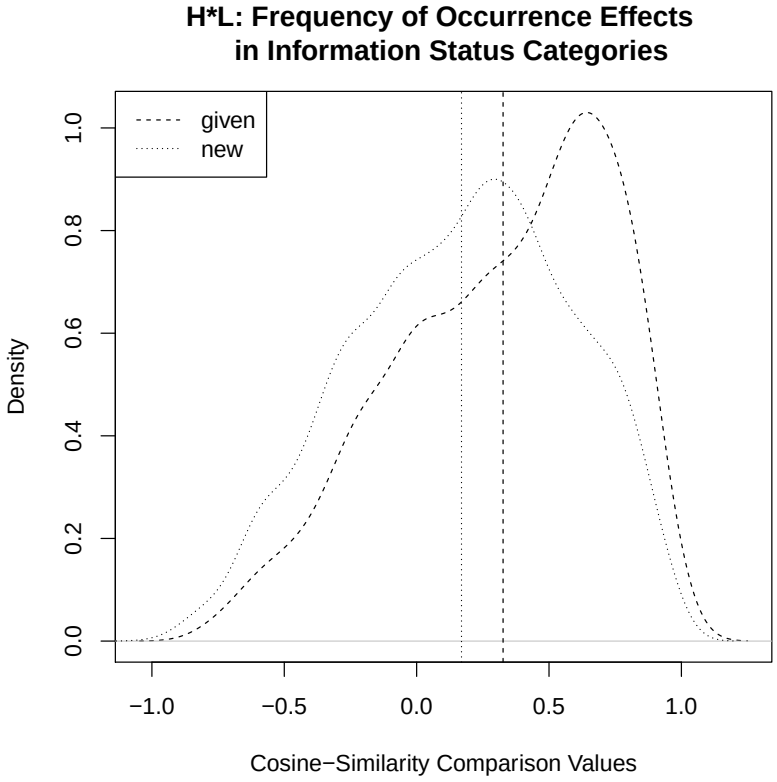


Figure 5.6: Density plots of similarity values of H∗L tokens. The tokens of the low-frequency information status category *given* display greater similarity to each other than those of the higher-frequency information status category *new*. Vertical lines mark the distributions' means.

.

distribution for *new* (dotted line) is centred left of the distribution for *given* tokens (dashed line). That is, the pairwise similarity comparisons of H*L accents in expressions that are contextually *new* generally result in lower values indicating less similarity of the pairs and therefore higher variability of the type "*new H*L*". The Kolmogorov-Smirnov test yielded a significant difference between the two distributions ($\alpha = 0.05, p \ll 0.001$), reflecting the clearly visible difference between the two curves.

It is noteworthy here that for H∗L the information status category *new* is more frequent than the category *given*. Indeed, approximately twice as many are labelled as *new* than those labelled *given* (87 vs. 44). Figure 5.6 illustrates that *new* H∗L accents are realised more variably than *given* ones. That is, again, an increase in frequency of occurrence co-occurs with an increase in variability, this time at the level of information status.

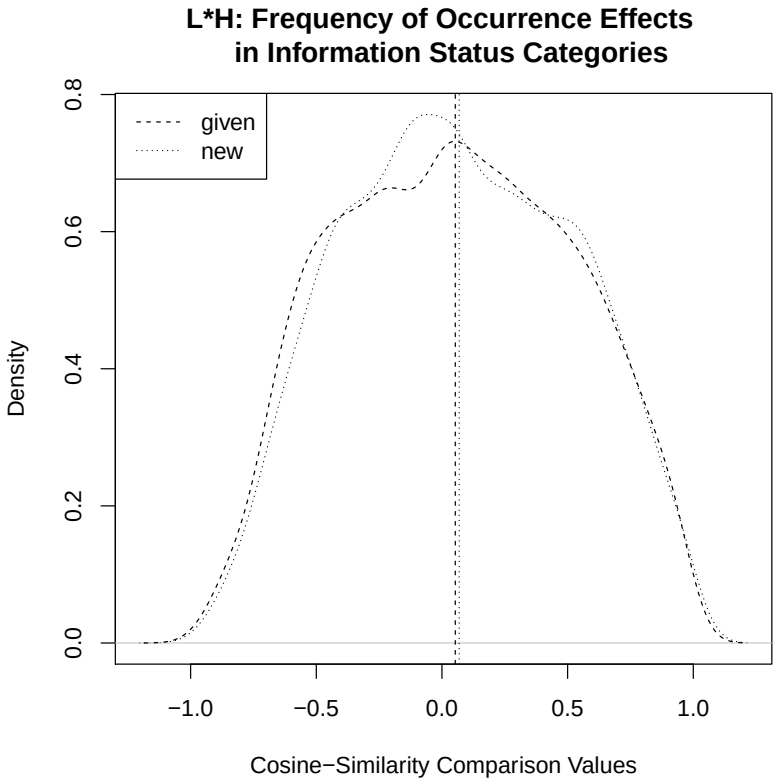


Figure 5.7: Density plots of cosine similarity comparison values of L∗H tokens. The difference between the curves is statistically significant, but the trend towards greater similarity for less frequent tokens is not observable. Unlike the case for H*L accents (displayed in figure 5.6) the number of tokens for both information status categories is comparable, here.

Variability of L*H by information status Figure 5.7 depicts the results for L*H accents. It is clearly visible that the two curves do not differ as much as those in the H∗L condition (figure 5.6). Both curves are centred around zero indicating that for both types the tokens are quite variable. However, the Kolmogorov-Smirnov test yields significance ($\alpha = 0.05, p = 0.044$), so the two distributions are still different. But the impact of information status in this case is unclear. Unlike in the case of H∗L, where it is obvious that the lower frequency of *given* tokens co-occurs with less variability, the distributions in figure 5.7 are alike. The mean values for the two distributions differ very slightly: The dashed vertical line marks the mean value for given tokens, the dotted line the mean for new tokens. The mean for new tokens is very subtly right of the mean for *given* tokens, i.e. there is a very slight trend towards less variability.

This result, which first looks as if it was contradictory to the result found for H∗L, can be explained in a straight-forward way, if frequency of occurrence is taken into consideration. The high frequency of L∗H accents in general results in a relatively high frequency of *given* L∗H tokens. The token number for both types is similar (102 *new* L∗H tokens vs. 114 *given* L∗H tokens), hence similar variability. Moreover, the slight trend towards less variability for *given* tokens could also be explained by the slight difference in token numbers (fewer tokens for *given* L∗H).

Again, the results can be well explained in the framework of Exemplar Theory. If the effect of information status is sensitive to frequency of occurrence, the effect of information status on H∗L can be explained as well as the "non-effect", that is no obvious difference between the two distributions on L∗H. In fact, looking at the frequencies of the four combined types *given L*H, new L*H, new H*L* and *given H*L* reveals a combined frequency effect; this is illustrated in section 5.1.4.5.

5.1.4.5 Conclusion: Combined Frequency Effects

Analysis 1 (presented in section 5.1.4.3) demonstrates that the variability of a pitch accent type is correlated with its frequency of occurrence: the more frequent an accent, the more variable its tokens. Frequency of occurrence also influences pitch accent variability if information status is taken into account: the more frequent the information status category for an accent, the more variable the accent's realisations. If the combined types of information status+pitch accent are similar in their number of tokens (as was the case for L∗H) no obvious changes in pitch accent variability were observed (cf. section 5.1.4.4).

This raises the question how these two factors, information status frequency and pitch accent frequency, interact. Figure 5.8 integrates the results. All combinations of the two information status categories *new* and *given* with the two pitch accents L∗H and H∗L are displayed. The number of tokens decreases from *new/given L∗H* (102/114 tokens) to *new H∗L* (87) and *given H∗L* (44). It is evident that an overall frequency of occurrence effect can be observed: the combined types *given L*H* and *new L*H*, which have similar numbers of instances, both are centred around zero; they are the most leftward curves in the graph. The distribution of the type *new H*L*, which is less frequent, shows a trend towards the right-hand side of the graph and thus represents greater similarity of the tokens, hence lower variability of the type. The distribution of similarity values for the least frequent combination of pitch accent and information status, *given H*L*, centres between 0.5 and 1.0 and is thus the most rightward curve in the graph, reflecting the

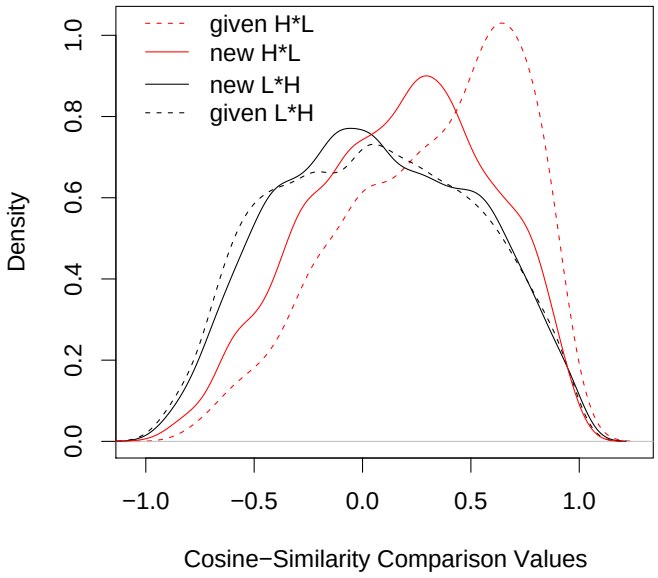


Figure 5.8: Overall frequency effect. There is a decrease in variability from *given* $H*L$ to *given/new* $L*H$ which co-occurs with an increase in the number of tokens.

highest similarity between the tokens. The graphed distributions demonstrate that, with decreasing frequency of a combined type, the variability of this type decreases.

Consequently, it can be concluded that the frequency of occurrence effect is not limited to pitch accents, nor to information status, but that it is the frequency of the combination of those two aspects that influences the similarity among the tokens of a type.

The finding that increasing frequency correlates with increasing variability is in keeping with exemplar-theoretic expectations: Firstly, as posited by Bybee (2006) and A. Schweitzer and Möbius (2004), high frequency of occurrence entails a large number of stored exemplars and a greater variety of different contexts that these tokens occur in. This gives the speaker the choice among a large number of production targets. The wider choice leads to a broader range of chosen targets for different productions and thus to more variable realisations of tokens of the same type.

Secondly, Exemplar-Theoretic models argue that while a category emerges, the variability among this category is expected to increase (see Pierrehumbert, 2001, and section 2.2.3) and would therefore predict the observed result. On the other hand, there is the phenomenon of entrenchment (see section 2.2.3, p.42), which predicts decreasing

variability for units that are highly frequent. See section 5.1.5.2 for a discussion of this controversy.

5.1.5 Discussion

The study described above sought to investigate whether frequency of occurrence effects can be found in intonation. More specifically, it examined whether the frequency of a tonal event – a pitch accent – influences the detailed acoustic realisation of that event. Furthermore, the experiment inspected the role of a broader linguistic context – information status – in the acoustic realisation of pitch accents. Therefore, a pre-study was conducted, which examined whether information status, as a property of utterances in a broader linguistic context (even over phrase boundaries), has an influence on tonal events in the analysed data (the IMS Radio News Corpus Rapp, 1998).

The pre-study revealed that there is a statistically significant relationship between pitch accent choice and information status in the data. Apart from that, it corroborated the semantic grounding for the applied information status labelling system (Riester, 2008b) in that the underlying semantic principles displayed different pitch accent preferences. Therefore, information status, was found to be suited to represent the linguistic context in which pitch accents occur.

The analyses of frequency of occurrence then demonstrated that pitch accents are subject to frequency of occurrence effects. The main findings are summed up in the following.

Pitch accent frequency. The frequency of occurrence of a pitch accent type correlates with the variability of the type. The more frequent a type, the more variability among its tokens. With decreasing frequency of a type, variability among the tokens also decreases. The presence of frequency of occurrence effects argues for a usage-based model of intonation, since such models encode the frequency of categories where autosegmental-metrical models of intonation would have to introduce a counter to handle frequency effects.

Frequency of linguistic context. The analyses highlight the importance of context and its frequency for the production of prosodic events: the frequency of occurrence of discourse context category correlates with the variability of the pitch accent tokens. The more frequent an information status category among the tokens of a pitch accent, the more variable the realisation of these tokens. In fact, a combined effect of pitch accent type frequency and information status frequency on the variability of pitch accent

tokens was observed. If a more functional approach to intonation, such as suggested by Batliner and Möbius (2005; cf. section 3.3), is adopted, pitch accents can be argued to reflect different linguistic functions: Specific linguistic functions influence the tonal contour and their effect results in a specific pitch accent label. The label is then rather a description of the actual acoustics than a linguistic category. Therefore the pitch accent label reflects a given linguistic context (a combination of several linguistic functions). The finding that the frequency of pitch accent types (labels) co-occurs with changes in pitch accent realisation can then be regarded as an indication that the frequency of occurrence of specific context configurations influences pitch accent realisation. Exemplar theory, building on the basic assumption that exemplars are stored in rich detail, would indeed predict an influence of all kinds of linguistic context because different contexts would be assumed to form distinct exemplar clouds from which the production targets are selected.

Summary Both findings add another piece of evidence to the many findings of frequency effects in language that were discussed in section 2.1.3. In addition, they extend these findings to tonal properties, demonstrating that the fundamental frequency contour is influenced by frequency of occurrence. This is important for the general question of how pitch accents are applied to words and phrases. While autosegmental-metrical theories of intonation (cf. chapter 3) do not make hypotheses on frequency of occurrence effects, exemplar theory is able to explain frequency effects naturally, since the different numbers of exemplars lead to different numbers of potential production targets and can therefore account for frequency of occurrence effects: exemplars of a type that occur often are more variable because they offer the speaker a wider selection of exemplars to choose from during production. While the existence of frequency effects on tonal parameters itself is a new finding and argues for a usage-based account of speech production, the explanation given above has some important consequences and also raises some questions that remain to be clarified. Both implications of, and objections to, the given explanation will be discussed in the following sections.

5.1.5.1 Implications

It is important to highlight some implications if Exemplar Theory is used to explain the results. Firstly, if pitch accent shape is subject to frequency effects then it must be considered that intonation could be stored as part of the lexical item, at least in some cases. Otherwise the effect would not be present. The idea that intonation can be stored

with the lexical items conflicts strongly with autosegmental-metrical theories that claim intonation to be post-lexical (cf. section 3.1) and adds a piece of evidence to a small but growing body of evidence that fundamental frequency can be stored (cf. section 4.2).

Secondly, since it has been shown that the frequency of occurrence of an information status category influences pitch accent shape, it has to be argued that information status should be considered as a dimension in the exemplar space. Given that information status has been used in this study to represent broader linguistic contexts that influence pitch accent shape, it can be suspected that linguistic context in general is encoded in the exemplar and that changes in the frequency of these contexts can lead to changes in the acoustic detail of the pitch accent productions.

5.1.5.2 Conclusions and further research questions

As mentioned above, it has been argued in the literature that frequent behaviours undergo entrenchment processes. That is, a very frequent category is expected to display less variability because the underlying processes (e.g. articulatory movements) are highly learned so that they are very efficient and precise. On the other hand, Exemplar Theory expects that the variability of exemplars increases as a category emerges, that is, as the category becomes more and more frequent (cf. e.g. Pierrehumbert, 2001, discussed in section 2.2.3).

Since pitch accents and information status categories in adult speech should be frequent enough to have reached a stable level of variability, one would probably expect to find an effect of entrenchment in the data. However, it has been argued that a greater number of contexts entails a greater number of production targets which leads to more variability among the exemplars of a frequent type. Though this might be a valid assumption, the results presented here can not necessarily considered as supporting it because of two reasons: firstly, the observed effect might be solely a consequence of different segmental contexts, that is of different word types on which the pitch accents were realised. Secondly, the selection of a production target is unlikely to be solely based on pitch accent and information status type. The following subsections elaborate on these two issues.

Increased number of segmental contexts The number of different contexts for frequent types is expected to be higher than the number of contexts for an infrequent type. A frequent pitch accent type can occur in many different segmental contexts – since the number of word types they occur on is supposedly greater than for infrequent types. Con-

sequently, word types should be included in the analysis so that firstly, micro-prosodic effects can be controlled for and that secondly, the frequency of the word is also taken into account. That is, instead of simply counting occurrences of pitch accent tokens, the occurrences of pairs of accents and words should be counted and a combined type of *pitch accent and word* should be analysed, similar to the combined types of pitch accent and information status that were examined in this experiment.

Word types and exemplar selection Assuming that for a frequent pitch accent type, or for a frequent combined type of pitch accent and information status, a speaker has a greater number of exemplars to choose from during production implies that pitch accent type and information status are the only criteria considered in the selection process. Hence, it implies that all exemplars of a certain pitch accent type (or of a combined type, that is only those ones on expressions with a certain information status) are activated and contribute to the construction of the production target. This seems unlikely since it would imply that the intended utterance is constructed according to pitch accent type but not according to the intended words. Therefore, word type is bound to be a selection criterion that narrows down the set of exemplars that contribute to constructing the production target. Consequently, when examining the exemplar-theoretic selection process, word identity should be taken into account.

In summary, experiment 1 lead to the following implications:

Implication 1: pitch accents are subject to frequency effects implying that tonal contours can be stored in the lexicon, at least in some cases – possibly depending on frequency of occurrence

Implication 2: the context of a pitch accent is important; discourse context frequency influences pitch accent realisation

Questions about the exemplar-theoretic production of pitch accents, however, remain. For instance, are there frequency effects on pitch accents if the segmental context is controlled for? And if so, how is the tonal realisation influenced? Experiment 2, presented in the next section, aims to answer those questions. Moreover, both criticisms described above are taken into account when analysing pairs of pitch accents and words as is done in experiments 2 and 3 (described ins sections 5.2 and 5.3).

5.2 Experiment 2: Absolute frequency of Pitch Accent+Word

Experiment 1 demonstrated that the frequency of both pitch accent type and information status category has a significant influence on the realisation of a pitch accent's contour. Also, exemplar theory has been shown to be suitable to explain frequency effects of tonal events. However, experiment 1 could be improved by performing a more fined grained analysis which considers the assumption that exemplars contain information about the word. Besides, if one assumes that intonation can – at least in certain cases – be part of the information that is being stored with the exemplars, then there should be dependencies between the lexical and the tonal level, contrary to the assumption that intonation is "post-lexical" (autosegmental-metrical theories of intonation, see Pierrehumbert, 1980, cf. chapter 3). To explore the relationship between words and their pitch contour more closely, another study was conducted. The aim of this study was to find potential frequency effects of words occurring with a specific tonal contour, i.e. with a specific pitch accent type. The reasoning behind this is the following: If exemplar-theoretic assumptions hold for intonation, and exemplars contain segmental as well as suprasegmental properties, then frequency effects for frequently occurring combinations of a lexical item with a particular tonal contour would be expected on the tonal level, viz. on fine phonetic detail of the tonal realisation. Specifically, the experiment aims to answer the question whether there are effects of the frequency of accent+word pairs on their tonal realisation, and if so, what these effects are. Therefore, pairs of pitch accent and word were examined with respect to characteristics of the pitch accent's shape (such as the range of the accent or its temporal alignment) and changes in their shape were related to the frequency of occurrence of the respective pair.

Since word types are typically distributed in a Zipfian distribution (Zipf, 1949), that is, there is a large number of rare types, a word-based examination is even more likely to run into data-sparseness problems. Therefore, a larger speech corpus was constructed. The process of creating the new resource is described in the following section (5.2.1). Then, section 5.2.2 describes the data extraction of both accent tokens and features describing their linguistic context. The statistical analysis and the methodology employed in visualising the results is described in section 5.2.3, before sections 5.2.4 and 5.2.5 present the results, which are interpreted and discussed in section 5.2.6.

5.2.1 Construction of new speech database: DRadio-corpus

To make a larger speech corpus available, recordings from a German radio station, *Deutschlandradio*, were processed. The radio station provides their news broadcasts as mp3s on their website, as well as the written transcript of the read news. To construct a large speech database, six days of radio news were downloaded. Since the written transcripts differed from the read news, they had to be corrected manually. After that, the speech files were aligned with phone, syllable and word boundaries using forced alignment (Rapp, 1998). Two and a half days were annotated prosodically using the GToBI(S)-scheme (Mayer, 1995, described in chapter 3.1) by three trained annotators. At the time when the experiment was conducted, the annotations of one day were labelled by one annotator and checked by the other two for mistakes or uncertainties. Therefore, the speech data used here consist of one day of radio news broadcasts (26/04/2007). It comprises 2 hrs 43 mins of read speech by 4 professional speakers (2 female, 2 male).

5.2.2 Data processing

5.2.2.1 Outlier Removal, Extraction of Pitch Accent Tokens and Normalisation

For each pitch accent, six parameters reflecting its shape were extracted from the smoothed F_0-contours[8] using the PaIntE model (see Möhler and Conkie, 1998, and section 3.2.2).

Data processing included several steps: manual labelling, F_0 approximation, F_0 smoothing and PaIntE parametrisation. Since each step increases the number of potential error cases[9], only pitch accent tokens that were clearly reasonable examples of either L*H or H*L were extracted. Reasonable examples were defined as meeting one of the following criteria:

1. Both of the function's sigmoids are used to model the accent, hence F_0 was approximated according to the two-sigmoid (standard) case (cf. section 3.2.2.3).

2. If only one sigmoid is used (referred to as one-sigmoid case in section 3.2.2.3), it is the falling sigmoid for H*L and the rising sigmoid for L*H.

[8] F_0-contours were estimated from the speech signal with the ESPS get_f0 algorithm; smoothing was done using smooth_f0 from Taylor et al. (1999)

[9] For instance, in manual GToBI labelling the inter-transcriber consistency has been determined to be at 71% (Grice et al., 1996). This value included the agreement on whether or not the word carried a pitch accent. Agreement on accent types, however, was only 51%.

Then, all pitch accent tokens for which one of the PaIntE-parameter's value was identified as an outlier were removed. Outliers were again defined as tokens that fell outside the whiskers of boxplots (R Development Core Team, 2006, see p. 102 for a detailed explanation of which tokens are located in- respectively outside the whiskers).

Since nuclear and prenuclear accents can differ in their shape (Mücke et al., 2006), only nuclear accents (defined here as the last accent in a tonal phrase) were extracted.

For each word type, the frequency of the combination of this type with an accent type (e.g. "Berlin+H*L"), referred to as *word+accent pair* in the following, was calculated.

To normalise for speaker differences, the PaIntE parameters were z-scored, on an accent-type basis, this time for each speaker separately according to equation 5.5 (page 103).

That is each of the z-scored values indicates if the PaIntE parameter value of the accent in question is increased or decreased with respect to the typical parameter values of the individual speaker. In this way, not only the different ranges of PaIntE parameters are normalised but also inter-speaker differences are controlled for.

The original dataset comprised 1757 nuclear L*H and 1226 nuclear H*L accents. Excluding accents which did not meet the above listed criteria reduced the numbers to 1668 L*H and 1098 H*L accents. Removing each accent for which one of the PaIntEvalues was an outlier, further reduced the data to 1098 nuclear L*H tokens (578 word+accent pairs) and 580 nuclear H*L tokens (385 word+accent pairs).

5.2.2.2 Extraction of features describing the lexical context

Three other factors that were expected to potentially influence pitch accent shape (see e.g. Van Santen and Möbius, 2000; Jilka and Möbius, 2007) were extracted using Festival (Taylor et al., 1998). They control for the differences between word types in terms of number of segments in the syllable and position of the syllable in the word

The factors were: number of segments in the accented syllable's onset and coda, respectively (0, 1, 2, or more than 2 segments) and the position of the accented syllable in the word, which distinguishes not only between different positions of the syllable in polysyllabic words, but also between mono- and polysyllabic words by assigning the value *single* to syllables in monosyllabic words (whereas syllables in polysyllabic words receive one of the values *initial, mid* or *final*).

5.2.3 Statistical analysis and visualisation of the effect

In an exploratory statistical approach, the two pitch accent sets (nuclear L*H and nuclear H*L accents) were analysed. To this end, 4-way ANOVAs were performed for each of the z-scored PaIntE parameters, where the parameter in question was the dependent variable and the above mentioned factors and word+accent pair frequency were incorporated as predictors.

It has to be noted that the data (coming from a natural speech corpus, not from a controlled experiment) is unbalanced. Since Type I ANOVAs were computed, a hierarchical series of regression equations are estimated, at each step adding an additional effect into the model (see StatSoft, 2011). Thus, the order of the factors is relevant for the result. The factors being incorporated last in the model are least likely to be found as having a significant influence on the dependent variable. Therefore, the control factors *onsetsize*, *codasize* and *syllable position in the word* were incorporated in the model before incorporating the factor *word+accent pair frequency*.

To visualise the effect of word+accent pair frequency isolated from the other analysed factors, the following methodology was applied: In those cases where word+accent pair frequency was found to have a significant main effect on the predicted values, 3-way ANOVAs were performed with word+accent pair frequency removed from the set of predictors. That is, the effects of the factors *onsetsize*, *codasize* and *syllable position in the word* were calculated. The residuals of these ANOVAs are the part of the data which cannot be explained by these three factors. Since *word+accent pair frequency* was known to have a significant effect on the data, the residuals were plotted against *word+accent pair frequency* to visualise how the frequency of occurrence of word+accent pairs affects the realisation of pitch accent contours (with the effect of the control factors removed).

5.2.4 Results on L*H accents

For L*H accents, main effects of the frequency of the word+accent pair were found for the range of the accent's rise (parameter c1) and the alignment of the peak within the syllable (parameter b). For the range of the rise the frequency of word+accent was the only main effect, with the control factors being significant mainly in interactions with the frequency value. Hence, for the range of the rise, i.e. the amplitude of the accent, the frequency of word+accent seems to be an important factor. For the alignment of the peak, however, all the control factors were significant as main effects. This might indicate that peak alignment is dependent on a variety of factors, in keeping with results from

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
onsetsize	3	3.69	1.23	1.2050	0.306734	
codasize	2	0.18	0.09	0.0868	0.916855	
sylpostype	3	0.80	0.27	0.2630	0.852098	
wordfreq	1	4.09	4.09	4.0163	0.045324	*
onsetsize:codasize	5	9.52	1.90	1.8673	0.097391	
onsetsize:sylpostype	9	10.20	1.13	1.1117	0.351245	
codasize:sylpostype	6	5.19	0.87	0.8487	0.532379	
onsetsize:wordfreq	3	5.21	1.74	1.7027	0.164760	
codasize:wordfreq	2	1.21	0.61	0.5947	0.551908	
sylpostype:wordfreq	3	3.80	1.27	1.2421	0.293176	
onsetsize:codasize:sylpostype	13	28.69	2.21	2.1645	0.009322	**
onsetsize:codasize:wordfreq	5	10.05	2.01	1.9709	0.080451	
onsetsize:sylpostype:wordfreq	6	19.48	3.25	3.1838	0.004225	**
codasize:sylpostype:wordfreq	6	1.88	0.31	0.3076	0.933148	
onsetsize:codasize:sylpostype:wordfreq	6	20.03	3.34	3.2751	0.003394	**
Residuals	1024	1043.97	1.02			

Table 5.6: ANOVA results (main effects) for L∗H dependent variable *c1*. Stars indicate significance (one star indicates significance at $\alpha = 0.05$, two stars at $\alpha = 0.01$). The frequency of word+accent pair is encoded as "wordfreq", main effects of which (or interactions with it) are bold faced.

e.g. Jilka and Möbius (2007) or Kohler (1990). The next sections describe the results in more detail.

Results on peak height Table 5.6 gives the results of the ANOVA for the range of the rise for L*H accents (parameter c1). As can be seen, the frequency of word+accent pair has a significant effect on the amplitude of the rise ($p < 0.05$). It is also significant in interactions, that is, the impact of the factors is dependent on the impact of other factors, highlighting the importance of context and, above all, the combination of contextual features, for the actual acoustic realisation of pitch accent tokens.

Figure 5.9 presents the visualisation of the effect of word+accent pair frequency. As described above, the figure displays the residual values of the 3-way ANOVA (with word+accent frequency removed from the predictors) for the z-scored parameter c1 for nuclear L*H accents, plotted against the frequency of the word+accent pair. In other words, it shows the portion of the data that cannot be explained with the effects of the control factors, as a function of the frequency of the word+accent pair. As can be seen, there is a subtle increase in the amplitude of the accents' rises, visualised by the regression line that slightly rises from the left hand side (where the low frequency values are located)

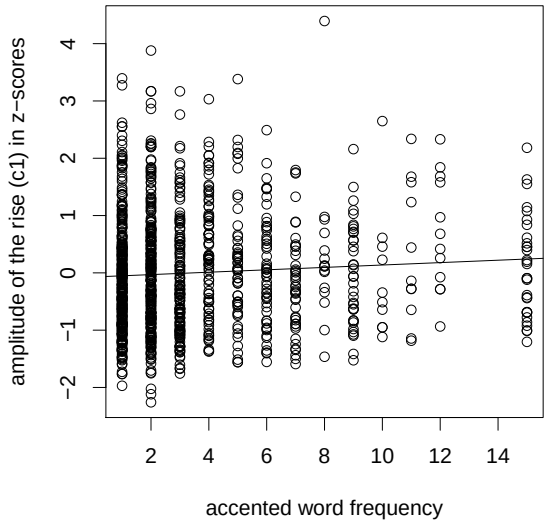


Figure 5.9: Residual-values for z-scored parameter *c1 (range of the rise)* for nuclear L*H accents plotted against the frequency of the word+accent (here word+L*H) pair.

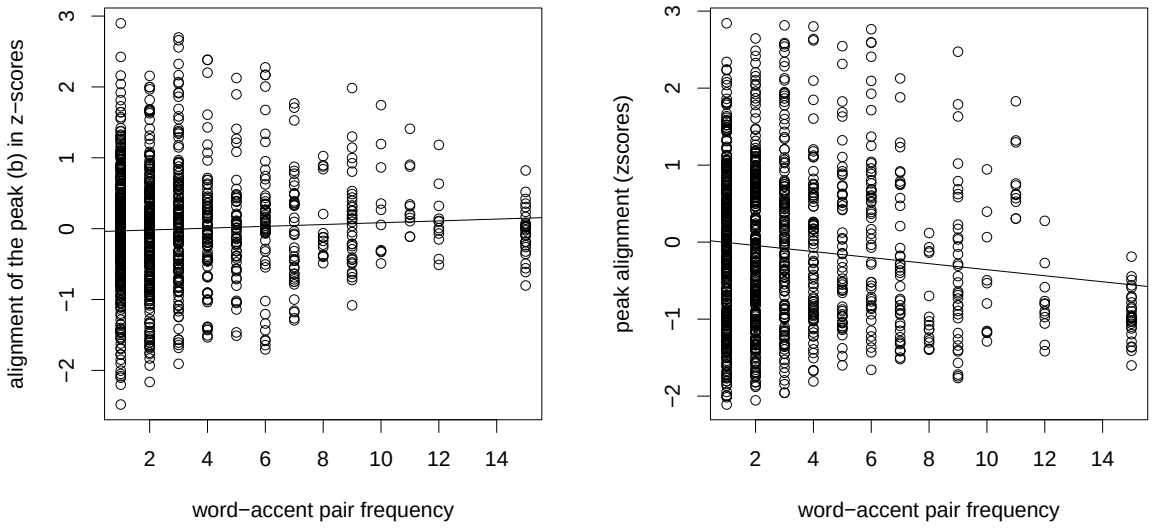
towards the right hand side of the graph (i.e. towards higher frequency). Thus, with increasing frequency of the word+accent pair, the range of the accents increases.

Results on peak alignment The results of the 4-way ANOVA for parameter b (peak alignment) can be seen in table 5.7. All predictors have a significant main effect, with $p < 0.001$ for the control factors and $p < 0.05$ for word+accent pair frequency. As mentioned above, this highlights the dependence of peak alignment on different contextual factors (see Jilka and Möbius, 2007, for details, e.g. effects of syllable structure).

Figure 5.10a presents the visualisation of the effect of word+accent pair frequency on peak alignment. Analogous to the visualisation for parameter c1 in figure 5.9, the graph displays the residuals of the 3-way ANOVA (without word+accent frequency amongst the predictors) plotted against the frequency values for word+accent pairs. That is, the graph shows the portion of the data that cannot be explained by the control factors and their interactions, as a function of word+accent frequency. As can be seen the regression line rises slightly from the left hand side (where the low frequency pairs are located) towards the right hand side of the graph (where higher frequency pairs are displayed). Thus, the peak alignment value slightly increases with increasing frequency. That is, the peak is realised later as word+accent frequency increases. Remember that this slight increase is statistically significant.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
onsetsize	3	22.12	7.37	11.0537	<0.001	***
codasize	2	55.14	27.57	41.3276	<0.001	***
sylpostype	3	321.36	107.12	160.5742	<0.001	***
wordfreq	1	2.71	2.71	4.0572	0.044246	*
onsetsize:codasize	5	7.80	1.56	2.3391	0.039967	*
onsetsize:sylpostype	9	6.69	0.74	1.1147	0.349060	
codasize:sylpostype	6	1.72	0.29	0.4287	0.860118	
onsetsize:wordfreq	3	3.83	1.28	1.9151	0.125396	
codasize:wordfreq	2	0.65	0.33	0.4887	0.613576	
sylpostype:wordfreq	3	2.26	0.75	1.1292	0.336176	
onsetsize:codasize:sylpostype	13	34.71	2.67	4.0024	0.000002	***
onsetsize:codasize:wordfreq	5	1.79	0.36	0.5361	0.749044	
onsetsize:sylpostype:wordfreq	6	6.65	1.11	1.6622	0.127017	
codasize:sylpostype:wordfreq	6	4.07	0.68	1.0157	0.413416	
onsetsize:codasize:sylpostype:wordfreq	6	9.12	1.52	2.2775	0.034420	*
Residuals	1024	683.12	0.67			

Table 5.7: ANOVA results (main effects) for L∗H, dependent variable *b* (peak alignment). Stars indicate significance (one star indicates significance at $\alpha = 0.05$, three stars at $\alpha = 0.001$). The frequency of word+accent pair is encoded as "wordfreq", main effects of which (or interactions with it) are bold faced.



(a) residuals of 3-way ANOVA

(b) z-score values

Figure 5.10: Parameter *b (peak alignment)* values for nuclear L*H accents plotted against the frequency with which the respective word occurs with a nuclear L*H accent.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
onsetsize	3	6.66	2.22	2.7315	0.043242	*
codasize	2	4.60	2.30	2.8274	0.060081	.
sylpostype	3	3.44	1.15	1.4084	0.239455	
wordfreq	1	13.45	13.45	16.5477	<0.001	***
onsetsize:codasizetype	5	10.31	2.06	2.5354	0.027865	*
onsetsize:sylpostype	8	16.19	2.02	2.4889	0.011742	*
codasizetype:sylpostype	6	8.30	1.38	1.7025	0.118299	
onsetsize:accwordfreq	3	5.31	1.77	2.1766	0.089853	
codasizetype:wordfreq	2	3.49	1.74	2.1452	0.118083	
sylpostype:wordfreq	3	3.80	1.27	1.5588	0.198465	
onsetsize:codasizetype:sylpostype	8	6.30	0.79	0.9685	0.459763	
onsetsize:codasizetype:wordfreq	3	2.16	0.72	0.8846	0.448862	
onsetsize:sylpostype:wordfreq	4	1.45	0.36	0.4454	0.775811	
codasizetype:sylpostype:wordfreq	6	5.52	0.92	1.1319	0.342305	
onsetsize:codasizetype:sylpostype:wordfreq	4	11.87	2.97	3.6515	0.006038	**
Residuals	518	421.14	0.81			

Table 5.8: ANOVA results (main effects) for nuclear H∗L dependent variable *c2*. Stars indicate significance, (one star indicates significance at $\alpha = 0.05$, two stars at $\alpha = 0.01$ and three stars at $\alpha = 0.001$). The frequency of word+accent pair is encoded as "wordfreq", main effects of which (or interactions with it) are bold faced.

Interestingly, the plot for the real values (i.e. for the z-scored b-parameter values as opposed to the residuals of the 3-way ANOVA), looks different: figure 5.10b displays the z-scored b-values as a function of word-accent pair frequency. In other words, here, the portion of the data that can be explained by the control factors is *not* removed from the plotted values. As can be seen, here the regression line slightly falls with increasing frequency. Since the effect of the control factors is not removed from the data in this graph and since we know that the effect of word+accent pair frequency causes the value to *increase* it can be concluded that the decrease in the b-values displayed here is an effect of the control factors overriding the effect of word+accent frequency. That is, different effects are "pulling" the peak in different directions: while increasing frequency of word+accent causes later peak alignment, the combination of the other factors causes earlier peak alignment for the more frequent pairs in the analysed data.

5.2.5 Results on H*L accents

For nuclear H*L accents, the only parameter that was significantly influenced by word+accent pair frequency was parameter *c2* – the range of the accents' fall. Table 5.8 presents the results of the 4-way ANOVA predicting the amplitude of the fall (parameter c2). As

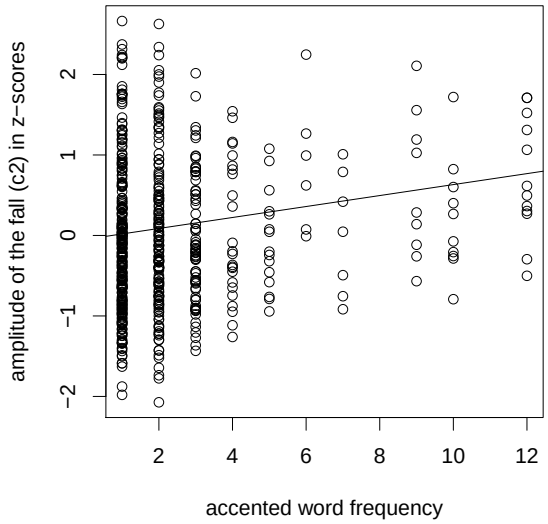


Figure 5.11: Residual-values for z-scored parameter *c2* for nuclear H*L accents plotted against the frequency of the word+accent (here word+H*L) pair.

can be seen, the frequency of word+accent has a significant main effect ($p < 0.001$), as has the interaction of all the control factors and word+accent frequency ($p < 0.01$).

Figure 5.11 visualises the effect of word+accent frequency according to the methodology explained above: the residuals of the 3-way ANOVA predicting the z-scored *c2* values without accent+word frequency are plotted against accent+word frequency. As can be seen, the regression line rises from left to right, hence, with increasing frequency of the respective word+accent pairs, the amplitude of the fall increases.

5.2.6 Discussion

The aim of this experiment was to test whether frequency effects of words occurring with a specific pitch accent can be observed. Analysing word+accent pairs instead of comparing all tokens of a specific pitch accent type (as was done in experiment 1) takes the word form into account and is therefore more suitable when interpreting the results in an exemplar-theoretic framework where the selection of an exemplar as production target should take the word that the pitch accent was realised on into account.

Teasing apart the effect of word+accent pair frequency from other influential factors is tricky, however the ANOVA results above indicate an effect of frequency of occurrence on the realisation of pitch accent range, and on the temporal alignment of L∗H peaks. It

is important to note however that while word+accent pair frequency affects L∗H peak location (visible when plotted against the residuals), this effect is not present in the raw data. In other words while frequency of occurrence aligns the peak later, the other factors rein it in.

With regard to pitch accent range, the rise tends to be larger with increasing word+accent pair frequency in nuclear L*H accents, and the fall is greater in nuclear H*L accents. Thus, for both accent types, the range of the accent increases, hence, the realisation of an accent becomes more prominent as the frequency of the word+accent pair increases.[10]

From the perspective of theories of lexical and prosodic production, the crucial point in these findings is that there is a dependency between the lexical and the tonal level. That is, combined word+accent frequency influences the tonal realisation of pitch accent tokens. In generative top-down-models of intonation this relationship is not expected or predicted. Within an exemplar-theoretic framework, the bias towards more pronounced accents with increasing frequency can be explained in terms of a production-perception loop. This interpretation and some implications of the results as well as further questions are discussed in the following sections.

5.2.6.1 Exemplar-Theoretic-Interpretation of the Results

The observed trend towards greater prominence for more frequent word+accent pairs can be explained and modelled in an exemplar-theoretic model. An exemplar-theoretic explanation relies on the production-perception loop in exemplar theory: during speech production a speaker selects a stored exemplar as a production target. Assuming that pitch accents can be stored with the word, the speaker would select an exemplar that best matches their communicative goal. The main purpose of pitch-accentuation is to make a word more prominent,[11] in order to draw the listener's attention to the word and make it more noteworthy. Moreover, the most prominent tokens are likely to have the greatest activation in the speaker's memory as they are more marked. Consequently, it seems plausible that most prominent exemplars are likely to be chosen as a production target. During articulation, imprecision inherent in the production process, i.e. production noise, will yield a pitch-accented word which is either more, or less, prominent than the production target itself. This new production will be perceived and stored as a new

[10] It has to be noted that the notion of *prominence* here refers to *acoustic* prominence.

[11] of course there are pitch accents that are mainly assigned to certain syllables due to rhythmical reasons, see e.g. Augurzky (2008), but even there, increasing the prominence is necessary to stress the word

exemplar, and, if more prominent, will be more likely to be selected as a production target than the previous target in subsequent productions (and possibly be produced more prominently again), and will be less likely to be selected for production if less prominent. Ultimately, this will yield more prominent productions and increase prominence within the cloud of word+accent pairs as a whole, as the less prominent exemplars will fade from memory due to lack of activation. This behaviour could become entrenched over time, preventing excessive peak prominence (Pierrehumbert, 2001).

Of course there are instances where high frequency word+accent pairs are produced less prominently than mean prominence. This emphasises again the importance of various aspects of context that can also influence pitch accent shape. This can be broader linguistic context such as discourse context (see the pre-study, section 5.1.3), phonological context such as the distance from a prosodic boundary, or segmental context, where tonal contours are influenced by micro-prosodic effects (e.g. Ladd et al., 2000; Jilka and Möbius, 2007).

5.2.6.2 Implications for exemplar selection

From an exemplar-theoretic perspective, the results of this study have some crucial implications. Firstly, the results and the exemplar-theoretic explanation lead to the conclusion that pitch accents, that is, the tonal contour on the word, are part of the exemplar and can therefore be stored in memory. Otherwise, frequency effects would not be expected. This is at odds with the assumption of autosegmental-metrical theories of intonation that the tonal contour is added by a separate component and that the placement and the realisation of pitch accents is solely rule-based.

Furthermore, the experiment demonstrated that there are frequency of occurrence effects depending on the frequency of occurrence of the word (incorporated as a combined frequency of occurrence of pitch accent and word). Hence, looking at exemplars in a word-based fashion is a reasonable approach. This is in a way not surprising since it seems reasonable to assume that for successful communication words are stored and form clouds of exemplars. The selection of production targets for an intended word would then be expected to happen from such a cloud, possibly restricted by other factors that form sub-clouds and further specify the properties the production target should have.

The most important implication, however, is that prominence, e.g. a tonal property of the exemplar, is considered in the selection process, as well. If exemplars can be selected according to their prominence, tonal properties are accessed in the selection process. More specifically: the linguistic function *prominence* is a relevant selection criterion in the

selection of production targets. Exemplars that match this criterion, that is, exemplars that are more prominent, have an increased pitch range[12]. Hence tonal properties must be accessible for the selection process.[13]

This argues not only for storage of F_0-contours, but also for a functional approach to prosody (see Batliner and Möbius, 2005, and chapter 3.3) in which the linguistic function *prominence* has a direct connection to the acoustic realisation: it is relevant for partitioning exemplar clouds into sub-clouds that are selected as production targets. These instances have certain acoustic specifics that are correlates of the linguistic function and that are accessible within the production perception loop of an exemplar-theoretic model. Over time, the cloud representing exemplars that match a certain linguistic function develops in a certain way, so that the exemplars match the linguistic requirements even better.

This connection between selection criteria and the development of a particular exemplar cloud – respectively a category – over time, has also been emphasised by Wedel (2006):

> Selection: Any factor that influences the likelihood that a given exemplar will participate in production or that influences the way a given percept is likely to be categorised will influence the direction in which the category system updates over time. Exemplars that are more 'fit' by these criteria will leave a greater trace in the future behaviour of the category than exemplars that are less fit. (Wedel, 2006, p. 253)

It also has to be noted that the idea of an entrenchment effect for highly frequent combinations to prevent excessive peak prominence has an implication for the selection process: following Pierrehumbert (2001) entrenchment can be modelled in an exemplar-theoretic model by assuming that it is not one exemplar that is (e.g. randomly) picked, but that a set of exemplars is selected and averaged over to construct the production target. Thereby, the respective value (in this case, pitch range) is caused to gravitate towards the mean of the selected set of exemplars. Even though there are other ways to model entrenchment, for the exemplar model incorporating intonation that will be outlined in section 6.3, averaging over a set of exemplars will be the mechanism suggested.

To conclude, with respect to the selection process in an exemplar-theoretic model, the results argue firstly for tonal properties to be considered in selection, and secondly for

[12] among other acoustic correlates of prominence such as increased intensity and lengthening

[13] Of course, one could also assume a less direct connection, where incoming exemplars are labelled as being *prominent* and then this label is accessed in selection. However, the conclusions are the same: tonal properties must be stored along with the exemplar.

selecting a set of exemplars rather than a single instance to form the production target for an intended new production.

5.2.6.3 Conclusions and further research questions

The results raise a few questions. First of all, increasing prominence with increasing frequency seems to be at odds with the idea of reduction for frequent units. There are findings on the segmental level documenting that frequent linguistic units are shortened and produced with less production effort (e.g. Bybee and Scheibman, 1999, and other studies described in chapter 2.1.3). So why are pitch accents enhanced instead of reduced as the exemplar (i.e. the combination of word+accent) becomes more frequent? The crucial difference between segmental and suprasegmental effects lies in the nature of pitch accents: whereas highly frequent words are often predictable and hence uninformative, the very reason to emphasise a word with a pitch accent is to draw the listener's attention to the word and to highlight the word. This underlying asymmetry between segmental and tonal events makes the two effects – segmental reduction and increased pitch excursion – fundamentally different. Moreover, as outlined in section 2.2.3 there is not only the aspect of reduction, i.e. reduced production effort to entrenchment, but there is also the idea of less variability as exemplars become entrenched. In other words, highly frequent pitch accented exemplars could be found to be enhanced in terms of their prominence but reduced in terms of lesser variability amongst exemplars of the same category.

So far, building on the 2 previous findings from experiment 1, experiment 2 leads to the following additional implications:

Implication 2: the context of a pitch accent is important; this includes discourse context as well as segmental context, i.e. the word on which the pitch accent was realised

Implication 3: the selection of production targets depends on a) word identity of the exemplar and b) tonal properties of the exemplar

Implication 4: in an exemplar model accounting for intonation, entrenchment has to be modelled, e.g. by selecting a number of exemplars and averaging over them to construct a production target

Questions about the nature of the selection process in an exemplar-theoretic model, however, remain. For instance, it is not clear yet, *how* the exemplars are selected. Is

the word the main decisive factor for selection? Or are exemplars selected according to the combination of word+pitch accent? That would imply that pitch accent categories are assigned to incoming exemplars as a label (Pierrehumbert, 2001), hence that they are categorically distinct and are classified after perception. Experiment 3, presented in the next section, will contribute to shedding light on the exemplar-theoretic selection process. Furthermore, it will investigate variability amongst pitch word+accent pair tokens to test for possible effects of entrenchment.

5.3 Experiment 3: Relative Frequency of Pitch Accent+Word

Experiments 1 and 2 demonstrated that pitch accents are subject to frequency effects – the frequency of pitch accent and information status category, and of accent+word pairs have been shown to influence pitch accent realisation. The next experiment tests if and how the acoustic realisation of pitch accent tokens on word+accent pairs is influenced by distributional factors.

Just as in experiment 2, the basic unit is again the word+accent pair. However, this time, not the absolute frequency but the relative frequency of a word+accent pair is looked at: the number of times a given word occurs with a specific pitch accent relative to all its pitch accented occurrences.

The acoustic parameter that is related to this relative frequency is again – similar to experiment 1 – the variability of pitch accent realisations. However, since this time word+accent pairs are looked at (as opposed to accent types as was the case in experiment 1), variability is calculated over all instances of a given combination of pitch accent and word (as opposed to over all instances of a particular accent – irrespective of the word). Thus, for each word+accent pair, a variability value is calculated measuring the variability among the tokens of this pair. In the statistical analysis, these variability values are then correlated with the relative frequency values of the word+accent pairs.

To understand the importance of relative frequency from an exemplar-theoretic perspective, it has to be noted that for a word+accent pair, the other instances of the same word with a different accent can be regarded as competing exemplars to the pair in question. Section 5.3.3 explains this idea and the hypotheses about exemplar selection derived from it in detail.

Section 5.3.1 describes the database – again news speech was analysed, however, this time the language under investigation was English, i.e. frequency of occurrence effects in another Germanic language could be examined. Sections 5.3.2 to 5.3.6 give an overview of the methodology employed in the analyses (how relative frequency and variability are calculated and which statistics were applied); the results are described in section 5.3.7. The last section (section 5.3.8) discusses the results and gives implications as well as open questions that link the work to the next experiment.

5.3.1 Data: Boston Radio News Corpus

The corpus used in this study is the Boston Radio News Corpus, a collection of radio news broadcasts (Ostendorf et al., 1995). A subset of the corpus has been labelled prosodically using ToBI (Silverman et al., 1992). The analysed data set contains approximately 1 h of speech from five professional speakers (3 female, 2 male). One speaker produced nearly half the tokens in the data set; however, as tokens from her were fairly evenly spread across the frequency bins (see below), it is assumed that this did not unduly influence results presented here.

The two most frequent accent types, H* and L+H*, were analysed. For each nuclear H* and L+H* accent (as well as the down-stepped versions of these accents), four parameters reflecting its shape were extracted using the PaIntE model (Möhler and Conkie, 1998, cf. section 3.2.2). Outlying tokens, i.e. tokens that fell outside the whiskers in a boxplot, hence that were more than 1.5 IQR away from the quartiles (see p. 102 for a detailed explanation), were removed.

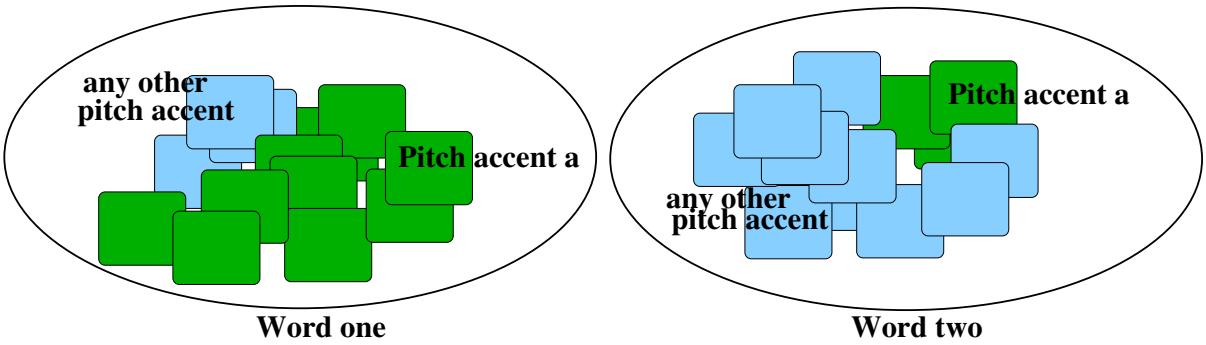
As the purpose of the study was to examine the variability between tokens of a word+accent pair, types with only one occurrence were not analysed so the final nuclear H* dataset comprised 1425 tokens and 465 types, and the nuclear L+H* set 306 tokens (123 types).

5.3.2 Calculating relative frequency

For each word type, the frequency of this word type with either H* or L+H* was calculated as was the frequency of the word with any pitch accent. The ratio of these two values measures the relative frequency of the word+accent pair:

$$h(w_a) = \frac{n(w_a)}{n(w_x)} \qquad , x \in Acc \tag{5.9}$$

where $h(w_a)$ is the relative frequency of word w and accent a, $n(w_a)$ is the absolute frequency of the combination of these two, and Acc is the set of accent types, hence $n(w_x)$ corresponds to the absolute frequency with which the w occurs with any pitch accent. Figure 5.12 illustrates two example cases of high and low relative frequency. The green boxes represent the cases of the word+accent pair under investigation, i.e. they correspond to $n(w_a)$. The blue boxes represent all the other instances of the same word w. Hence, both types of boxes – blue and green – taken together represent all instances where the word w occurred with a pitch accent and therefore correspond to



(a) high relative frequency (b) low relative frequency

Figure 5.12: Schematic depiction of (a) high and (b) low relative frequency cases. Green boxes represent instances of the respective word+accent pair, blue boxes stand for instances of the same word with any other pitch accent.

$n(w_x)$. The word "school" for instance, occurred 19 times with a pitch accent, in eight of these occurrences the accent was labelled as H*. Thus, the relative frequency of the word-accent pair "school+H*" is $h(school_{H*}) = \frac{8}{19} \approx 0.421$.

5.3.3 Relative frequency in an exemplar-theoretic model

As summarised in section 2.2.4, the exemplar selection process differs from model to model. Especially when attempting to incorporate tonal events into an exemplar-theoretic model, it is not clear how selection works.[14] As mentioned before (section 5.1.5.2), it seems unreasonable to assume that exemplars are selected as production targets solely on the basis of pitch accent type, i.e. that, when constructing the production target, any instances matching the intended accent are selected – even if they do not match the intended word. It seems more likely that the word restricts the set of exemplars that can be selected (because only exemplars of the intended word or even phrase are activated).

Experiment 2 demonstrated that tonal *properties* are likely to be considered in the selection process. It is, however, not clear whether tonal *categories* are a selection criterion. That is, it is not clear, whether exemplars of a word in its pitch-accented form (regardless of the pitch accent type) are activated or whether the appropriate combination of word+accent type is selected.[15] In other words: are pitch accents part of the

[14] Another open question that is not addressed by many exemplar-theoretic models is the question of the size of the underlying unit, stored in memory (i.e. phones, words, phrases, ...). But see Walsh et al. (2010) for a unified model that deals with different unit sizes.

[15] Since experiment 2 demonstrates the importance of tonal properties and indicates that prominence is a characteristic that is considered in the selection process, the assumption for experiment 3 is that

exemplars in the form of (categorical) labels, assigned to the exemplar in perception, that can be accessed in production?

From the perspective of exemplar-theoretic production, the relative frequency of a specific pitch accent type on a given pitch accented word can be seen as the ratio of "correct" production targets to "wrong" production targets. More precisely, with the definition given above (equation (5.9) in section 5.3.2), relative frequency is the number of times that the word occurs with the intended pitch accent (instances of the word with the "correct" pitch accent) related to the number of times it occurs with any pitch accent (all pitch-accented instances of the word). Note that here the notion of "correct" or "intended pitch accent" should be understood as "the pitch accent which is appropriate in the given linguistic context". Referring to an accent type as "correct" (or incorrect, respectively) should not be taken as starting out from the prerequisite that pitch accent labels are assigned and are used in the definition of the production goal. Pitch accents are rather regarded as a description (which might be categorical or not, cf. Batliner and Möbius, 2005) of the specific tonal contour, which is evoked by the linguistic context. The same linguistic contexts are assumed to evoke the same (or at least similar) tonal contours that would be labelled with the same ToBI category by human annotators. Therefore, if the contour is appropriate in the linguistic context, it should have the same pitch accent type as other exemplars in the same linguistic context. This pitch accent type is hence the "correct" or "intended" one for the respective production of the word.

The exemplars that do not match the intended pitch accents (that is, the ones represented as blue boxes in figure 5.12) can be seen as "competitors" to the intended pitch accent: they are realisations of the respective word where the word was realised with a different pitch accent. Relative frequency can therefore be interpreted as a means of encoding the interference of competing exemplars for the production of a word+accent pair. If, in production, the targets are selected according to the word *plus a specific pitch accent label*, then these competing exemplars should *not* have an influence on the production of the new target. If, however, *all* (pitch accented) instances of a word are selected, that is, if "wrong" competitors as well as "correct" instances are selected, an interference of the competitors with the correct production is expected.

the overall prominence of a word compared to its local context, that is its being pitch-accented or not, is accessible in the exemplar-theoretic prosody production process.

5.3.4 Balancing the data

The frequency distribution of the word-accent pairs is a Zipfian distribution, where few types occur often and many types occur rarely (Zipf, 1949). Due to this, the datasets are highly unbalanced with respect to both the number of tokens per type and the number of types that occur with a given frequency. Therefore, three different data reductions were carried out to create more balanced versions of the dataset. The original datasets, will be referred to as *HN* (nuclear H* accents) and *LHN* (nuclear L+H* accents), henceforth.

HNmod and LHNmod One reduced version of the datasets excludes types which were the only ones occurring with a specific frequency. For instance, in *HN*, only one type ("Melnicove+H*") occurred 14 times. These tokens were therefore excluded. This reduction affected the very rare high-frequency types that were likely to influence the regression line in the statistical analysis massively. For H*, this applied to 3 types: "Melnicove" (occurring 14 times), "WBUR" (occurring 15 times) and the word "says" (58 times). For L+H*, only one type had to be excluded ("one", 8 times). The modified datasets are referred to as *HNmod* and *LHNmod* henceforth.

HNbal A different data reduction aimed to balance the data with respect to the number of high- and low-frequency types. A balanced dataset for nuclear H* (*HNbal*) was created, where some of the low frequency types (with only two or three tokens) were randomly excluded so that the number of low frequency types and higher frequency types (with 4 or more tokens) was the same. This was done to prevent the many low-frequency tokens from outweighing effects of the fewer tokens with higher frequency. For L+H* no such analysis could be carried out, since this set would have consisted of only 32 types, and would thus have lacked sufficient statistical explanatory power.

HNequ For the last data reduction, the data was divided into frequency sets, i.e. sets of word-accent types occurring with the same frequency. Then an equal number of types (10) was randomly selected from each frequency set. From each of these types, an equal number of tokens (2) was randomly selected to create the new dataset *HNequ*. It has to be noted that this equalisation process left only a small subset of the original dataset: only frequency sets 2-6 comprised enough different word-accent types for 10 to be selected. The new dataset comprised only 196 types (and 392 tokens).

5.3.5 Calculation of pitch accent variability

To measure the variability amongst tokens of word+accent pairs, first, 4 PaIntE parameters (see section 3.2.2) were calculated for each accent from the smoothed F_0 contours.[16] The parameters encoded the accent's peak alignment (*b*), the range of the rise and fall (*c1,c2*), as well as the height of its peak (*d*). Unlike in experiment 1, the methodology was refined in that the *a* parameters (gradient of the rise and fall, respectively) were excluded from the analysis. The reason for this is that in cases where only one sigmoid is used, the *a* value of the other sigmoid can be regarded as meaningless, since there is no slope (cf. section 3.2.2.3). For *c* parameters, this is not the case, because the value for the sigmoid that is not used is set to 0, which reflects the actual properties of the accent: if there is for instance only a falling part in an accent, it is reasonable to set the range of the rising part to zero. It does, however, not make sense to set the gradient of the rise to $a1 = -1$. Therefore, only the other four parameters were included in the analysis. Another option would have been to exclude tokens for which the approximation uses only one sigmoid. But not only would this have resulted in having very little data (and the amount of data is already not sufficient for some analyses, see above), it would also have meant excluding a significant portion of pitch accent tokens that have a specific shape. Consequently the actual data would not have been represented properly.

5.3.5.1 Outlier Removal and Normalisation

Analogous to the preceding experiments, for each PaIntE dimension and each accent type, outliers were removed so that values did not fall outside the whiskers in a boxplot (see again p. 102 for a detailed explanation).

As in experiment 2, the PaIntE parameters were z-scored for each speaker and accent separately according to equation 5.5 (page 103).

5.3.5.2 Euclidean Distance

To measure the variability of the tokens of an accent+word pair (e.g. all exemplars of the word "school" where "school" is accented with an H* accent), each of the accent tokens was represented as a four-dimensional vector (one dimension for each z-scored PaIntE parameter). Figure 5.13 illustrates this in the two-dimensional space: 3 accent tokens

[16] F_0 contours were estimated from the speech signal with the ESPS get_f0 algorithm; smoothing was done using smooth_f0 from Taylor et al. (1999)

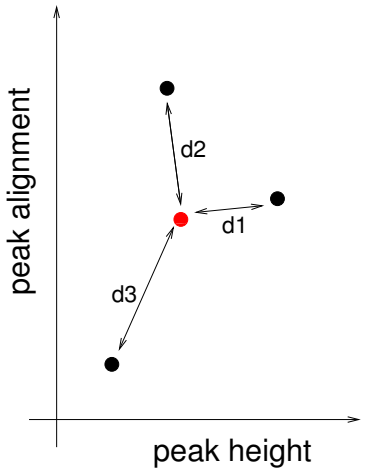


Figure 5.13: Illustration of three vectors and their centroid in a two-dimensional space. The vectors are marked as locations in the space (black dots). The vector composed of the mean values for each of their dimensions is the centroid (red dot). The distance between two vectors gives an indication of how similar the two vectors are. The bigger the distance between two points, the smaller the similarity. To measure variability within the set of these three vectors, the distance between each vector and the centroid is measured ($d1, d2$ and $d3$). The average of these three values then indicates the variability in the set.

are represented as vectors with two dimensions, e.g. pitch height and peak alignment. The three resulting vectors are marked by black dots.

For a given word type and a given accent, the average distance between the tokens of this word+accent pair in the four-dimensional space was calculated. This was done by first calculating a centroid (i.e. a vector composed of the mean values of each dimension). For that, for each of the four PaIntE dimensions, the mean value was calculated according to

$$\bar{p}_{dim} = \frac{\sum_{i=1}^{n} p_{dim,i}}{n} \qquad , dim \in Painte := \{b, c1, c2, d\} \tag{5.10}$$

where p_{dim} stands for the respective PaIntE dimension, $\bar{p}_{dim}$ is the mean value and n is the number of tokens on the basis of which the average is calculated (e.g. the number of instances of "school" occurring with H*: 8 instances). This four-dimensional vector is then the centroid for the analysed type, that is the centroid of the respective word+accent pair ("school"+H*). In figure 5.13 the centroid for the three vectors is marked by a red dot.

The Euclidean distance d between the centroid $\bar{x}$ of a word+accent pair and a token x_i of that type was then calculated according to equation (5.11):

$$d(x_i, \bar{x}) = \sqrt{\sum_{dim \in Painte} (x_{dim,i} - \bar{x}_{dim})^2} \tag{5.11}$$

The average distance of all tokens of a type to their centroid gives a measure of the variability of the type.

That is, for the two-dimensional example in figure 5.13, the distances between each of the three vectors and their centroid (marked as $d1, d2$ and $d3$) would be calculated. The average of these distance values would then give an indication how variable the set of three accents is.

In the case of the word "school" in the *HN* dataset, where the centroid represents eight instances of "school+H*", the distance of each of the single tokens to the centroid in the four-dimensional space was calculated and the average distance of those eight distances is interpreted as the variability value of the type "school+H*".

5.3.6 Statistical Testing

Linear regression models were fitted to assess the relationship between the relative frequency of a word+accent pair and the variability of the tokens of this type (there were no other variables in the models). This was done for all the datasets described in section 5.3.4.

For the two datasets that involved random selection of data-points, HNbal and HNequ, a multiple-fold cross-validation was carried out to ensure that the randomisation did not distort the result. For HNbal, the random set was validated in a 100-fold cross-validation. For the drastically reduced random dataset HNequ the statistical analysis was repeated 1000 times.

5.3.7 Results

Table 5.9 gives an overview of the models' results. For both analysed accent types, the regression models for the complete dataset (*HN* and *LHN*) yielded p-values indicating a significant correlation between the relative frequency of a word+accent type and the variability among the tokens of this type. This effect also held for the modified datasets *HNmod* and *LHNmod* where extremely frequent types were removed.

Figure 5.14a illustrates the effect for the dataset *HN*. Each point in the graph represents a word+H* type (e.g, "school+H*"). The variability of the tokens of the respective type is plotted against its relative frequency. The regression line falls from the left hand

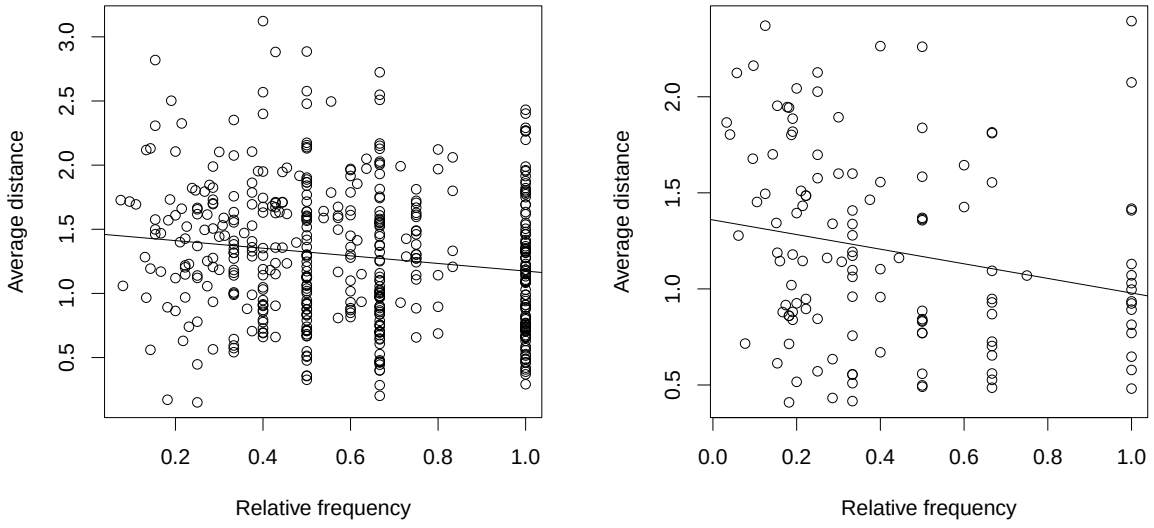
	HN	*LHN*	*HNmod*	*LHNmod*
p-value	< 0.01	< 0.05	< 0.01	< 0.05
coefficient	-0.2967	-0.3804	-0.2901	-0.3637
std. error	0.0918	0.1591	0.0922	0.1591

(a) Datasets that did not involve randomisation

	HNbal	*HNequ*
repetitions	100	1000
significance	79%	7%
tendency	7%	5%

(b) Randomised datasets

Table 5.9: Overview of the linear regression results for the tested datasets. For datasets that were created by random selection of subsets of the data, the regression was repeated. The number of repetitions is given as well as the proportion of significant cases ($\alpha = 0.05$) and of cases that showed a tendency ($p < 0.08$).



(a) word+H* pairs

(b) word+L+H* pairs

Figure 5.14: Relationship between the relative frequency of word+accent pairs and the variability among the tokens of each of the types.

side of the graph, where the low relative frequency pairs are located, towards the right hand side, where the high relative frequency pairs can be found. This illustrates a decrease of variability with increasing relative frequency. Figure 5.14b demonstrates that the same effect can be observed for L+H* (dataset *LHN*): the greater the relative frequency of a word+L+H* type, the lower the variability of this type.

For H*, where it was possible to balance the dataset (so that the same number of tokens per type was analysed) and still keep a reasonable number of tokens, the random selection of a smaller number of low-frequency types (i.e. the generation of the dataset *HNbal*)

and the subsequent calculation of the model was done 100 times. It yielded significant correlations between relative frequency and variability in 79 of 100 repetitions. In all the significant cases, variability of the types decreased with increasing relative frequency, just as in the graphs displayed in figure 5.14.

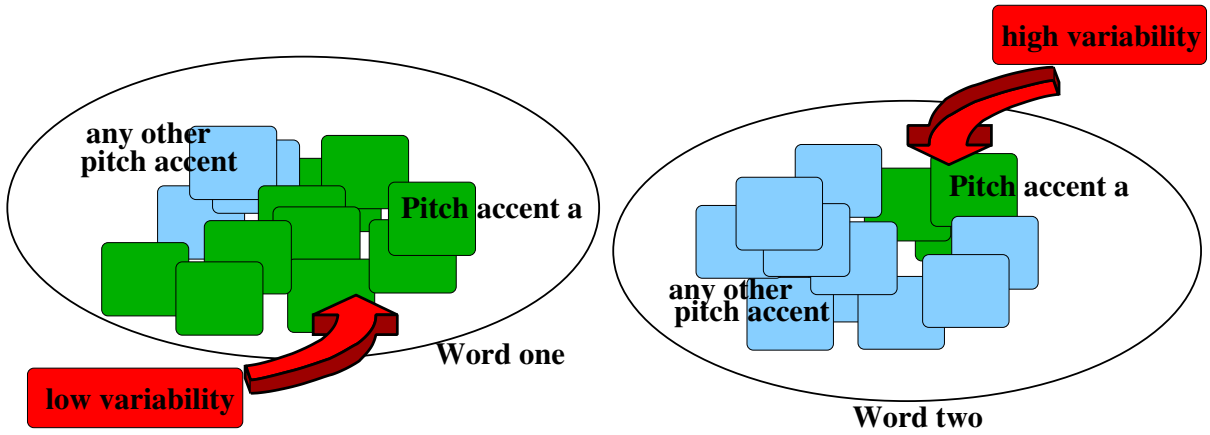
For the fully equalised set *HNequ*, however, the effect did not hold. Since the selection of a subset of the original dataset involved two randomisation processes (first the random selection of two tokens per type and second the random selection of 10 types per frequency bin), the number of repetitions was increased to 1000 repetitions. The regression model was significant in only 66 cases. It has to be noted though that in those significant cases the effect was the same as for the other datasets: decreasing variability for increasing relative frequency. In addition, this dataset is reduced drastically, compared to the original set. While the original *HN* dataset comprises 1425 tokens of 465 types with tokens ranging in their frequency between two and 58 (though only one type is as frequent as this), the reduced dataset *HNequ* consists only of 196 types and only two instances of each type went into the analysis. Furthermore, the types ranged in their frequency only between two and six, i.e. the higher frequency types were not analysed. It is therefore debatable whether such a drastic reduction of the data still represents the original dataset in an appropriate way.

5.3.8 Discussion

The aim of this experiment was to test if and how the acoustic realisation of pitch accent tokens on word+accent pairs is influenced by the relative frequency the pair. Since it was argued that relative frequency encodes the interference of competing exemplars in production (section 5.3.3), the experiment also aimed to shed light on the exemplar-theoretic selection process.

The analyses described above indicate a relationship between the relative frequency with which a word occurs with a certain pitch accent and the variability of the realisations of this accent. The greater the relative frequency of a word+accent pair, the less variable the realisations of the accent. This result is schematically depicted in figure 5.15.

The following section gives an exemplar-theoretic interpretation of the result. The findings have some important implications for models of prosody production as well as, specifically, for the selection process in exemplar-theoretic models. Likewise, it raises some questions for further investigation. Both implications and further questions are discussed.



(a) high relative frequency (b) low relative frequency

Figure 5.15: Schematic depiction of the result for (a) high and (b) low relative frequency cases. With decreasing relative frequency of the combination of word and accent, the realisations of the accent (depicted as green boxes) become more variable, since the competing exemplars (depicted as blue boxes) interfere with the production of the intended accent.

5.3.8.1 Exemplar-theoretic interpretation of the result

The analyses indicated that with increasing relative frequency of a word+accent pair, the variability of the realisations of the accent decreases. The exemplar-theoretic explanation for this effect is as follows: In production, exemplars of the target word are activated (all boxes in figure 5.15(a) or (b), respectively). If many of these exemplars are accented with pitch accents other than the intended one (i.e. many exemplars come from linguistic contexts that differ from the current one), then there are more competitors to the intended tonal contour (cf. section 5.3.3). These competitors with deviating pitch accents are depicted as blue boxes in figure 5.15.

If all pitch-accented instances of a word are selected during production, then for word+accent pairs that have many competitors (such as in figure 5.15b), the set of tonal contours from which the production target is derived is more variable, because it includes not only realisations with the intended accent but also realisations with competing pitch accents. The competing realisations interfere and thereby cause "noise" in the production of the intended accent. Hence, the larger the proportion of competing pitch accents, the greater the variability among the realisations of the target accent.

For words that are often realised with the same accent, however, i.e. where the proportion of competitors is smaller (such as in figure 5.15a), the realisation of the pitch accent on this word is less variable. That is because the set of F_0-shapes from which the

production target is derived is less variable, since it includes mainly realisations with the intended accent.

In other words, word+accent pairs with a small relative frequency (more competitors) are more variably realised than word+accent pairs with a high relative frequency (less competitors).

Here, it is worth mentioning again the results from Luce and Pisoni (1998; cf. section 2.1.3) on lexical access: words that have many similar neighbours are identified less accurately. Likewise, words with high frequency neighbours are more often reported as non-words than words with low frequency neighbours. Thus, words with many competitors were classified worse than words with few competitors. This result can be explained assuming that similar pronunciations are activated similarly strongly in perception. If then a fixed number of exemplars or a fixed neighbourhood is activated, the competing (similar) exemplars would be expected to cause noise in classification. Other production studies found an influence of competing contexts (Jurafsky et al., 2001) and words (similarly pronounced words; Wright, 1997) on segmental production: the fewer competitors, the greater segmental reduction.

5.3.8.2 Implications

As mentioned before, frequency effects on tonal events generally present a challenge for traditional, autosegmental theories of intonation (cf. section 3.1). These theories are silent on the effect of frequency on pitch accent type and realisation. A relationship between the word types and the accent types they occur with, would not be predicted. Neither would an influence of the relative frequency on the actual acoustic realisation of pitch accents be expected. It is instead assumed there that, once a particular accent type has been assigned to a given word, the realisation of the pitch contour on that word is related purely to the phonological context, e.g. the position in the phrase and how near other accents are (Ladd, 2008). While these factors are undoubtedly still relevant, the results presented here show that the frequency of word+accent pairs also plays a part in explaining pitch contour variation. Along with the results from experiment 1 and 2, the findings suggest that pitch contour realisation cannot be considered to be purely post-lexical in English.

An exemplar-theoretic view of pitch accenting, on the other hand, expects word-(possibly also phrase- or multiple-word-)based storage of accent contours, and resultant frequency effects. The findings from this experiment therefore speak in favour for an

exemplar-theoretic model of prosody production with storage of fundamental frequency along with the exemplar.

Another aspect of the result is its implications for exemplar selection. As described in section 5.3.3, an influence of competing exemplars on the production of the intended accents points to a model in which the selection of production targets is based on words, not on pitch accent types. If pitch accent *type* would be considered as well, and if therefore, during selection, only exemplars with the "correct" accent type would be activated, then the competitors (that is, exemplars with a different accent type) could not influence the new production. However, since in the present experiment they were found to do so, one can conclude that also exemplars with other pitch accent types are activated, and that consequently pitch accent *type* does not seem to be a selection criterion. It therefore seems likely that pitch accent type is not explicitly assigned to the exemplar as a label.[17] This would suggest that what is being stored are rather the actual pitch *contours* than abstract pitch accent *categories*.

This view is supported by functional approaches to intonation models such as Batliner et al. (2001) and Batliner and Möbius (2005) which argue for the omission of an intermediate, categorical level and instead assume "a functional representation of the positions of accents and phrase boundaries" (Batliner et al., 2001, p. 2288). Such an approach is, firstly, highly compatible with an exemplar-theoretic model, and secondly, it argues for the idea that the selection takes place over all *pitch accented* instances of a word, regardless of the *type* of accent: pitch accenting signals several linguistic functions and is therefore considered in the selection process. This is supported by the result of experiment 2, where the linguistic function "prominence" or its tonal properties, respectively, was shown to be relevant in exemplar selection.

5.3.8.3 Conclusions and further research questions

In conclusion, building on the 4 previous findings from experiments 1 and 2, experiment 3 leads to the following additional implication:

Implication 5: pitch accent type is not a selection criterion and probably not assigned to the exemplar; what is being stored are pitch contours rather than categories

[17] Of course the type could be assigned but then not used in production. However, principles of economy make this option unlikely – why should something be assigned as a label but not be used in later processing steps?

Open questions, however, remain: The results of this experiment indicate that the *word* is an important unit in pitch accent production. Additionally, along with the results of experiment 2 they demonstrate a relationship between the segmental and the suprasegmental level. More precisely, the experiments demonstrate a dependency between acoustic realisations of pitch accents and a combination of the segmental level (the word) with the suprasegmental level (the pitch accent type). This raises the question whether there is a dependency between the suprasegmental level and the segmental level alone (as opposed to a combined level). In other words: does word frequency or the frequency of word sequences have an influence on the realisation of pitch accents? Are there also effects on the tonal context? These questions will be addressed in the following experiment.

5.4 Experiment 4: Relative Word Frequency

The experiments presented so far demonstrated various frequency effects on pitch accent realisation. Pitch accent type frequency, information status type frequency, the frequency of the word+accent pair and the relative frequency of such a pair have been shown to influence the acoustic realisation of pitch accents. The analyses on word+pitch accent pairs demonstrated that there is a dependency between the word level and the tonal level – this is highly surprising if one follows the common assumption that accenting is solely based on top-down information and is realised after accessing the lexicon to derive the correct word form (see Ladd, 2008). The previous experiments, however, all incorporated word frequency as a function of pitch accenting - either the absolute frequency of word+accent pairs was investigated (experiment 2) or the relative frequency of those pairs (experiment 3).

The last experiment presented in this thesis goes a step further: it looks for a dependency between the word level alone (without being related to a tonal event) and the tonal level. More specifically, the analysis investigates possible frequency effects of words on the *prosodic* properties of these words (see section 2.1.3 for literature on frequency effects on segmental properties). Since the experiments so far demonstrated the importance of linguistic context for exemplar-based pitch accent production, experiment 4 investigates effects of the relative frequency of word sequences, as opposed to single, isolated word tokens.

Much of language is made up of word sequences that have a high relative frequency, or collocations (Erman and Warren, 2000), e.g. *I don't know* or *make a mistake*: phrases

where the component words occur together more frequently than would be expected by chance. Within the framework of Exemplar Theory, these frequent phrases are assumed to be stored as single exemplars in memory (Bybee, 2006). As the production of frequent exemplars becomes entrenched over time, they become less variable (Pierrehumbert, 2001; Bybee, 2006) in their segmental properties. However, the detailed *prosodic* properties of such frequent lexical sequences have not been investigated, yet[18] – even though it has been argued that *idioms* are expected to occur with a narrow range of tonal contours (Bolinger, 1985).

The experiments presented so far indicate that the production of intonation operates – at least in some cases – in a usage-based fashion and argue therefore for the integration of intonation into exemplar-based production models. Hence, intonation is assumed to be stored as part of the exemplar, and to be accessible in the selection process during production.

If the acoustic information stored with exemplars includes prosodic information (e.g. the pitch contour and/or the pattern of accents and phrasing), then lexical sequences that have a high relative frequency in their context, i.e. that are highly probable, are expected to exhibit less prosodic variation than less probable sequences, just as they exhibit less phonetic variation. The experiment presented here was designed to test this assumption. Prosodic variation was examined on two levels: pitch accent variability (similar to experiments 1 and 3) on a pitch accented word, and the variability of the prosodic context of a word.

Specifically, the experiment aimed to answer two questions: Does the variability of a pitch accent contour on a given word decrease as the relative frequency of the word in its lexical context increases? And does the same hold for the variability of the prosodic context of this word?

As the results from experiment 3 indicate that pitch accent *type* might not be a necessary source of information within an exemplar-theoretic prosody production model, the larger Switchboard database could be used (Godfrey et al., 1992) which is only annotated for accent *position* and phrase boundaries (Calhoun et al., 2010). This database has the additional advantage that, other than the databases used in experiments 1-3, it is not news speech, but conversational speech. So the experiment also complements the analyses presented here with an investigation of exemplar-theoretic effects in another (probably more natural) speaking style.

[18] But see e.g. Pan and Hirschberg (2000), and other studies summed up in section 4.2.2, for the relationship between relative frequency (within a bigram) and the presence of a pitch accent: the lower the predictability of the word, the more likely it carries a pitch accent.

Section 5.4.1 describes the database. In section 5.4.2 the calculation of relative frequency is defined and section 5.4.3 explains how the two measures of prosodic variability were calculated. The statistical analysis is explained in section 5.4.5, and section 5.4.6 describe its results. Section 5.4.7 discusses the implications for a model of pitch accent production.

5.4.1 Data: Switchboard

The Switchboard corpus consists of a collection of spontaneous telephone conversations between American English speakers (Godfrey et al., 1992). 76 conversations, or around 6h of speech from 114 speakers, are annotated for pitch accent and prosodic boundary location using the ToBI (Beckman and Hirschberg, 1999) standard (Calhoun et al., 2010). Pitch accent type is not marked.

Prosodic realisation For the two analyses presented here, two different datasets were extracted.

1. For the investigation of pitch accent variability, those trigrams were extracted that occurred at least *4 times with an accent on the middle word* in the prosodically annotated set. 95 trigram types met this requirement. However, they varied in their token frequency (from 4 to 55 tokens), so 100 datasets were created where for each trigram type 4 tokens were randomly selected in order to balance the design.

2. For the examination of prosodic pattern variability, those trigrams were extracted that occurred *at least 4 times* in the prosodically annotated part of switchboard (3705 tokens, 124 word types in 541 trigram types).

Lexical frequency To calculate relative word frequency, lexical frequencies for words and trigrams were extracted from the whole Switchboard corpus, along with the Callhome American English corpus (Kingsbury et al., 1997), a smaller corpus of spontaneous telephone conversations. The combined corpus comprised just over 3 million words. Trigrams containing fillers like "uh-hum" or incomplete words were not included for the calculation of lexical frequency.

5.4.2 Calculation of relative word frequency

The relative word frequency, that is the relative frequency of a word in its lexical (trigram) context was calculated in a similar fashion as relative word+accent pair frequency in section 5.3.2.

For each trigram, its frequency was divided by the number of times the middle word occurred (in any trigram context), thus the measure denotes the probability of a word given its neighbours:

$$P_{Lex}(l_i w_i r_i) = \frac{n(l_i w_i r_i)}{n(l w_i r)} \tag{5.12}$$

$P_{Lex}(l_i w_i r_i)$ is the probability (i.e. the relative frequency) of the word w_i given its left neighbour l_i and its right neighbour r_i. $n(l_i w_i r_i)$ is the number of times the trigram $l_i w_i r_i$ was found in the combined Switchboard and Callhome corpus, and $n(l w_i r)$ is the number of any trigrams in the corpus where w_i was the middle word.

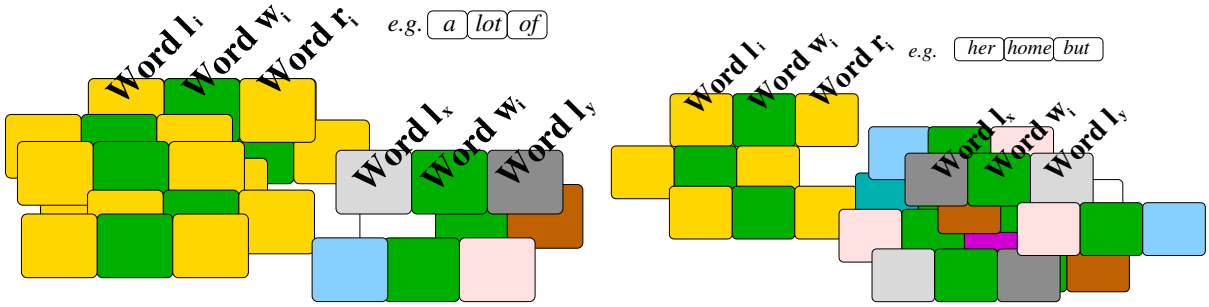
Figure 5.16 schematically depicts a case of high and a case of low relative word frequency. The green boxes represent one word, occurring in different contexts. Cases where the context words are depicted as yellow boxes illustrate the trigram for which the relative frequency is calculated. Other trigram contexts with the same middle word are depicted as boxes in another colour. Whereas in the left hand panel (a), which represents a case of high relative word frequency the proportion of trigram tokens with the same (yellow) context is high, it is low in the left right hand panel (a), which illustrates a low relative frequency case.

The probability measure accounts for the variability of lexical contexts. The greater the value of P_{Lex}, the less likely the word occurs in other contexts, i.e. less opportunity for diverse lexical contexts. For example, the word *lot* occurs 10059 times in Switchboard and Callhome. Of these, it occurs 6631 times in the trigram *a lot of* and 3428 times in other lexical contexts, yielding the relative frequency of *lot* (given its left neighbour *a* and its right neighbour *of*) $P_{Lex}(a\ lot\ of) = \frac{6631}{10059} \approx 0.66$.

Table 5.10 lists the trigrams with the highest relative word frequency in Switchboard and Callhome. It is not surprising that most of them would be considered collocations in English.

5.4.3 Calculation of pitch accent variability

For each trigram that occurred at least 4 times with an accent on the middle word, pitch accent variability was calculated. The calculation of pitch accent variability is similar



(a) high relative frequency (b) low relative frequency

Figure 5.16: Schematic depiction of (a) high and (b) low lexical probability/relative frequency cases. Green boxes represent instances of the middle word. Yellow boxes represent a particular trigram, coloured boxes represent any other contexts of the respective middle word.

Table 5.10: List of the trigrams with the highest relative word frequency of the middle word in combined Switchboard and Callhome.

P_{lex}	trigram	P_{lex}	trigram
0.74	the rest of	0.40	a nursing home
0.66	a lot of	0.40	as soon as
0.61	i grew up	0.28	i don't know
0.54	as far as	0.28	i used to
0.54	be able to	0.27	it seems like
0.49	a couple of	0.27	a little bit
0.48	to worry about	0.22	i ended up
0.43	a matter of	0.20	look forward to

to the methodology employed in experiment 3, but now, a slightly more sophisticated method was employed, which did not exclude the PaIntE-parameters *a1* and *a2*. The argumentation for the new method is as follows: Recall that in experiment 3, the *a* parameters were excluded, because in those cases where the PaIntE function used only one sigmoid to approximate the accent's shape, they are assigned a dummy value (-1) for the other sigmoid which is not meaningful (cf. section 5.3.5). Also recall that this problem does not apply to the c-parameters, because they are assigned a value of $c1 = 0$ or $c2 = 0$, respectively, which is reasonable: a non-present slope has a range of zero. In experiment 3, therefore, the a-parameters were excluded from the analysis of variability.

However, one could argue that by leaving out these parameters, important information is lost: accents where only one sigmoid is used (that is, accents that only have a rise

or a fall, but not a pronounced peak), are fundamentally different in their shape from accents that are described best by two sigmoids, namely accents that have both a rise and a fall. In fact, it could be argued that the difference on the a-dimension of two such accents – a one-sigmoid one and a two-sigmoid one – is a categorical one: in one case, the slope exists, in the other one, it does not.

Therefore, in experiment 4, the methodology of measuring pitch accent shape variability has been adapted, so that for those cases where one of the accents being compared has a normal, meaningful *a* value and the other one has no *a* value for the respective sigmoid, the difference between the *a* values was not measured on a continuous scale but rather on a categorical scale. That is, the methodology encoded the difference between "existent" and "non-existent". This difference should be the same for all cases where a non-existent *a* parameter is compared to an existent one. Following this reasoning, the distance between such two *a* values was represented by a constant, encoding a large difference between the two values.

To achieve this, changes had to be made to both the normalisation of the PaIntE values, and the calculation of the average Euclidean distance accounting for the variability among accent tokens. The modified methodology is described in detail below.

5.4.3.1 Outlier removal and normalisation

Analogous to the preceding experiments, for each PaIntE dimension and each accent type, outliers were removed (see p. 102) and the PaIntE parameters were z-scored for each speaker and accent separately (equation (5.5) on p. 103).

Since in the present experiment the *a*-parameters were analysed as well, the calculation of the z-scores for *a1* and *a2* was modified: when calculating the z-scores for the respective dimension, those cases where the value was a (meaningless) dummy value were not taken into account .

5.4.3.2 Euclidean distance

The calculation of the average Euclidean distance was also modified to be able to deal with the cases in which one of the *a* parameters had a non-meaningful value.

Firstly, a pairwise comparison was carried out where each token is compared to each other token of the same trigram type by measuring their distance in the Euclidean space (figure 5.17 illustrates this approach in the two-dimensional space). The average distance of all comparisons for a trigram type is then the variability value for this type. That is, unlike in experiment 3, there was no centroid that was calculated and compared to each

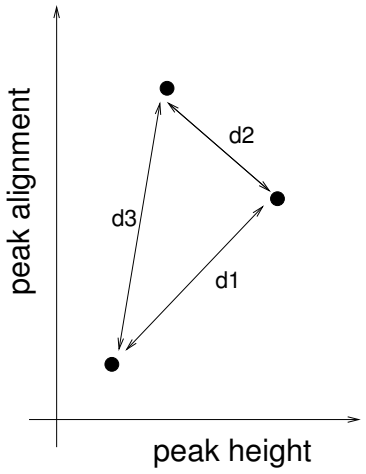


Figure 5.17: Illustration of three vectors in a two-dimensional space. The vectors are marked as locations in the space (black dots). The distance between two vectors gives an indication of how similar the two vectors are. The bigger the distance between two points, the smaller the similarity. To measure variability within the set of these three vectors, the distance between each pair is measured ($d1, d2$ and $d3$). The average of these three values then indicates the variability in the set.

token. The reason for this is that it does not make sense to calculate a centroid for all tokens with the a parameters included in the analysis (and thus the "dummy-values", which do not encode a meaningful number, but rather a categorical difference to accents which were assigned a real, meaningful value).

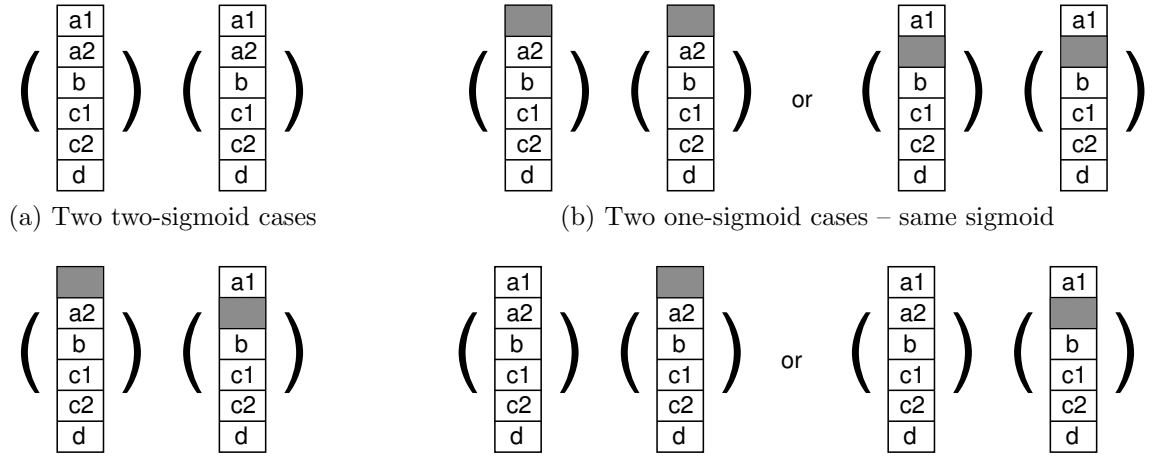
Secondly, comparing the vectors directly to each other made it possible to adapt the calculation of the Euclidean distance depending on what kind of tokens were compared:

Two two-sigmoid cases In the default case, that is the case where both accents were approximated with two sigmoids (as illustrated in figure 5.18a), the Euclidean distance between those two accent tokens was calculated according to the following formula

$$d(x,y) = \sqrt{\sum_{dim \in Painte} (x_{dim} - y_{dim})^2} \qquad , Painte := \{a1, a2, b, c1, c2, d\} \tag{5.13}$$

Where $d(x, y)$ is the Euclidean distance between the vectors x and y.

The same formula was used in experiment 3 (equation (5.11), p. 138), but there the Euclidean distance between a pitch accent token and the centroid was measured; and the vectors were only four-dimensional since the a parameters were excluded from the analysis.



(a) Two two-sigmoid cases

(b) Two one-sigmoid cases – same sigmoid

(c) Two different sigmoids

(d) One one-sigmoid case and one two-sigmoid case

Figure 5.18: Possible types of combinations when calculating the Euclidean distance between two accent tokens. In the PaIntE-approximation, each accent token can either be approximated by two sigmoids, by a falling sigmoid, or by a rising sigmoid. In the two-sigmoid-case, all PaIntE-values are assigned a meaningful value, in the one-sigmoid case, one of the a-dimensions is assigned a dummy value. The calculation of the Euclidean distance between the two respective tokens varies for the different combinations.

Two one-sigmoid cases – same sigmoid For cases where both accents had been approximated with the same sigmoid, that is cases where in both accents the same a parameter was assigned the dummy value (cf. figure 5.18b), the distance in the a-dimension is zero, since in both cases the same sigmoid was missing. The other values were summed up according to formula (5.11).

Two different sigmoids or one one-sigmoid and one two-sigmoid case For those cases where the accents were either approximated with two different sigmoids (as illustrated in figure 5.18c) or where one of the two accents was approximated with two sigmoids and the other one with only one sigmoid (as illustrated in figure 5.18d), the difference between the meaningless a value and its meaningful counterpart in the other accent's vector was set to a constant.

The constant was determined by calculating the ranges r_1 and r_2 of all meaningful (z-scored) a values for each of the a dimensions. For a1, the range was $r_1 \approx 22$, for a2, $r_2 \approx 19$. The constant was chosen so that it was higher than the largest measurable difference; it was set to the constant value $c = 25$.[19]

[19]However, it has to be noted, that for middle range distance value ($c = 6$) the results remain the same, so changes in variability are not only due to the one-sigmoid cases.

The values for the other PaIntE dimensions (b, c1, c2 and d) were summed up according to formula (5.11).

After determining the pairwise distances for each analysed type in the way described above, the distances between all pairs were summed up and an average was calculated. Analogous to experiment 3, this average distance of all pairwise comparisons gives a measure of the variability of pitch accents on the middle word of the trigram type: tokens with more similar accent contours have a smaller average distance.

As mentioned in section 5.4.1, this procedure was applied to 100 randomly selected datasets that were balanced for the number of tokens of each analysed trigram type. Hence, for each type, 4 tokens form the basis for the average Euclidean distance.

5.4.4 Calculation of prosodic context variability

To capture the prosodic context of a word, a prosodic pattern was determined for each trigram token. To this end, each word in a trigram was classified as being accented or not, and as carrying a boundary or not. Then, the prosodic pattern of the token was represented by the sequence of classifications of the three words. For example, for the trigram *a lot of*, if there was an accent on *lot* and a boundary after *of*, the prosodic pattern of the word sequence *a—lot—of* was encoded as *NoAcc-NoBound—Acc-NoBound—NoAcc-Bound*.

To encode the variability of the prosodic context of a word, for each middle word of a trigram, the probability of it to occur with the given prosodic pattern was determined, comparable to to the calculation of its relative word frequency (which denotes the probability to occur in the given lexical context): the number of times the word w_i occurred with the particular pattern p_i was divided by the number of times the word occurred at all:

$$P_{pros}(w_i, p_i) = \frac{n(w_i, p_i)}{n(w_i)} \tag{5.14}$$

For instance, of the 100 tokens of *lot* in the prosodically annotated part of Switchboard, 48 were with the prosodic pattern $p1$ (*NoAcc-NoBound—Acc-NoBound—NoAcc-NoBound*) yielding a probability value of $P_{p1} = \frac{48}{100} = 0.48$.

Then, for each trigram type, the likelihood of the most dominant prosodic pattern in which the middle word occurs, was encoded. That is, the probability value of the most probable prosodic pattern for the middle word was calculated.

This value reflects the cohesion between the word and the most dominant prosodic context it is realised in. Thus, it gives an indication of how variably the word's prosodic context is realised. If the probability value is high, a large proportion of the trigram tokens in which the word occurs is realised with one prosodic pattern. Thus, a high value indicates a strong coupling between the word and its prosodic context and hence low variability. A low value on the other hand indicates that there is greater variability: the most dominant context does not occur often around the word, and all the other patterns have even smaller values, so there are more different types and no other type is dominant.[20] For example, consider a word type W which occurs in 12% of the cases with prosodic pattern P1. There is no other prosodic pattern with which W occurs more often, so P1 is the most dominant prosodic context for W. Consequently all the other prosodic patterns in which W occurs, cover a proportion of W tokens which is smaller than 12%. Hence, for a given word, the number of prosodic pattern types is larger if the probability value for the most dominant type is small.

To conclude: a high probability value of the most dominant prosodic context indicates a strong coupling of word and prosodic pattern, a low value indicates greater variability. Consequently, the suggested probability scale is proportional to the cohesion between word and prosodic context, and it is *inversely proportional* to the variability of the prosodic context.

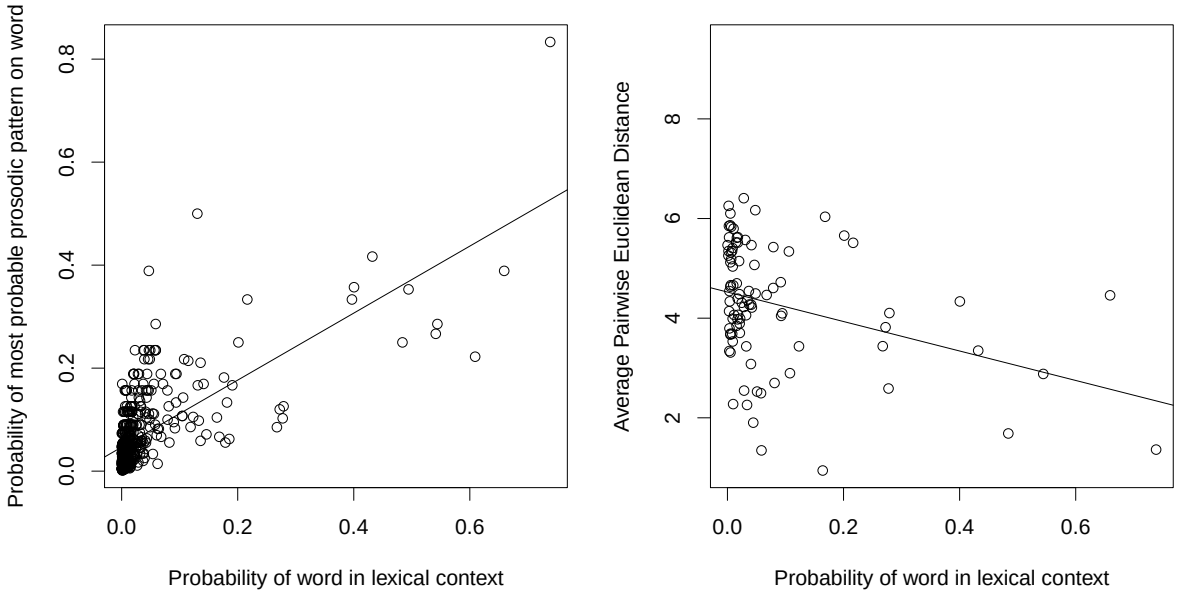
5.4.5 Statistical Analysis

Linear regression models were fitted to assess the relationship between the relative frequency of a word in its trigram context and the prosodic variability of the respective word. Variability was represented as a) variability of the prosodic context of the word, and b) variability of the pitch accent tokens on the word in a given trigram.

The relative frequency of a word in a given trigram context was tested as a predictor of these two indicators of prosodic variability.

For pitch accent variability, where 100 randomised datasets with equal number of tokens for each type were created, a regression model was fitted for each dataset, to predict pitch accent variability by the probability of the word in its lexical context. The

[20] This measure of prosodic context variability does not account for whether the remaining tokens are realised with many or few prosodic pattern types: Consider two word types W1 and W2. Both word types occur in 70% of the cases with prosodic pattern P1. W1 occurs with a second pattern P2 in 30% of the cases, the remaining 30% of the realisations of W2 are realised with 10 different patterns. The measure suggested here would give the same measure of variability for both types.



(a) prosodic context variability (b) pitch accent variability

Figure 5.19: The relative frequency of a word in its trigram context plotted against prosodic variability. In (a) the y-axis displays the probability of the most dominant prosodic pattern around the word. A high probability reflects low prosodic context variability. With increasing relative frequency, the probability of the pattern increases, hence the prosodic variability of the words' context decreases. In (b) the y-axis displays the average Euclidean distance of pitch accent tokens on the middle word. A high average distance reflects high variability among the tokens. With increasing relative frequency, the variability decreases, as well.

number of times the test yielded significance as well as the direction of the effect was kept track of.

5.4.6 Results

Prosodic context variability The regression model predicting the probability of a word occurring in its most dominant prosodic context by the relative frequency of the word in its lexical context, yielded a significant p-value of $p < 0.001$ ($Adj.R^2 = 0.48$).

Figure 5.19a displays the result. Each point in the graph represents a trigram type. The probability of the middle word in the given lexical context is displayed on the x-axis, the probability of the middle word's most dominant prosodic context is shown on the y-axis. As can be seen, with increasing relative frequency of a word in its lexical context, the probability of its most prominent prosodic context, increases as well. Since the probability of the most dominant prosodic context is inversely correlated to the

variability of the prosodic contexts of the word, in general, the graph illustrates how prosodic context variability decreases with increasing relative lexical frequency.

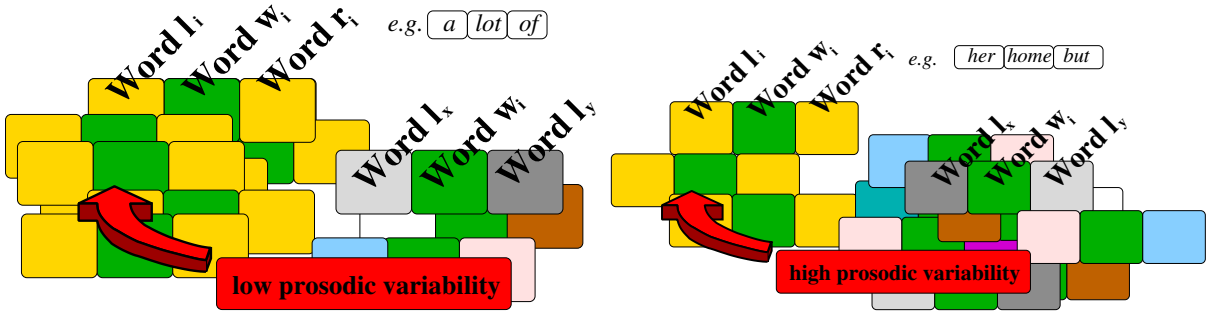
Pitch accent variability The regression models predicting pitch accent variability by the relative frequency of the word in its lexical context for the randomised datasets yielded significant p-values ($\alpha = 0.05$) in 80% of the cases, and in 7 cases there was a tendency ($p < 0.08$). Figure 5.19b displays the result for a sample dataset (again each point represents one trigram type): with increasing relative frequency of the word in its lexical context, the average distance (i.e. the variability) of pitch accent tokens on this word decreases. That is, the more probable a word in its context, the more similar the realisations of pitch accent tokens on the trigram. This relationship (a negative correlation) was the same in all significant regression models. As can be seen in the figure, the correlation is not a strong one ($Adj.R^2 \approx 0.11$), however the models yield significance in a vast majority of cases.

Summary To conclude, for both parameters measuring prosodic variability on or around the word, respectively, a negative correlation of the relative frequency with which the word occurs in its lexical context and the prosodic variability was found: with increasing relative frequency, the prosodic variability decreases. Figure 5.20 illustrates this result.

It is worth mentioning that in none of the models the correlation is a strong one, indicated by the dispersion in the graph and the relatively low R-squared correlation coefficient, nevertheless, there is a significant relationship between the respective two factors, with variability decreasing as lexical relative frequency increases. Given that different parameters are expected to influence variability, as well (for instance, syntactic properties of the context), the change that can be ascribed to frequency of occurrence is expected to be subtle.

5.4.7 Discussion

The experiment tested the assumption that sequences that occur together relatively often, that is, sequences were the middle word has a high relative frequency in its given trigram context, display less prosodic variation. The acoustic realisation of such sequences has been shown to be less variable on the segmental level (e.g. Bybee, 2006; Wright, 1997) and, in the framework of Exemplar Theory, it has been argued that such collocations are stored in memory as single units. As the results from experiments 1–3 presented here indicate that pitch accents are subject to frequency of occurrence effects



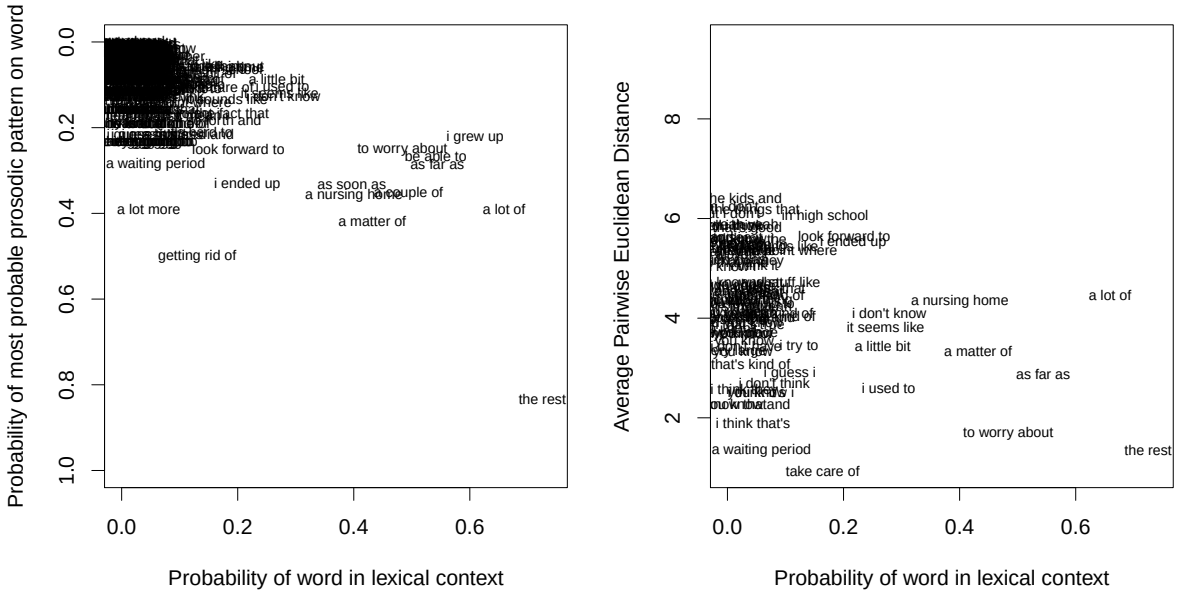
(a) high relative frequency (b) low relative frequency

Figure 5.20: Schematic depiction of (a) high and (b) low lexical probability/relative frequency cases. Green boxes represent instances of the middle word. Prosodic variability was found to be negatively correlated with relative frequency, hence for high relative frequency cases (that is, cases where the word occurs relatively often in the respective trigram context) prosodic variability was found to be low. With decreasing relative frequency, prosodic variability was found to increase.

and might be stored along with the exemplar, the current experiment was set up to test whether prosodic variation decreases with increasing relative frequency of a word in a given lexical context. Such a dependency between distributional properties of words and the tonal realisation present a challenge for traditional models of intonation that assume that the production of pitch accents is post-lexical and operates as an autonomous component in speech production (e.g. Ladd, 2008, and see Pierrehumbert, 2001).

Both analyses confirmed the hypotheses based on exemplar-theoretic expectations. The results showed that words that have a high relative frequency in their context display less prosodic variation than words that occur in a less fixed context. The first analysis demonstrated that such words are more likely to occur with a certain prosodic pattern. Consequently there is less probability mass for other prosodic patterns to occur around the word. The second analysis confirmed that pitch accent contours on the accented form of such words are less variable with increasing relative frequency of the word in the trigram context. The observed decrease in variability might come from a decrease in the number of accent types on the word or from more similar realisations of the same pitch accent type. Since the dataset is not annotated for accent type, these different cases could not be distinguished. However, the conclusions are the same in both cases: higher lexical relative frequency leads to lesser tonal variability.

Figure 5.21 illustrates the result by giving the trigram types employed in the two analyses. To make the two graphs easier to compare, the scale for the graph displaying prosodic pattern variability (sub-figure 5.21a) has been inverted, so that the graph de-



(a) prosodic context variability (b) pitch accent variability

Figure 5.21: Trigram types in the graphs displayed in figure 5.19. There is an overlap in the trigram types demonstrating that the trigrams where the middle word is relatively frequent given its lexical context, are realised less variably in both analysed aspects of variability: the prosodic context of the middle words is less variable, the same is true for the pitch accent shapes on it.

picts the decrease in variability. As can be seen, the trigram types are similar (obviously prosodic pattern analysis can contain more types since the trigrams were not restricted to ones with an accent on the middle word as was the case for pitch accent variability analysis). The most probable type ("the rest of") is in both cases also the least variable one. The other types vary in the degree of measured variability, but overall the two graphs demonstrate that there is an overlap of trigram types that display less variability on both aspects of variability: pitch accent shape and prosodic context.

5.4.8 Implications

The relationship between the lexical level and the tonal level revealed by the two analyses makes it unlikely that prosodic realisation is solely determined by a combination of 'top-down' syntactic, semantic and pragmatic factors (e.g. given/new status), and the phonological context (e.g. how close together accents are). While these factors are undoubtedly still relevant, the results presented here show that in collocations, at least,

stored information about the prosodic patterns and accent realisation of that collocation also play a role.

As mentioned above, within an exemplar-theoretic framework, frequent collocations are expected to be stored as exemplars along with phonetic and contextual information. As is shown here, this includes the prosodic realisation in the form of pitch contours. In production, the exemplars containing a word serve as production targets. If most of these come from the same lexical context, then there should be less variability amongst them than if they come from a variety of contexts. Hence, productions of that word should show less variability. For example, 66% of *lot* tokens come from the collocation *a lot of*. If these collocations are realised with little variability, because they are stored together and entrenched, then productions of *lot* overall should show less prosodic variability than words occurring in a trigram context that is less fixed. This was confirmed in the above analyses where both the accent realisation of, and the prosodic pattern around, words like *lot* in *a lot of* were found to vary less.

The correlation between accent contour variability and lexical probability was much lower than for prosodic pattern variability. This could be because the entrenchment effect is weaker for accent realisation, or because the pitch accented datasets for the second analysis were too small to illustrate the strength of the effect.

In conclusion, this experiment corroborates and extends experiments 1–3, which all demonstrated exemplar-theoretic effects on the realisation of pitch accents. It also demonstrates that the word level is important for pitch accent production and that tonal properties have a word-specific aspect reminiscent of the effects of "word-specific phonetics" reported by Pierrehumbert (2002). Hence, one might even speak of "word-specific prosody" which adds to other, known effects of linguistic top-down information influencing prosodic realisation. autosegmental-models of intonation could not explain such word-specific tonal effects that go beyond the assignment of weak or strong word stress (cf. section 3.1.3).

5.4.9 Conclusions and further research questions

The analyses revealed an effect of relative lexical frequency on prosodic realisation. Table 5.10 demonstrated that trigrams in the middle and higher relative frequency bands could be considered as collocations. Nevertheless, in cases of very low relative frequency of the middle word in the given lexical context, the trigrams were less coherent. For instance, the least probable trigrams were trigrams like "walk the streets", "what you see" or "rock and roll", but also trigrams that are less syntactically coherent, e.g. "kids

and i" ore "that as being". Hence, the variability in cases of very low relative word frequency (with relative frequency values of $P_{Lex} < 0.00001$) could also be ascribed to less syntactic coherence in those cases. In future studies, the analysed tokens could be controlled for syntactic features or part-of-speech tags of the trigram tokens.

To conclude, building on the 5 findings from experiments 1 to 3, experiment 4 leads to the following additional implication:

Implication 6: frequency of occurrence of the lexical context influences tonal variability: the intonation of word sequences can be entrenched

Open questions remain. For example, how do other factors interact with frequency of occurrence? And which factors are most relevant for changes in variability? Are there frequency of occurrence effects for more contextual factors such as syntactical constructions, emotional states or speech acts? Concerning specifically experiment 4, it would also be interesting to test different approaches to identify collocations (Erman and Warren, 2000); the one suggested here (trigrams where the middle word is probable given its neighbours) is a simple one, more sophisticated approaches might yield stronger results.

Moreover, even though a larger database was used, the sparseness of the data is still a problem: there are few frequent types and many infrequent ones. Therefore, a single type which is very frequent influences the analysis more than a single infrequent type. Both analyses presented here relied on manually annotated data. If automatic approaches to intonation annotation (e.g. A. Schweitzer, 2010; Rosenberg, 2010; Fernandez and Ramabhadran, 2010) were used, larger databases could be analysed to get a clearer picture of frequency of occurrence effects on intonation.

6 General Discussion and Outlook

The experimental work presented here sought to answer the question whether there are exemplar-theoretic effects in intonation and to investigate potential effects and their implications for existing exemplar-theoretic models.

Experiment 1 gave evidence that pitch accents are subject to frequency of occurrence effects, generally supporting the idea of an exemplar-theoretic approach to intonation. Experiment 2 showed that the sensitivity to frequency of occurrence is also present when word+accent pairs are analysed. The results from this experiment can be explained with the exemplar-theoretic production-perception loop. Experiment 3 then demonstrated a relationship between the relative frequency of word+accent pairs and the acoustic realisation of accents – the results can be explained with the selection process and give more insight into the criteria that are relevant for selecting exemplars to construct a production target. Finally, experiment 4 demonstrated that pitch accents and prosodic context are sensitive to lexical probability, which speaks for lexical entrenchment of intonation.

In sum, the experiments support the idea of Exemplar Theory and argue for an integrated approach, where intonation is part of an exemplar-theoretic production model.

The following sections offer concluding remarks about the main findings. Firstly implications of the experiments with respect to exemplar-theoretic storage will be given (section 6.1). Then the selection process as it is indicated by the findings presented here will be outlined (section 6.2). Here, properties of exemplars that are relevant for selection will be in focus. Finally, section 6.3 outlines the essentials of an exemplar-theoretic model that incorporates intonation.

6.1 Pitch Accents and Exemplar-Theoretic Storage

With regard to the question of whether exemplar-theoretic effects can be found in pitch accents, the main findings of the work presented here are, firstly, that pitch accents are subject to frequency of occurrence effects (all experiments), secondly, that they can

undergo lexical entrenchment (experiment 4), and, thirdly that they are influenced by the frequency of their linguistic context (experiments 1 and 4). Each of these three main findings will be discussed in the following sections.

6.1.1 Frequency Effects on Pitch Accents

Each of the four experiments presented here revealed that pitch accents are subject to frequency of occurrence effects on various levels. As has been outlined in chapter 2, usage-based models of language and speech production and perception provide a framework that does not only deal effectively with frequency effects, but that is built upon the idea that the number of instances of linguistic units significantly affects the way these units are produced. By assuming un-normalised storage of linguistic units as exemplars, distributional properties of language are taken to be reflected in a speaker's memory. Consequently frequency effects are expected in such a framework. Unlike models of language acquisition that solely rely on the acquisition of rules and parameters, in exemplar-theoretic systems, frequency effects have not to be dealt with by introducing counters that keep track of how often a unit was perceived and/or produced.

Naturally, only properties of linguistic units that are stored with the exemplar can display frequency effects and hence, if a frequency effect for a particular linguistic or non-linguistic property is observed and explained within the exemplar-theoretic framework, this argues for the storage of the respective property with the exemplar.

Experiment 1 showed that the acoustic realisation of pitch accent tokens is related firstly, to how often a pitch accent type (i.e. a particular tonal contour) occurs, and secondly, to how often an information status category (a feature of the linguistic context) occurs. If both of these types of information are assumed to be stored with the word, the effect can be explained in an exemplar-theoretic model. Therefore, the experiment argues for lexical storage of pitch accents in context.

Experiments 2 and 3 demonstrated that the frequency of word+accent pairs is also a factor determining pitch accent shape. This argues for storage of intonation contours along with the word. Furthermore, it reveals a dependency of tonal parameters on lexical items which is unexpected if one follows the traditional, autosegmental view of pitch accents to be assigned post-lexically.

Finally, experiment 4 demonstrated that the word level is even more influential on pitch accent production: the (relative) frequency of words affects the variability of pitch accent tokens on them. Hence, the purely lexically (and not tonally) based frequency of words and word sequences is related to their prosodic realisation. Tonal variability was

found to decrease with increasing relative frequency, in other words, the more common a particular sequence of words, the less variable their tonal realisation: they tend to occur with similar tonal patterns. This is in accordance with the exemplar-theoretic assumption that acoustic detail is stored and that therefore, tonal contours should be stored along with lexical items. The finding supports a view of what one might call "word-specific prosody": word sequences can have a specific prosody, thus, not only the lexical item is stored, but also a specific tonal realisation. This idea adds tonal realisation to the idea of "word-specific phonetics" which has been coined by Pierrehumbert (2002).

6.1.2 Lexical Entrenchment of Intonation

Experiment 4 found that the variability of pitch accents on the middle word of trigrams is related to the relative frequency with which the trigram occurs in its lexical context: the higher the frequency, the smaller the variability. In addition, words that tend to occur often in the same trigram context, also occur in less variable prosodic patterns. Together, these findings demonstrate a decrease of prosodic variation with increasing collocation strength of the trigram.

On the segment level, words that have a high relative frequency in their context have been found to be more reduced (e.g. Wright, 1997; Jurafsky et al., 2001). This reduction is taken to be one effect of entrenchment. Entrenchment is the phenomenon that highly learned units are produced with less production effort and more precision. Whereas less production effort would be expected to be reflected by phonetic reduction (as has been found in the mentioned studies) more precision is manifested by less variability. Hence, the main finding of experiment 4, namely that collocations are realised with less prosodic variation, shows that intonation is subject to lexical entrenchment. Therefore, the experiments presented here argue not only for the storage of F_0-contours along with the word and other properties of the linguistic unit, but also for the joint storage of fixed word sequences (i.e. sequences that occur together frequently). It seems reasonable to assume that these sequences can vary in their length and that their storage depends on how often the particular combination of words occurs together. As a consequence, an exemplar model should be able to deal with different unit sizes, as is for instance the Multi-Level Exemplar Model (Walsh et al., 2010, see section 2.2.3, p. 43).

6.1.3 Frequency of linguistic context

Experiment 1 and experiment 4 both explicitly tackled the question whether the frequency of different linguistic contexts has an influence on pitch accent realisation. In experiment 1, properties of the discourse contexts were analysed by investigating frequency effects of two information status categories. In experiment 4 probabilistic factors of the word contexts were related to prosody by testing for effects of the relative frequency of the word given its neighbours. Both experiments demonstrated a dependency between context frequency and pitch accent shape. Experiment 4 also related prosodic context variation to the relative frequency of the word. These findings have some implications for the question of which information is part of the stored exemplar. Whereas some researchers argue that everything is stored with little if any information loss and without any abstraction or categorical labelling (e.g. Wade et al., 2010), others assume that there is information which is stored as raw information (e.g. duration or formant frequencies) and that there is information which is stored as a "label" with which the exemplar is marked (Pierrehumbert, 2001).

The results from experiments 1 and 4 imply that both, discourse context and lexical context can be part of the information that is being stored – otherwise no frequency effects could be observed for these parameters.

Moreover, the results from experiments 1 and 2 could be interpreted as an indication of any linguistic context influencing pitch accent realisation. The reasoning for this is as follows: Experiments 1 and 2 showed that the frequency of occurrence of a pitch accent type influences the acoustic realisation of accent tokens. Following a more functional approach to intonation, as for instance suggested by Batliner and Möbius (2005; cf. section 3.3), pitch accents can be considered to encode different linguistic functions. The pitch accent label is then just a description of the tonal contour. That is, the pitch accent describes the actual acoustics, which are correlates of different context configurations. Thus, the accent reflects the linguistic context. For instance, an accent could encode a context such as "a word that is given in the discourse, located at the end of a question". If accents are taken to reflect specific contexts, the frequency of a pitch accent type can be regarded as encoding the frequency of a specific context. The results from experiment 1 and 3 can then be reformulated: the frequency of occurrence of linguistic contexts affects pitch accent realisation. Consequently the properties of those contexts must be stored along with the exemplar, as well.

Note that the findings from these two experiments could also be explained by assuming that pitch accent category labels are assigned to the exemplar and that the frequency

with which a pitch accent category occurs, has an influence on the realisation of the accent token. However, experiment 3 indicated that what is stored is the actual pitch contour and not a pitch accent category (cf. also section 6.2). Therefore, all experiments considered, a more functional viewpoint seems to fit the results best.

6.2 (Properties of) Pitch Accents and the Selection Process

Experiment 2 demonstrated that frequent pairs of pitch accent+word have a greater accent range than less frequent ones. This finding can be explained by the exemplar-theoretic production-perception loop (cf. section 5.2.6.1): to produce a pitch accented word, prominent exemplars are selected as production targets. The new exemplar is consequently prominent, as well. Due to the storage of self-produced exemplars, the number of prominent exemplars increases as the exemplar cloud grows. Hence, as the number of exemplars increases, the accents are realised with greater accent ranges.

This implies that the linguistic function *prominence* is relevant to the selection process: exemplars that fulfil the selection criterion of being *prominent*, are selected. Other exemplars that, for example due to other contextual factors or because of extralinguistic reasons, are realised with reduced peaks, are not selected to construct the production target. Accent range, as one of the correlates of prominence, is therefore accessed during the selection process.[1] For this to be possible, fundamental frequency has to be stored along with the exemplar. And, moreover, this information must be accessible in the selection process. As highlighted by Wedel (2006), selection criteria are expected to influence the development of an exemplar cloud, over time (cf. section 5.2.6.2).

Even though these effects are very subtle, and the increased pitch ranges differ only slightly, it does not seem reasonable to assume that the development of the respective exemplar clouds will go on infinitely. Even an assumed slow ongoing language change would be expected to stop, eventually. Therefore, a mechanism modelling entrenchment of the effect is needed: following Pierrehumbert (2001), entrenchment can be modelled by selecting a set of exemplars and averaging over it to construct a production target as opposed to the random selection of a single instance. Hence, the selection process in an exemplar-theoretic model that incorporates intonation should be *averaging over several exemplars* when constructing production targets.

[1] Or, a label for prominence is attached as exemplars are perceived, during the classification process, cf. footnote on p. 128.

While experiment 2 demonstrated that *prominence* is a selection criterion, experiment 3 indicated that *pitch accent type* is not: Pitch accent+word pairs that occur together relatively frequently, viz. where the word occurs rarely with a different pitch accent, are realised less variably than pairs with a lower relative frequency. Relative frequency of pitch accent+word pairs reflects the number of competitors to an intended production – instances where the word was accented with a different pitch accent. These instances seem to interfere in the production of the word with the intended accent, resulting in greater variability of the pitch accent tokens on the respective word.

An influence of competing exemplars in production can only be explained by assuming that these competing exemplars are selected to participate in the construction of a production target. Therefore, the selection mechanism does not seem to select according to pitch accent category.[2] The results rather suggest that many *pitch accented* instances of the given *word* are selected, regardless of the pitch accent type. This is in keeping with the finding from experiment 2 which indicated that *prominence* is a selection criterion. Following Batliner and Möbius (2005) one can argue that *pitch accent category* as such is not considered in selection because it is not a linguistic function. Prominence, however, is. Moreover, other linguistic functions, for instance *givenness*, *focus* and other functions as suggested by Batliner and Möbius (cf. section 3.3.1) are then assumed to be taken into account in the selection process.

6.3 Incorporating Intonation in Exemplar Models

Taken together, the findings and implications of the experiments presented here specify some criteria that an exemplar-theoretic model that also accounts for intonation, has to meet. An example model is outlined in the following. The description of the model is not assumed to be exhaustive; it is rather supposed to present a starting point for further research and simulations.

The following criteria have to be met:

1. **Exemplars contain F_0:** As explained in detail in the previous sections, the findings from all four experiments argue for the storage of fundamental frequency as part of the exemplar. This implies that properties of the pitch contour are dimensions in the multidimensional space in which the exemplars are stored.

[2] Or, alternatively, the category labels are inaccurate or not useful.

2. **(Discourse) context is stored:** Discourse context and probably other linguistic context is assumed to be stored (experiments 1 and 4). The experiments here do not make assumptions about how it is stored. Contextual properties could be assigned as categorical labels during perception, that is the exemplar could be indexed with that information (cf. e.g. Pierrehumbert, 2001; Hay and Bresnan, 2006). Alternatively, the characteristics of the context could be implicitly there if the model looks at long stretches of raw speech (e.g. Wade et al., 2010, however, here the size of the context window would have to be increased, see section 2.2.3).

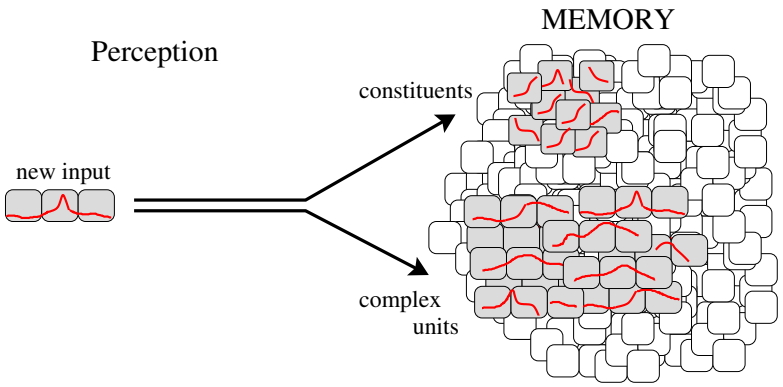
3. **Complex units can be stored as one exemplar:** Since the intonation of word sequences can be entrenched (experiment 4), the model must enable lexical collocations to be stored as one unit. Therefore, the model should provide alternative processing routes, which allow it to flexibly choose between processing the unit and constituent level as is suggested in the Multi-Level Exemplar Theory (Walsh et al., 2010).

4. **Selection considers word identity and prominence:** The findings from experiments 2 and 3 indicate that prominence and word identity are selection criteria. Supposedly, the selection takes place over pitch accented instances of words. The model should therefore store this kind of information (or derive it from the signal) and consider it in the selection process.

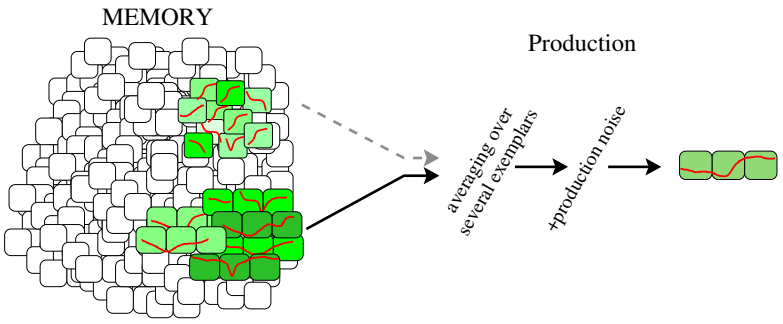
5. **Bias for greater accent ranges:** Experiment 2 indicated that there is a bias towards greater pitch accent ranges as a unit becomes more frequent. Hence, the model should simulate a production bias for greater accent ranges.

6. **Averaging over a set of exemplars:** Experiment 4 found entrenchment of word sequences. Moreover, the production bias (see above) should be counterbalanced. Therefore, the model should average over a set of exemplars when constructing the production target to model entrenchment.

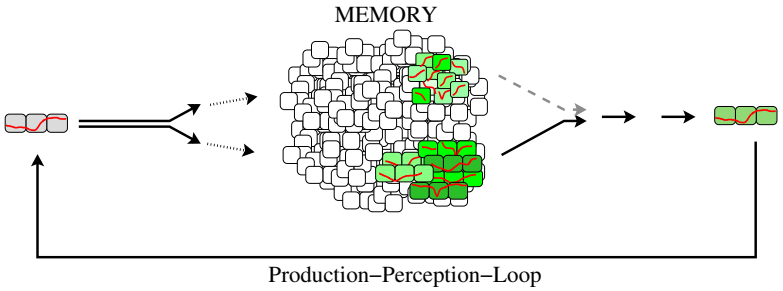
The example model outlined in section 2.3 (figures 2.2 and 2.3) has to be slightly modified to fulfil these criteria. First of all, fundamental frequency has to be part of the exemplar representation. Moreover, in perception, a complex input should be compared to both exemplars of the complex unit (if existing) and exemplars of the underlying smaller units. Figure 6.1a schematises the two alternatives.



(a) New input is stored in memory. The percept is compared to both exemplars of the unit and exemplars of its constituents.



(b) In production, a set of exemplars which meet the selection criteria is activated. For complex units, the model decides between complex and simplex exemplars as proposed by Walsh et al. (2010). In the figure, a case where the complex unit was realised frequently enough to be chosen, is displayed. The selected set is averaged over to construct a production target. Production noise is added and a new exemplar is produced.



(c) Self-produced exemplars are stored in memory, as well. Thereby, the tonal production biases found in experiment 2 can be explained.

Figure 6.1: Novel exemplar-theoretic model of speech production and perception incorporating intonation. Fundamental frequency is part of the exemplar representation.

The new exemplar is stored close to similar exemplars. In doing so, tonal features are included in the calculation of similarity. Here, a model like the one suggested by Walsh et al. (2010) could benefit from the additional features: their model calculates the similarity scores for the set of complex exemplars and for the sets of decomposed ones. If the activation within the complex exemplar cloud reaches a certain threshold, the unit is judged as grammatically valid and is stored as a complex exemplar along with the correct decomposition into simplex exemplars (cf. section 2.2.3). If tonal features are part of the exemplar, they can be employed in calculating the similarity scores and can thereby enhance the naturalness of the judgement.

Analogously to Walsh et al.'s (2010) model, if a cloud of valid complex units is found, both representations are stored, the complex and the decomposed one. If not, only the decomposed one is stored. The stored exemplars are indexed with categorical information about information status, accent status and prominence and probably with more labels describing the linguistic and extra-linguistic context such as information about speaker identity, speaking situation, emotion, etc.

In production, the model activates a set of exemplars that meet the selection criteria – for instance, word identity, accent status and information status. These criteria could be task-dependent. For complex units, the model should be able to choose, possibly depending on frequency of occurrence, between selecting the exemplar cloud formed by exemplars of the unit as a whole or by the multiple clouds formed by the decomposed sub-parts.[3] Walsh et al.'s (2010) model offers the possibility to do so: the activation for both sets of exemplars is calculated. If there are enough complex exemplars so that their joint activation based on similarity reaches a certain threshold, the production target is constructed from the complex instances. Otherwise, it is constructed from the several sets of simplex instances. Figure 6.1b displays the case were enough similar complex exemplars are found. The construction target in the model suggested here is then constructed based on these exemplars by averaging over them. Thereby entrenchment can be modelled. Production noise is added and the new exemplar is produced.

The self-productions are perceived and stored in memory (cf. figure 6.1c. In this way, production biases can be explained. Since fundamental frequency is assumed to be part

[3]Ideally, the model should be able to choose for any unit size, if the unit is produced holistically or if it is broken down. That is, the unit size could be a phrase, a word sequence, a word, a syllable or a segment. The decision unit could decide based on similarity and frequency of occurrence. Potentially it could also incorporate other aspects in the calculation of the best unit size, comparable to how it is done in unit selection synthesis, where "concatenation costs" are weighed up against similarity scores to find the most appropriate unit.

of the exemplar representation, tonal production biases like the one observed in experiment 2, can be modelled.

6.4 Outlook

This section provides a brief overview of open questions and possibilities for future work. There are questions that should be investigated on speech corpora, but also other types of experiments such as production or perception studies would be desirable to fully understand the mechanisms involved in intonation production. In addition to that, a detailed specification of an exemplar model that incorporates fundamental frequency could be obtained by running computational simulations.

Other types of experiments The evidence for frequency of occurrence effects on pitch accent realisation presented here is purely corpus-phonetic. Analysing speech corpora has some advantages: firstly, frequencies of tonal units can be retrieved, secondly, relatively large amounts of data can be analysed and thirdly, natural speech is analysed. The results from corpus experiments enable researchers to draw conclusions about production but very little can be said about perception. In any case, it is desirable for both research on production and perception to use the corpus studies as a starting point for further experiments. For the production side, effects found on a corpus might be even greater if the data is controlled for other context and thereby noise in the data is reduced. Also, controlled laboratory experiments can contrast influencing factors explicitly to give insight into a hierarchy of aspects that influence pitch accent realisation. The backside of the coin is that production studies have the disadvantage of comprising speech which is relatively artificial, so a combination of corpus studies and controlled laboratory experiments is desirable. Perception experiments could give a better insight into the perception of tonal units. For instance, are the observed effects (the production bias in experiment 2 or changes in variability in the other experiments) perceivable? Furthermore, psycholinguistic experiments could provide insight into lexical access of pitch accented words and word sequences. Lexical access is facilitated by familiar tonal contours (Braun et al., 2011). Are words or word sequences accessed more easily if a frequent combination of words and accents is used? To provide a full specification of an exemplar-theoretic model incorporating intonation, such additional experiments to examine the interplay between the tonal and the segmental layer are crucial.

Other corpus experiments However, there are additional aspects, that can and should be investigated on corpus data: Since the results indicate that a functional view on intonation is firstly desirable and secondly highly compatible with an exemplar-theoretic model of intonation, it would be interesting to test if other linguistic functions are subject to frequency of occurrence effects and if they are also selection criteria in exemplar-theoretic models accounting for intonation. The catalogue provided by Batliner and Möbius (2005) suggests different kinds of linguistic functions that could be tested. In doing so, larger datasets should be employed. Since the experiments presented here suggest that tonal contours rather than categorical pitch accent labels are stored, speech corpora could be processed automatically and be annotated for pitch accent placement (e.g. A. Schweitzer, 2010; Sridhar et al., 2008; Nenkova et al., 2007). Then, PaIntE parameters could be used to describe the contour. An independence from manually-labelled data would solve many of the described sparse-data problems.

Computational simulations The exemplar-theoretic model which has been outlined in section 6.3 is based on existing exemplar-theoretic models. It is very similar to the model proposed by Walsh et al. (2010). A computational simulation of the observed effects is expected to provide further insights into how such a model has to be specified in detail and is thus a main goal for future work.

A usage-based grammar of tonal contours Although the experiments argue that in some cases "word-specific prosody" should be an option, it seems unavoidable that a "grammar" of tonal contours emerges from the exemplar-based storage of intonation. Approaches to exemplar-theoretic syntax and semantics argue that a grammar can emerge by detecting analogies between new input and existing exemplars (Bod, 1998, 2009; Bybee and Eddington, 2006). Calhoun and A. Schweitzer (forthcoming) found evidence that collocations of tonal contours in combination with lexical items can spread their meaning. Therefore, in future work it would be desirable to examine longer stretches of speech and their tonal realisation as a function of frequency of occurrence to obtain a comprehensive picture of exemplar-based intonation production.

7 Conclusion

The objective of this thesis was to bridge the gap between research on exemplar-theoretic phenomena on the one hand and research on intonation on the other. The two areas were, up to now, to some extent separate from each other since their basic assumptions are in a way at odds with each other: Exemplar Theory assumes episodic storage of highly detailed stretches of speech in a speaker's memory – which basically corresponds to the assumption that "everything" is in the lexicon (and rules are learned by detecting analogies in the stored instances), however, the basic assumption of the most widespread intonation models, is that in Germanic languages intonation is assigned post-lexically according to top-down information.

To investigate if exemplar-theoretic effects can be found in intonation, four corpus experiments sought to examine potential frequency of occurrence effects on pitch accent realisation. Specifically, the following questions were answered:

Does frequency of occurrence affect pitch accent realisation? Frequency changes of different linguistic units were investigated, such as pitch accent type, information status type, the combination of pitch accent type and word type, and the frequency of the word type alone. The experiments demonstrated that changes in frequency of occurrence on all those levels correlate with changes in the tonal realisation: pitch accent shape, pitch accent variability and the tonal context of word tokens has been shown to be sensitive to frequency of occurrence. The main findings were

1. Pitch accent variability is influenced by
   - a – the frequency of occurrence of pitch accent type
   - b – the frequency of occurrence of a combined type of pitch accent and information status category
   - c – the relative frequency of occurrence of pitch accent+word pairs
   - d – the relative frequency of occurrence of words in their context

2. Pitch accent range is influenced by the frequency of occurrence of a combined type of pitch accent and word

3. Tonal context variability is influenced by the relative frequency of occurrence of words in their context

Exemplar Theory was shown to be able to account for those frequency-related changes. If fundamental frequency is assumed to be part of the exemplar realisation, existing exemplar models should be easy to augment to incorporate intonation. This is not the case for traditional generative models of intonation, such as autosegmental-metrical models, where counters or word-specific rules would have to be introduced.

How do the word and tonal level interact in exemplar-theoretic selection? The relevance of pitch accent labels in exemplar-theoretic selection was examined by investigating if competitors to a word-accent combination can interfere with a new production. The results indicated that

1. Tonal properties can be relevant to the selection process

2. Word identity is relevant for selecting exemplars

3. Pitch accent *type* is not considered in selection

The findings were shown to be in keeping with functional approaches to model intonation. Supporters of these intonation models argue that an intermediate level, which describes intonation with the help of categorical labels, is unnecessary, and that a direct mapping of linguistic functions to acoustic or articulatory features should be determined. That is, these models provide a way of capturing intonation without assuming an additional layer of phonological categories that have a specific, invariable surface realisation. Generally, these approaches to intonation are highly compatible with Exemplar Theory. The results obtained in this thesis suggest that linguistic functions that have an influence on tonal parameters contribute to the formation of exemplar clouds. These clouds are assumed to specify the set of exemplars which is selected when a new instance is to be produced. Thus, linguistic functions are considered in the selection process. Hence, functional intonation models might be the preferable way of analysing intonation – especially when intonation is to be integrated into exemplar models of speech production and perception.

What are the essential features of an exemplar model accounting for intonation? The results of the experiments were interpreted from an exemplar theoretic viewpoint. Starting out from a simple exemplar-theoretic model, the following adaptations were suggested in order to account for the effects found in the corpus studies.

1. Storage of tonal contours with the exemplar
2. Storage of contextual information
3. Flexible unit size
4. Selection of production targets according to tonal parameters
5. Ability to model production biases for tonal parameters
6. Ability to model entrenchment of tonal parameters

These adaptations led to a novel exemplar model, which incorporates intonation. The proposed mechanisms in the model are based on existing exemplar-theoretic implementations which should make a computational simulation of the observed effects straightforward.

Summary The research presented here constitutes a step towards an integrated exemplar-based model of speech production in which intonation is accounted for. Although open questions concerning the detailed implementation of such a model remain, the studies suggest that Exemplar Theory is highly suited to model intonation production. The carefully crafted corpus experiments presented here yielded results which highlight in considerable detail the impact that frequency of occurrence has on supra-segmental parameters just as it has been shown to be influential on the segmental level. Therefore, if intonation is to be modelled in other frameworks, frequency effects should be taken into account in order to attain a comprehensive picture of the production of intonation.

Bibliography

Abebe, A., Daniels, J., McKean, J. W., Kapenga, J. A., 2001. Statistics and data analysis, textbook for a course offered by the Department of Statistics at Western Michigan University.
URL `http://www.stat.wmich.edu/s160/book/`

Augurzky, P., 2008. Prosodic balance constrains argument structure interpretation in German. Poster presented at AMLaP 2008, Cambridge.

Baayen, H., 1992. Quantitative aspects of morphological productivity. Yearbook of Morphology 1991, 109–149.

Baayen, H., 1994. Productivity in language production. Language and Cognitive Processes 9 (3), 447–469.

Baker, R. E., Bradlow, A. R., 2009. Variability in Word Duration as a Function of Probability, Speech Style and Prosody. Language and Speech 52 (4), 391–413.

Bard, E. G., Aylett, M. P., 1999. The dissociation of deaccenting, givenness, and syntactic role in spontaneous speech. In: Proceedings of the 14th International Congress of Phonetic Sciences (San Francisco). Vol. 3. pp. 1753–1756.

Batliner, A., Möbius, B., 2005. Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground? In: Barry, W. J., van Dommelen, W. A. (Eds.), The Integration of Phonetic Knowledge in Speech Technology. Springer, Dordrecht, pp. 21–44.

Batliner, A., Möbius, B., Möhler, G., Schweitzer, A., Nöth, E., 2001. Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground. In: Proceedings of the European Conference on Speech Communication and Technology. Vol. 4. Aalborg, Denmark, pp. 2285–2288.

Baumann, S., 2006. The Intonation of Givenness – Evidence from German. Vol. 508 of Linguistische Arbeiten. Niemeyer, Tübingen, Ph.D. thesis, Saarland University.

Baumann, S., Grice, M., 2006. The intonation of accessibility. Journal of Pragmatics 38 (10), 1636 – 1657, prosody and Pragmatics.
URL http://www.sciencedirect.com/science/article/B6VCW-4JMVHR2-3/2/4c31f695b09fb445e16f753bf47384e8

Baumann, S., Grice, M., Benzmüller, R., 2001. GToBI - a phonological system for the transcription of German intonation. In: Puppel, S., Demenko, G. (Eds.), Prosody 2000. Speech Recognition and Synthesis. Poznan: Adam Mickiewicz University, Faculty of Modern Languages and Literature, pp. 21–28.

Baumann, S., Grice, M., Steindamm, S., 2006. Prosodic marking of focus domains - categorical or gradient? In: Proceedings of the third International Confrerence on Speech Prosody. Dresden, Germany, pp. 301–304.

Beckman, M., Hirschberg, J., 1999. The ToBI annotation conventions. http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html.

Beckman, M., Pierrehumbert, J., 1986. Intonational structure in English and Japanese. Phonology Yearbook 3, 255–310.

Bien, H., Levelt, W. J., Baayen, R. H., 2005. Frequency effects in compound production. Proceedings of the National Acadademy of Sciences of the United States of America 102 (49), 17876–17881.

Bod, R., 1998. Beyond Grammar: An Experience-Based Theory of Language. CSLI Publications / Cambridge University Press, Cambridge, UK.

Bod, R., 2009. From exemplar to grammar: A probabilistic analogy-based model of language learning. Cognitive Science 33 (4).

Bolinger, D. L., 1985. Intonation and its parts: Melody in spoken English. Arnold, London.

Braun, B., Dainora, A., Ernestus, M., 2011. An unfamiliar intonation contour slows down online speech comprehension. Language and Cognitive Processes 26 (3), 350–375.
URL http://www.ingentaconnect.com/content/psych/plcp/2011/00000026/00000003/art00002

Braun, B., Johnson, E. K., 2011. Question or tone 2? How language experience and linguistic function guide pitch processing. Journal of Phonetics, 585–594.
URL http://dx.doi.org/10.1016/j.wocn.2011.06.002

Braun, B., Kochanski, G., Grabe, E., Rosner, B. S., 2006. Evidence for attractors in English intonation. The Journal of the Acoustical Society of America 119 (6), 4006–4015.
URL http://link.aip.org/link/?JAS/119/4006/1

Brenier, J. M., Nenkova, A., Kothari, A., Whitton, L., Beaver, D., Jurafsky, D., 2006. The (non)utility of linguistic features for predicting prominence in spontaneous speech. In: in IEEE/ACL 2006 Workshop on Spoken Language Technology. pp. 54–57.

Browman, C. P., Goldstein, L., 1990. Tiers in articulatory phonology, with some implications for casual speech. In: Kingston, T., Beckman, M. E. (Eds.), Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech. Cambridge University Press, pp. 341–376.

Brown, G., 1983. Prosodic structure and the given/new distinction. In: Cutler, A., Ladd, D. R. (Eds.), Prosody: Models and Measurements. Springer, New York, pp. 67–77.

Bybee, J., 2000. The phonology of the lexicon: Evidence from lexical diffusion. In: Barlow, M., Kemmer, S. (Eds.), Usage-based models of language. Stanford:CSLI, pp. 65–85.

Bybee, J., 2006. From usage to grammar: The mind's response to repetition. Language 84, 529–551.

Bybee, J., Scheibman, J., 1999. The effect of usage on degrees of constituency: The reduction of don't in English. Linguistics 37 (4), 575–596.

Bybee, J. L., 1995. Diachronic and typological properties of morphology and their implications for representation. In: Feldman, L. (Ed.), Morphological aspects of language processing. Lawrence Erlbaum Associates, Hillsdale, N.J., pp. 225–246.

Bybee, J. L., Eddington, D., 2006. A Usage-based Approach to Spanish Verbs of 'Becoming'. Language 82 (2), 323–355.

Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., Beaver, D., 2010. The NXT-format Switchboard Corpus: A rich resource for investigating

the syntax, semantics, pragmatics and prosody of dialogue. Language Resources and Evaluation 44 (4), 387–419.

Calhoun, S., Nissim, M., Steedman, M., Brenier, J., 2005. A framework for annotating information structure in discourse. In: Pie in the Sky: Proceedings of the workshop, ACL. pp. 45–52.

Calhoun, S., Schweitzer, A., forthcoming. Can intonation contours be lexicalised? Implications for discourse meanings. In: Elordieta Alcibar, G., Prieto, P. (Eds.), Prosody and Meaning (Trends in Linguistics). Mouton DeGruyter.

Carreiras, M. A., Perea, M. B., 2004. Naming pseudowords in Spanish: Effects of syllable frequency. Brain and Language 90, 393–400.

Celex, 1993. The CELEX lexical database - Dutch, English, German. CD-ROM. Centre for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen.

Chafe, W. L., 1994. Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing. University Of Chicago Press.

Cholin, J., Levelt, W. J. M., Schiller, N. O., March 2006. Effects of syllable frequency in speech production. Cognition 99 (2), 205–235.
URL http://dx.doi.org/10.1016/j.cognition.2005.01.009

Chomsky, N., 1957. Syntactic Structures. Mouton.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper & Row, New York.

Cohen, W. W., 1995. Fast effective rule induction. In: In Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann, pp. 115–123.

Cosi, P., Avesani, C., Tesser, F., Gretter, R., Pianesi, F., sept. 2002. A modified "PaIntE" model for Italian TTS. In: Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on. pp. 131 – 134.

Croot, K., Rastle, K., 2004. Is there a syllabary containing stored articulatory plans for speech production in English? In: Proceedings of the 10th Australian International Conference on Speech Science and Technology (Sydney). pp. 376–381.

Crystal, D., 1969. Prosodic Systems and Intonation in English, 1st Edition. Cambridge University Press, 32 East 57th Street, New York, N.Y. 10022.

Cutler, A., Donselaar, W. V., June 2001. Voornaam is not (really) a homophone: Lexical prosody and lexical access in Dutch. Language and Speech 44, 171–195.

Dogil, G., Möbius, B., 2001. Towards a model of target oriented production of prosody. In: Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark). Vol. 1. pp. 665–668.

Erman, B., Warren, B., 2000. The idiom principle and the open choice principle. Text Interdisciplinary Journal for the Study of Discourse 20 (1), 29–62.
URL http://www.reference-global.com/doi/abs/10.1515/text.1.2000.20.1.29

Fernandez, R., Ramabhadran, B., 2010. Discriminative training and unsupervised adaptation for labeling prosodic events with limited training data. In: Proceedings of Interspeech. pp. 1429–1432.

Féry, C., 1993. German Intonational Patterns (Linguistische Arbeiten). Max Niemeyer Verlag.

Festival, I., 2010. IMS German Festival home page. Insitut für Maschinelle Sprachverarbeitung, Universität Stuttgart.
URL www.ims.uni-stuttgart.de/phonetik/synthesis

Francis, N. W., Kučera, H., 1982. Frequency Analysis of English Usage: Lexicon and Grammar. Houghton Mifflin, Boston, MA.

Fujisaki, H., 1988. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In: Fujimura, O. (Ed.), Vocal Physiology: Voice Production, Mechanisms and Functions. Raven, New York, pp. 347–355.

Fujisaki, H., Ohno, S., 1995. Analysis and modeling of fundamental frequency contours of english utterances. In: Proceedings of EUROSPEECH. ESCA, Madrid, Spain, pp. 985–988.

Fujisaki, H., Ohno, S., Wang, C., 1998. A command-response model for f0 contour generation in multilingual speech synthesis. In: Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis (SSW3). Blue Mountains (Australia), pp. 299–304.

Godfrey, J., Holliman, E., McDaniel, J., 1992. Switchboard: Telephone speech corpus for research and development. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-92. Vol. 1. pp. 517–520.

Goldinger, S. D., 1997. Words and voices—perception and production in an episodic lexicon. In: Johnson, K., Mullennix, J. W. (Eds.), Talker variability in speech processing. Academic Press, San Diego, pp. 33–66.

Goldinger, S. D., 1998. Echoes of echoes? A episodic theory of lexical access. Psychological Review 105 (2), 251–279.

Green, K., Kuhl, P., Meltzoff, A., Stevens, E., 1991. Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the mcgurk effect. Attention, Perception, & Psychophysics 50, 524–536.
URL http://dx.doi.org/10.3758/BF03207536

Greenberg, S., Ellis, D., Hollenback, J., 1996. Insights into spoken language gleaned from phonetic transcription of the switchboard corpus. In: ICSLP-96. Philadelphia, PA, pp. 24–27.

Grice, M., Benzmüller, R., 1995. Transcription of German Intonation using ToBI tones; The Saarbrücken System. Phonus 1, 33–51.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J., Batliner, A., 1996. Consistency in transcription and labelling of German intonation with GToBI. In: Proceedings of ICSLP. pp. 1716–1719.

Halliday, M. A. K., 1967. Intonation and Grammar in British English. Mouton, The Hague.

Harrington, J., 2006. An acoustic analysis of 'happy-tensing' in the queen's christmas broadcasts. Journal of Phonetics 34, 439–457.

Harrington, J., Palethorpe, S., Watson, C., 2000a. Does the queen speak the queen's english? Nature, 927–928.

Harrington, J., Palethorpe, S., Watson, C., 2000b. Monophthongal vowel changes in received pronunciation: an acoustic analysis of the queen's christmas broadcasts. Journal of the International Phonetic Association, 63–78.

Hawkins, S., Smith, R., 2001. Polysp: a polysystemic, phonetically-rich approach to speech understanding. Italian Journal of Linguistics - Rivista di Linguistica 13 (1), 99–188.

Hay, J., 2001. Lexical frequency in morphology: is everything relative? Linguistics 39 (6), 1041.

Hay, J., Bresnan, J., 2006. Spoken syntax: The phonetics of *giving a hand* in New Zealand English. The Linguistic Review 23, 321–349.

Hay, J., Jannedy, S., Mendoza-Denton, N., 1999. Oprah and /ay/: lexical frequency, referee design and style. In: Proceedings of the 14th International Congress of Phonetic Sciences. San Francisco, pp. 1389–1392.

Hendriks, H., McQueen, J. (Eds.), 1996. Annual Report 1995. Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

Hintzman, D. L., 1986. 'schema abstraction' in a multiple-trace memory model. Psychological Review 93, 328–338.

Jilka, M., Möbius, B., 2007. The influence of vowel quality features on peak alignment. In: Proceedings of Interspeech 2007 (Antwerpen). pp. 2621–2624.

Johnson, K., 1997. Speech perception without speaker normalization: An exemplar model. In: Johnson, K., Mullennix, J. W. (Eds.), Talker Variability in Speech Processing. Academic Press, San Diego, pp. 145–165.

Johnson, K., 2005. Speaker normalization in speech perception. Blackwell Publishing, Ch. 15, pp. 363–389.

Johnson, K., 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. Journal of Phonetics 34 (4), 485–499.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W. D., 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In: Bybee, J., Hopper, P. (Eds.), Frequency and the Emergence of Linguistic Structure. Benjamins, Amsterdam, pp. 229–254.

Jusczyk, P., Hohne, E., Jusczyk, A., Redanz, N., 1993. Do infants remember voices? JASA, 2373.

Kingsbury, P., Strassel, S., McLemore, C., McIntyre, R., 1997. Callhome American English transcripts. Linguistics Data Consortium, Philadelphia, No. LDC97T14.

Kohler, K. J., 1990. Macro and micro F0 in the synthesis of intonation. In: Kingston, J., Beckman, M. E. (Eds.), Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech. Cambridge University Press, Cambridge, UK, pp. 115–138.

Kruschke, J., 2011. Models of Attentional Learning. In: Pothos, E., Wills, A. (Eds.), Formal Approaches in Categorization. Cambridge University Press, pp. 120–152.

Kruschke, J. K., 1992. ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review 99 (1), 22–44.

Kruschke, J. K., Johansen, M. K., 1999. A model of probabilistic category learning. Journal of Experimental Psychology: Learning, Memory and Cognition 25, 1083–1119.

Kučera, H., Francis, W. N., 1967. Computational Analysis of Present-Day American English. Brown University Press, Providence, RI.

Ladd, D., 2008. Intonational Phonology, 2nd Edition. Cambridge University Press, Cambridge, UK.

Ladd, D. R., 1980. The Structure of Intonational Meaning—Evidence from English. Indiana University Press, Bloomington, IN.

Ladd, D. R., Mennen, I., Schepman, A., 2000. Phonological conditioning of peak alignment in rising pitch accents in Dutch. The Journal of the Acoustical Society of America 107 (5), 2685–2696.
URL `http://link.aip.org/link/?JAS/107/2685/1`

Ladd, R., 1996. Intonational phonology. Cambridge University Press.

Lambrecht, K., 1994. Information Structure and Sentence Form. Cambridge University Press, Cambridge.

Large, N. R., Pisoni, D. B., 1998. Subjective familiarity of words: Analysis of the hoosier mental lexcion. In: Research on spoken language processing progress report no. 22. Indiana University, Bloomington, Indiana, pp. 215–231.

Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. Journal of the Acoustical Society of America 105 (3), 1455–1468.

Levelt, W. J. M., 1989. Speaking: from intention to articulation. MIT press, Cambridge.

Levelt, W. J. M., Roelofs, A., Meyer, A. S., 1999. A theory of lexical access in speech production. Behavioral and Brain Sciences 22, 1–75.

Levelt, W. J. M., Wheeldon, L., 1994. Do speakers have access to a mental syllabary? Cognition 50, 239–269.

Liberman, A. M., Mattingly, I. G., 1985. The motor theory of speech perception revised. Cognition 21, 1–36.

Liberman, M., 1975. The Intonation System of English. Ph.D. thesis, Massachusetts Institute of Technology, reproduced by the University of Indiana Linguistics Club, Bloomington.

Liberman, M. Y., Pierrehumbert, J., 1984. Intonational invariants under changes in pitch range and length. In: Aronoff, M., Oehrle, R. T. (Eds.), Language Sound Structure. MIT Press, Cambridge, MA, pp. 157–233.

Liberman, M. Y., Prince, A., 1977. On stress and linguistic rhythm. Linguistic Inquiry 8, 249–336.

Lindblom, B., 1984. Economy of Speech Gestures. In: MacNeilage, P. (Ed.), The Production of Speech. Springer, New York, pp. 217–245.

Lindfield, K. C., Wingfield, A., Goodglass, H., 1999. The role of prosody in the mental lexicon. Brain and Language 68 (1-2), 312 – 317.
URL http://www.sciencedirect.com/science/article/B6WC0-45JK39J-1J/2/f792fec9b5fe5393bd9382bebe27b4bb

Losiewicz, B. L., 1992. The effect of frequency on linguistic morphology. Ph.D. thesis, University of Texas, Austin, TX.

Luce, P. A., 1986. Neighborhoods of words in the mental lexicon. Ph.D. thesis, Indiana University, Bloomington, Dept. of Psychology.

Luce, P. A., Pisoni, D. B., 1998. Recognizing spoken words: The neighborhood activation model. Ear and Hearing 19 (1), 1–36.

Mandel, D. R., Jusczyk, P. W., Nelson, D. G. K., 1994. Does sentential prosody help infants organize and remember speech information? Cognition 53 (2), 155 – 180.
URL http://www.sciencedirect.com/science/article/pii/0010027794900698

Marsi, E., Reynaert, M., van den Bosch, A., Daelemans, W., Hoste, V., 2003. Learning to predict pitch accents and prosodic boundaries in Dutch. In: Proceedings of the ACL-2003 Conference (Sapporo, Japan). pp. 489–496.

Mayer, J., 1995. Transcribing German intonation – the Stuttgart system. Tech. rep., Universität Stuttgart.
URL http://www.ims.uni-stuttgart.de/phonetik/joerg/labman/STGTsystem.html

McQueen, J. M., 2005. The Handbook of Cognition. Sage Publications, London, Ch. Speech perception, pp. 255–275.

Medin, D. L., Schaffer, M. M., 1978. Context theory of classification learning. Psychological Review 85 (3), 207 – 238.
URL http://www.sciencedirect.com/science/article/B6X04-4NN6WB6-5/2/cb2dda0c11adb6a6beef3ce1b7dc2058

Meyer, A. S., Roelofs, A., Levelt, W. J. M., 2003. Word length effects in object naming: The role of a response criterion. Journal of Memory and Language 48 (1), 131 – 147.
URL http://www.sciencedirect.com/science/article/B6WK4-47K2R3P-2/2/98aca502da829f67d89f64df934761ee

Möbius, B., 1995. Components of a quantitative model of German intonation. In: Proceedings of the 13th International Congress of Phonetic Sciences (Stockholm). Vol. 2. pp. 108–115.

Möbius, B., Pätzold, M., Hess, W., 1993. Analysis and synthesis of German F0 contours by means of Fujisaki's model. Speech Communication 13, 53–61.

Möhler, G., 2001. Improvements of the PaIntE model for F_0 parametrization. Tech. rep., Institute of Natural Language Processing, University of Stuttgart, draft version.

Möhler, G., Conkie, A., 1998. Parametric modeling of intonation using vector quantization. In: Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia). pp. 311–316.

Mücke, D., Grice, M., Becker, J., Hermes, A., Baumann, S., 2006. Articulatory and acoustic correlates of prenuclear and nuclear accents. In: Proceedings of Speech Prosody 2006 (Dresden). pp. 297–300.

Müller, K., 2002. Probabilistic syllable modeling using unsupervised and supervised learning methods. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 8 (3). University of Stuttgart.

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. In: in Proceedings of NAACL-HLT. pp. 9–16.

Nooteboom, S. G., Kruyt, J. G., 1987. Accents, focus distribution, and the perceived distribution of given and new information: an experiment. Journal of the Acoustical Society of America 82, 1512–1524.

Norris, D., McQueen, J. M., Apr. 2008. Shortlist B: A Bayesian model of continuous speech recognition. Psychological Review 115 (2), 357–395.
URL http://dx.doi.org/10.1037/0033-295X.115.2.357

Nosofsky, R. M., 1986. Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General 115 (1), 39–57.

Nosofsky, R. M., Kruschke, J. K., McKinley, S. C., 1992. Combining exemplar-based category representations and connectionist learning rules. Journal of Experimental Psychology: Learning Memory and Cognition 18 (2), 211–233.

Nusbaum, H., Pisoni, D., Davis, C., 1984. Sizing up the hoosier mental lexcion: Measuring the familiarity of 20,000 words. In: Research on Speech Perception Progress Report No. 10. Indiana University, Bloomington, Indiana, pp. 257–276.

Nygaard, L. C., Pisoni, D. B., 1998. Talker-specific learning in speech perception. Perception and Psychophysics 60, 355–376.

Nygaard, L. C., Sommers, M. S., Pisoni, D. B., January 1994. Speech perception as a talker-contingent process. Psychological Science 5 (1), 42–46.

Ostendorf, M., Price, P. J., Shattuck-Hufnagel, S., 1995. The Boston University Radio News Corpus. Tech. Rep. ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA.

Palmeri, T. J., Goldinger, S. D., Pisoni, D. B., 1993. Episodic encoding of voice attributes and recognition memory for spoken words. Journal of Experimental Psychology: Learning, Memory, and Cognition 19 (2), 309–328.

Pan, S., Hirschberg, J., 2000. Modeling local context for pitch accent prediction. In: ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, USA, pp. 233–240.

Pan, S., McKeown, K. R., 1999. Word informativeness and automatic pitch accent modeling. In: In Proceedings of EMNLP/VLC 99. pp. 148–157.

Pernegger, T. V., 1998. What's wrong with Bonferroni adjustment. British Medical Journal 316, 1236–1238.

Pierrehumbert, J., 1999. What people know about sounds of language. Studies in the Linguistic Sciences 29 (2), 111–120.

Pierrehumbert, J., 2000. Tonal elements and their alignment. In: Horne, M. (Ed.), Prosody: Theory and Experiment—Studies Presented to Gösta Bruce. Kluwer, Dordrecht, pp. 11–36.

Pierrehumbert, J., 2001. Exemplar dynamics: Word frequency, lenition and contrast. In: Bybee, J., Hopper, P. (Eds.), Frequency and the Emergence of Linguistic Structure. Benjamins, Amsterdam, pp. 137–157.

Pierrehumbert, J., 2002. Word-specific phonetics. In: Gussenhoven, C., Warner, N. (Eds.), Laboratory Phonology 7. Mouton de Gruyter, Berlin, pp. 101–140.

Pierrehumbert, J., Beckman, M., 1988. Japanese Tone Structure. MIT Press, Cambridge, MA.

Pierrehumbert, J., Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P. R., Morgan, J., Pollack, M. E. (Eds.), Intentions in Communication. MIT Press, Cambridge, MA, pp. 271–311.

Pierrehumbert, J. B., September 1980. The phonology and phonetics of English intonation. Ph.D. thesis, Massachusetts Institute of Technology.

Prince, A., Smolensky, P., 1993. Optimality theory: Constraint interaction in generative grammar. Tech. rep., Rutgers University.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rabiner, L. R., Juang, B.-H., January 1986. An introduction to hidden markov models. IEEE ASSP magazine.

Rapp, S., 1998. Automatisierte Erstellung von Korpora für die Prosodieforschung. Ph.D. thesis, IMS, Universität Stuttgart, aIMS 4 (1).

Riester, A., 2008a. A Semantic Explication of Information Status and the Underspecification of the Recipients' Knowledge. In: Grønn, A. (Ed.), Proceedings of Sinn und Bedeutung 12. Oslo, pp. 508–522.

Riester, A., 2008b. The components of focus and their use in annotating information structure. Ph.D. thesis, University of Stuttgart.

Rosenberg, A., 2010. AutoBI - a tool for automatic tobi annotation. In: Proceedings of Interspeech 2010. pp. 146–149.

Schacter, D. L., Church, B. A., 1992. Auditory Priming: Implicit and Explicit Memory for Words and Voices. Journal of Experimental Psychology: Learning, Memory, and Cognition 18 (5), 915–930.
URL `http://www.sciencedirect.com/science/article/B6X09-46V0823-2/2/b316fee3b6de98db80499f8ae8d41681`

Schweitzer, A., Möbius, B., 2004. Exemplar-based production of prosody: Evidence from segment and syllable durations. In: Speech Prosody 2004 (Nara, Japan). pp. 459–462.

Schweitzer, A., 2010. Production and perception of prosodic events – evidence from corpus-based experiments. Doctoral dissertation, Universität Stuttgart.

Schweitzer, K., Calhoun, S., Schütze, H., Schweitzer, A., Walsh, M., 2010a. Relative frequency affects pitch accent realisation: Evidence for exemplar storage of prosody. In: Proceedings of the Thirteenth Australasian International Conference on Speech Science and Technology (SST) 2010. Melbourne, Australia, pp. 62–65.

Schweitzer, K., Riester, A., Walsh, M., Dogil, G., 2009a. Pitch Accents and Information Status in a German Radio News Corpus. In: Proceedings of Interspeech 2009. Brighton, UK, pp. 877–880.

Schweitzer, K., Walsh, M., Calhoun, S., Schütze, H., 2011. Prosodic variability in lexical sequences: Intonation entrenches too. In: Proceedings of ICPhS 2011. Hong Kong, pp. 1778–1781.

Schweitzer, K., Walsh, M., Möbius, B., Riester, A., Schweitzer, A., Schütze, H., March 2009b. Frequency matters: Pitch accents and information status. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). Association for Computational Linguistics, Athens, Greece, pp. 728–736.
URL http://www.aclweb.org/anthology/E09-1083

Schweitzer, K., Walsh, M., Möbius, B., Schütze, H., 2010b. Frequency of occurrence effects on pitch accent realisation. In: Proceedings of Interspeech 2010. Makuhari, Japan, pp. 138–141.

Sekiguchi, T., 2006. Effects of lexical prosody and word familiarity on lexical access of spoken Japanese words. Journal of Psycholinguistic Research 35, 369–384.

Siepmann, R., 2001. Phonetische Intonationsmodelle und die Parametrisierung von kontrastiven Satzakzenten im Deutschen. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation (München), FIPKM 38, 3–111.

Silverman, K., Backman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. ToBI: A standard for Labeling English Prosody. In: Proceedings of the 1992 International Conference on Spoken Language Processing. Vol. 2. Banff, Canada, pp. 867–870.

Silverman, K., Pierrehumbert, J., 1990. The timing of prenuclear high accents in English. In: Kingston, J., Beckman, M. E. (Eds.), Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech. Cambridge University Press, Cambridge, UK, pp. 71–106.

Soto-Faraco, S., Sebastián-Gallés, N., Cutler, A., 2001. Segmental and suprasegmental mismatch in lexical access. Journal of Memory and Language 45 (3), 412 – 432.
URL http://www.sciencedirect.com/science/article/B6WK4-457VF2D-M/2/3f503c96a09433aa1f0e8ff15b29e944

Sprenger, S., Levelt, W. J. M., Kempen, G., 2006. Lexical access during the production of idiomatic phrases. Journal of Memory and Language 54 (2), 161–184.
URL http://pubman.mpdl.mpg.de/pubman/item/escidoc%3A58711%3A2/component/escidoc%3A58712/Sprenger_2006_lexical%20access.pdf

Sridhar, V. K. R., Nenkova, A., Narayanan, S., Jurafsky, D., 2008. Detecting Prominence in Conversational Speech: Pitch Accent, Givenness and Focus . In: Proceedings of Speech Prosody (SP-2008). Campinas, Brazil, pp. 453–456.

StatSoft, 2011. Electronic statistics textbook. http://www.statsoft.com/textbook/.

Strand, E. A., 2000. Gender Stereotype Effects in Speech Processing. Ph.D. thesis, Ohio State University.

Syrdal, A., Möhler, G., Dusterhoff, K., Conkie, A., Black, A., 1998. Three methods of intonation modeling. In: Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia). pp. 305–310.

Taylor, P., Black, A. W., Caley, R., 1998. The Architecture Of The Festival Speech Synthesis System. In: Proceedings of the third ESCA workshop in speech synthesis. pp. 147–151.

Taylor, P., Caley, R., Black, A. W., King, S., 1999. Edinburgh speech tools library. [http://www.cstr.ed.ac.uk/projects/speech_tools/], system Documentation Edition 1.2, for 1.2.0 1999/06/15.

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.

Van Lancker, D., Canter, G. J., 1981. Idiomatic versus literal interpretations of ditropically ambiguous sentences. J Speech Hear Res 24 (1), 64–69.
URL http://jslhr.asha.org/cgi/content/abstract/24/1/64

Van Lancker, D., Canter, G. J., Terbeek, D., 1981. Disambiguation of ditropic sentences: Acoustic and phonetic cues. Journal of Speech and Hearing Research 24 (3), 330–335.
URL http://jslhr.asha.org/cgi/content/abstract/24/3/330

Van Santen, J. P. H., Möbius, B., 2000. A quantitative model of F0 generation and alignment. In: Botinis, A. (Ed.), Intonation—Analysis, Modelling and Technology. Kluwer, Dordrecht, pp. 269–288.

Vigário, M., Freitas, M. J., Frota, S., 2006. Grammar and frequency effects in the acquisition of prosodic words in european portuguese. Language and Speech 49 (2), 175–203.
URL http://las.sagepub.com/content/49/2/175.abstract

Wade, T., Dogil, G., Schütze, H., Walsh, M., Möbius, B., 2010. Syllable frequency effects in a context-sensitive segment production model. Journal of Phonetics 38 (2), 227–239.

URL http://www.sciencedirect.com/science/article/B6WKT-4YVG80R-2/2/0e6e2c74dc9eb8449cb38ff406f37982

Wade, T., Möbius, B., 2007. Speaking rate effects in a landmark-based phonetic exemplar model. In: Proceedings of Interspeech 2007 (Antwerpen). pp. 402–405.

Walsh, M., Möbius, B., Wade, T., Schütze, H., 2010. Multi-level exemplar theory. Cognitive Science 34, 537–582.

Walsh, M., Schütze, H., Möbius, B., Schweitzer, A., 2007. An exemplar-theoretic account of syllable frequency effects. In: Proceedings of the International Congress of Phonetic Sciences (Saarbrücken). pp. 481–484.

Walsh, M., Schweitzer, K., Möbius, B., Schütze, H., 2008. Examining pitch-accent variability from an exemplar-theoretic perspective. In: Proceedings of Interspeech 2008. Brisbane, Australia, pp. 877–880.

Wedel, A., 2006. Exemplar models, evolution and language change. The Linguistic Review 23, 247–274.

Wells, J., 1997. SAMPA computer readable phonetic alphabet. In: Gibbon, D., Moore, R., Winski, R. (Eds.), Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin and New York, pp. 684–732.

Whiteside, S. P., Varley, R. A., 1998. Dual-route phonetic encoding: Some acoustic evidence. In: Proceedings of the 5th International Conference on Spoken Language Processing (Sydney). Vol. 7. pp. 3155–3158.

Witten, I. H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition. Morgan Kaufmann, San Francisco.

Wright, R., 1997. Lexical competition and reduction in speech: A preliminary report. In: Research on Spoken Language Processing Progress Report No. 21. Indiana University, Bloomington, Indiana, pp. 471–485.

Xu, Y., 2004. Understanding tone from the perspective of production and perception. Language and Linguistics 5, 757–797.

Xu, Y., 2005. Speech melody as articulatorily implemented communicative functions. Speech Communication, 220–251.

Yule, G., 1980. Intonation and Givenness in Spoken Discourse. Studies in Language, 271–286.

Zipf, G. K., 1949. Human Behavior and the Principle of Least Effort—An Introduction to Human Ecology. Hafner, New York.