Catching words in a stream of speach

Coltekin, Cagri

Catching Words in a Stream of Speech:

Computational Simulations
of
Segmenting Transcribed Child-Directed Speech

Çağrı Çöltekin

university of groningen
faculty of arts
CLCG

bcn

The work presented here was carried out under the auspices of the School of Behavioural and Cognitive Neuroscience and the Center for Language and Cognition Groningen of the Faculty of Arts of the University of Groningen.

Groningen Dissertations in Linguistics 97
ISSN 0928-0030

Cover art, titled *auto*, by Franek Timur Çöltekin

Document prepared with LaTeX 2$_\varepsilon$ and typeset by pdfTeX
Printed by Wöhrmann Print Service, Zutphen

RIJKSUNIVERSITEIT GRONINGEN

Catching Words in a Stream of Speech

Computational Simulations of Segmenting Transcribed Child-Directed Speech

Proefschrift

ter verkrijging van het doctoraat in de
Letteren
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
donderdag 8 december 2011
om 14.30 uur

door

Çağrı Çöltekin
geboren op 28 februari 1972
Çıldır, Turkije

Promotor:	Prof. dr. ir. J. Nerbonne
Beoordelingscommissie:	Prof. dr. A. van den Bosch
	Prof. dr. P. Hendriks
	Prof. dr. P. Monaghan

ISBN (electronic version): 978-90-367-5259-6
ISBN (print version): 978-90-367-5232-9

Preface

I started my PhD project with a more ambitious goal than what might have been achieved in this dissertation. I wanted to touch most issues of language acquisition, developing computational models for a wider range of phenomena. In particular, I wanted to focus on models of learning linguistic ‘structure’, as it is typically observed in morphology and syntax. As a result, segmentation was one of the annoying tasks that I could not easily step over because I was also interested in morphology. So, I decided to write a chapter on segmentation. Despite the fact that segmentation is considered relatively easy (in comparison to learning syntax, for example) by many people, and it is studied relatively well, every step I took for modeling this task revealed another interesting problem I could not just gloss over. At the end, the initial ‘chapter’ became the dissertation you have in front of you. I believe I have a far better understanding of the problem now, but I also have many more questions than what I started with.

The structure of the project, and my wanderings in the landscape of language acquisition did not allow me to work with many other people. As a result, this dissertation has been completed in a more independent setting than most other PhD dissertations. Nevertheless, this thesis benefited from my interactions with others. I will try to acknowledge the direct or indirect help I had during this work, but it is likely that I will fail to mention all. I apologize in advance to anyone whom I might have unintentionally left out.

First of all, my sincere thanks goes to my supervisor John Nerbonne. Here, I do not use the word sincere for stylistic reasons. All PhD students acknowledge the supervisor(s), but if you think they all mean it, you probably have not talked to many of them. Besides the valuable comments on the content of my work, I got the attention and encouragement I needed, when I needed it. He patiently read all my early drafts on short notice, even correcting my never-ending English mistakes.

I would also like to thank Antal van den Bosch, Petra Hendriks and Padraic Monaghan for agreeing to read and evaluate my thesis. Their comments and criticisms improved the final draft, and made me look at the issues discussed in the thesis from different perspectives. In earlier stages of my PhD project, I also received valuable comments and criticisms from Kees de Bot and Tamás Biró. Although focus of the project changed substantially, the benefit of their comments remain. Later, regular discussions with fellow PhD students Barbara Plank, Dörte Hessler and Peter Nabende

kept me on track, and I particularly got valuable comments from Barbara on some of the content presented here. Barbara and Dörte also get additional thanks for agreeing to be my paranimphs during the defense.

A substantial part of completing a PhD requires writing. Writing at a reasonable academic level is a difficult task, writing in a foreign language is even more difficult and writing in a language in which you barely understand the basics is almost impossible. First, thanks to Daniël de Kok for helping me with the impossible, and translating the summary to Dutch. Second, I shamelessly used some people close to me for proof reading, on short notice, without much display of appreciation. My many mistakes in earlier drafts of this dissertation were eliminated by the help of Joanna Krawczyk-Çöltekin, Arzu Çöltekin, Barbara Plank and Asena and Giray Devlet. I am grateful for their help, as well as their friendship.

My interest in language and language acquisition that lead to this rather late PhD goes back to my undergraduate years in Middle East Technical University. I am likely to omit many people here because of many years past since. However, the help I got and things that I learned from Cem Bozşahin and Deniz Zeyrek are difficult to forget. I am particularly grateful to Cem Bozşahin for his encouragement and his patience in supervising my many MSc thesis attempts.

This thesis also owes a lot to people that I cannot all name here. I would like to thank to those who share their data, their code, and their wisdom. This thesis would not be possible without many freely available sources of information, tools and data that we take for granted nowadays—just to name a few: GNU utilities, R, LaTeX, CHILDES.

There is more to life than research, and you realize it more if you move to a new city. Many people made the life in Groningen more pleasant during my PhD time here. First, I feel fortunate to be in Alfa-informatica. Besides my colleagues in the department, people of foreign guest club and international service desk of the university made my life more pleasant and less painful. I am reluctant to name people individually because of certainty of omissions. Nevertheless here is an incomplete list of people that I feel lucky to have met during this time, sorted randomly: Ellen and John Nerbonne, Kostadin Cholakov, Laura Fahnenbruck, Dörte Hessler, Ildikó Berzlánovich, Gisi Cannizzaro, Tim Van de Cruys, Daniël de Kok, Martijn Wieling, Jelena Prokić, Aysa Arylova, Barbara Plank, Martin Meraner, Gideon Kotzé, Zhenya Markovskaya, Radek Šimík, Jörg Tiedemann, Tal Caspi, Jelle Wouda.

My parents, Hoşnaz and Selçuk Çöltekin, have always been supportive, but also encouraged my curiosity from the very start. My interest in linguistics likely goes back to a description of Turkish morphology among my father's notes. I still pursue the same fascination I felt when I realized there was a neat explanation to something I knew intuitively. My sister, Arzu, has always been there for me, not only as a supportive family member, but I also benefited a lot from our discussions about my work, sometimes making me feel that she should have been doing what I do. Lastly, many thanks to two people who suffered most from my work on the thesis by being closest to me, Aska and Franek, for their help, support and patience.

Contents

List of Abbreviations

List of Abbreviations

ANN	Artificial Neural Network	17
APS	Argument from poverty of stimulus	35
BF	Boundary F-score	73
BP	Boundary precision	73
BR	Boundary recall	73
BTP	Backwards or reverse TP	99
CDS	Child Directed Speech	61
CHILDES	Child language data exchange system	46
H	Boundary entropy	94
IIL	Identification in the limit also used as ‘Identifiable in the limit’	27
LAD	Language Acquisition Device	6
LF	Lexical (or word-type) f-score	73
LP	Lexical (or word-type) precision	73
LR	Lexical (or word-type) recall	73
MDL	Minimum description length	65
MI	(Pointwise) Mutual Information	90
PAC	Probably approximately correct (learning)	29
PM	Predictability-based Segmentation model	113
PUM	Segmentation model that uses predictability and utterance boundaries	123
PUWM	Segmentation model that uses predictabilty, utterance boundaries and know words	135

PUWSM	Segmentation model that uses all cues: predictabilty, utterance boundaries, known words and stress	143
PV	Predecessor Value	99
RM	Random Model	76
SM	Stress-based Segmentation model	142
SRN	Simple recurrent network	18, 63
SV	Successor Variety	91
TP	Transitional probability	88
UG	Universal Grammar	6
UM	Utterance boundary-based Segmentation model	123
VC dimension	Vapnik-Chervonenkis dimension	30
WF	Word F-score	73
WM	Lexicon-based Segmentation model	132
WP	Word precision	73
WR	Word recall	73

1 Introduction

> A witty saying proves nothing.
>
> Voltaire

The aim of language acquisition research is to understand how children learn languages spoken in their environment. This study contributes to this purpose by investigating one of the first steps of the language acquisition process, the discovery of words in the speech stream directed to children, by means of computational simulations.

We take words for granted, we identify them effortlessly when listening to others speaking a language we understand, and we use them to construct utterances possibly never uttered before. We learn the sound forms of the words, associate them with meanings, discover how to use them appropriately in company of other words, and in presence of different people. Despite apparent ease with which we process and learn words, learning a proper set of words, a lexicon, to effectively communicate with our environment is a challenging task. The challenge starts with identifying these words in a continuous speech stream. Unlike written text where we typically put white spaces between the words, the speech signal does not contain analogous reliable markers for word boundaries.

A competent language user is aided, to some extent, by his/her knowledge of words to extract them from a continuous stream: *itisannoyingbutyouprobablycanfigure-outthewordsinthissequence*. However, at the beginning of their journey to becoming competent speakers, children do not know the words in the language they are acquiring. As a result they cannot make use of words. If this is not convincing, try to locate the word boundaries in this sentence: *eğertürkçebilmiyorsanızbudizidekisözcük-leribulmanızçokzor*. This is approximately what happens when you hear an unfamiliar language (in this case, Turkish). Without knowing the words of the input language, discovering words in a continuous speech stream does not seem possible, which leads to a chicken-and-egg problem. In spoken language, we are not as helpless as in the written stream of letters. There are several acoustic cues that indicate word boundaries. However, although these cues correlate with the boundaries, they are known to be insufficient, noisy and sometimes in conflict with each other. Furthermore, the cues are language dependent, that is, one needs to know the boundaries to learn when and how these cues correlate with the boundaries. We are back to the chicken-and-egg problem again.

Fortunately, there are also some simple and general segmentation strategies that seem to work universally for all languages. A byproduct of the fact that the natural speech stream is formed by concatenating words, the flow of basic units (such as syllables or phonemes) in an utterance follows certain statistical regularities. Particularly, the basic units within words predict one another in sequence, while units across boundaries do not. It is even more encouraging that children seem to be sensitive to these statistics at a very young age. Another source of information for word boundaries that does not require knowledge of words in advance comes from utterance boundaries. Utterance boundaries are also word boundaries, and words are formed by certain regularities, for example they share common beginnings and endings. This provides another source for discovering words before knowing them. Once we start discovering words using these general strategies, we can also learn to use the language-specific cues.

This is a good point at which to summarize the problem:

> Given a list of unsegmented utterances formed using an unknown set of words, and a set of incomplete, noisy and sometimes conflicting cues that correlate with the word boundaries in unknown ways, find the word boundaries.

If you have ever taken a programming class, this might look familiar: it looks like a rather tricky programming exercise. And if you have taken a class in machine learning, you may already have some ideas on how to go about solving it. Regardless of whether it is solved by a human brain or a computer, this is a computational (or information processing) problem. This statement is true for many cognitive processes. As a result, a common method of studying cognitive processes, including language acquisition, is to model them formally, and study the model using computational simulations. The methodology in this study follows this general practice. In a nutshell, a computational model helps us understand the natural phenomenon it models by: (1) finding parallels between the natural phenomenon and the computational model, (2) testing hypotheses that are difficult or impossible to test directly, and (3) providing more insight into the problem by describing it in detail.

I will present computational models of segmentation that offer solutions to the segmentation problem guided by the strategies mentioned in the discussion above, namely, predictability statistics and utterance boundaries followed by cues that are available from the words discovered previously. Special attention will be paid to be consistent with what we do know about child language acquisition. The models will be tested using transcriptions of actual child-directed speech. The modeling effort will follow the cues mentioned in the discussion above. I will start presenting models and results of computational simulations with language-neutral methods, or cues, and demonstrate their usefulness in combination with other language-specific cues.

Before presenting the computational models of segmentation outlined above, the next two chapters will discuss some general issues in the field of language acquisition literature and computational modeling of language acquisition processes. The problem of language acquisition in general, and the debates and issues in the broader field will be discussed Chapter 2. After a discussion of a central debate in the field, the nature–nurture debate, the chapter will review a number of general theories about language acquisition and the solutions they offer for the problem.

Chapter 3 will discuss the computational modeling practice in detail, and how this approach can be helpful in answering questions about cognitive phenomena in general, and language acquisition in particular. The chapter will discuss the differences and similarities between two separate but related methods to study computational models, namely, mathematical analysis of the models and computational simulations. I will argue that these two methods are complementary, yet, in some cases computational simulations may avoid the difficulties faced by analytic methods by adopting sometimes loose, and sometimes more accurate formalizations of certain aspects of the problem being modeled. In studying language acquisition processes, the computational simulations are at an advantage modeling the input to the learner. Even though it is difficult to model the utterances a child hears during language acquisition by mathematical formulas, it is relatively easier and more accurate to model them using appropriately large amounts of child-directed speech corpora. There will be some discussion on the nature–nurture debate in this chapter as well, this time focusing on the formal aspects of it.

Chapter 4 will focus on the problem of segmentation. I will demonstrate the problem in detail, review the relevant developmental psycholinguistic literature, and introduce the cues that are known or believed to be used by children in solving this problem.

Chapter 5 discusses the computational problem of segmentation in more detail. Along with descriptions of different ways of modeling the segmentation problem computationally, I will review relevant previous studies. The chapter discusses the general issues with computational models of segmentation such as the questions they answer, how to evaluate their performance, and how to interpret their results. Furthermore, this chapter will present a reference computational model of segmentation that follows a successful strategy different from the strategy advocated in this study. This model will be used as a reference throughout the rest of the thesis for comparison of the performance of the models developed during this study.

Chapter 6 takes the first step towards the intended computational models of segmentation in this work. After a detailed analysis of a number of measures of predictability (or uncertainty), a predictability-based segmentation model will be presented. The model has two main components. First, given a certain predictability measure based on statistical information extracted from previous input utterances, the model uses an unsupervised method for finding word boundaries in the current utterance. Second, the model uses a method to combine the decisions obtained from a set of individual boundary indicators, or measures. These two components will be used in the following two chapters while incorporating additional indications of word boundaries.

Chapter 7 and Chapter 8, extend the model described in Chapter 6 using information from utterance boundaries and already discovered words, respectively. These two chapters demonstrate that information from different sources are useful in combination, and the cues that are language specific may start being useful once we start learning some words in the input language.

Chapter 9 summarizes the segmentation models presented in the preceding three chapters, provides a qualitative analysis, compares the results among these models and the other models presented in the literature, and finally suggests possible extensions in the future work.

Chapter 10 gives a brief general summary and concludes.

2 The Problem of Language Acquisition

> The most essential characteristic of scientific technique is that it proceeds from experiment, not from tradition.
>
> Bertrand Russell

A typical introduction in many books and articles on language acquisition starts with defining language acquisition with expressions like, ‘the greatest intellectual achievement of one’s lifetime’, ‘endlessly fascinating’, ‘a snap’, ‘an astonishing process’, a ‘fascinating feat’, ‘a monumental achievement’, or ‘a great gift’.[1] Clearly, we are impressed with the way children acquire the languages spoken around them. The difficulty of learning languages in general, and the apparent ease with which children acquire them is what lies behind these big words of appreciation. For those of us who have tried learning a second language, it is clear that learning a language is a difficult task. Children, on the other hand, seem to learn languages, even multiple languages, spoken in their environment in an effortless way. They do not rehearse word lists, they do not need aid from teachers, they do not spend time in language labs, they do not do any grammar exercises, nor do they use any other training material that adult second language learners typically use.

The difficulty of learning languages and the impressive performance children show in this task make research on language acquisition an interesting inquiry. Our knowledge about how children achieve this impressive task is limited, and the theories in the field drastically disagree. As well as providing a broader introduction to main issues in language acquisition literature, the aim of this chapter is to clarify the place of the present study in respect to influential theories or viewpoints in the broader field of language acquisition. The next section will have a closer look at the major disagreement about what enables children to learn languages quickly and effortlessly.

[1] The phrases quoted above are only a few of the words of astonishment that reoccur in the literature with slight variation in wording. The first one can be sourced to Bloomfield (1933, p.29). The others, at least, appeared in Pinker (1995, p.175), Crain and Pietroski (2002), Guasti (2002, p.2) and Akmajian et al. (2010, p.481), in order presented above, and Saxton (2010, p.3) successfully fits the last three into a single paragraph.

Section 2.2 will briefly summarize some of the popular theories of language acquisition, and Section 2.3 will conclude after a brief discussion.

2.1 The nature–nurture debate

Stating that 'the aim of language acquisition research is understanding how children acquire languages' may sound like a tautology. However, to a large extent, the research in language acquisition focuses on providing evidence for or against the existence of an innate language capacity. The underlying purpose of this divergence is to support one of the two philosophical viewpoints on human cognition: *nativism* or *empiricism*. These viewpoints have been under constant debate (also known as the *nature–nurture* debate) as far as known human intellectual history extends, and language acquisition research has been the battlefield of this debate for the last 50 years. This thesis takes no side in this debate, and I find it counter-productive to keep the debate at the main focus of the research agenda. Nevertheless, the debate is too central to the field to go unmentioned.[2] This section presents a brief discussion of the nature–nurture debate in the context of language acquisition, and provides arguments against taking an a priori side in it.

Nativism is the view that certain skills, abilities or knowledge are innate, that they are not learned from the environment. The roots of nativism can be traced, at least, back to Plato, and Descartes was probably the most influential thinker for the modern nativist (or rationalist) standpoint. However, modern *linguistic nativism* gained popularity because of Chomsky's ideas on language acquisition (Chomsky, 1959b, 1965). According to linguistic nativism, humans are born with an innate endowment specific to language, commonly referred to as *language faculty*, *language acquisition device* (LAD) or *universal grammar* (UG). The UG enables acquisition of languages, while environmental factors are regarded as making a minor contribution. As Chomsky (1980) puts it,

> [. . .] in certain fundamental respects we do not really learn language; rather, grammar grows in the mind.
>
> When the heart, or the visual system, or other organs of the body develop to their mature form, we speak of growth rather than of learning. [. . .] In both cases, it seems, the final structure attained and its integration into a complex system of organs is largely predetermined by our genetic program, which provides a highly restrictive schematism that is fleshed out and articulated through interaction with the environment (embryonical or postnatal). (Chomsky, 1980, p.134)

[2] The terms 'acquisition', 'learning' and 'development' often indicate the side a researcher has taken in this debate. In parallel with the arguments in this chapter and Chpater 3, this thesis does not make any clear distinctions between these terms.

Empiricism, on the other hand, is the opposing position that individuals are born without any innate knowledge (the term *tabula rasa* or *blank slate* is frequently used to define this state), that the knowledge comes from perception and experience. This view can be traced as far back as Aristotle, and Locke is the most influential philosopher in this camp. However, current empiricist (or non-nativist) theories of language acquisition diverge from the historical empiricism. In the language acquisition literature, connectionist models and theories of language acquisition (see, e.g., Elman et al., 1996) have been the main representatives of this viewpoint. Contemporary non-nativist theories do not exclude all forms of innate capacities. However, the role of environment and domain-general learning mechanisms are regarded as more important from this perspective.

Indisputably, acquiring languages requires some biological mechanism that we are born with: any normally developing child learns the language(s) he/she is exposed to, but the kitten born and raised in the same environment does not. Likewise, languages are learned: children born in different language environments learn different languages. Hence, besides the domain-specificity of the innate capacity, the disagreement is on the degree—rather than existence— of the innate knowledge or mechanisms.

As can be guessed from phrases like '*largely* predetermined' or '*more* important', the distinction is a fuzzy one.[3] In language acquisition, the role of genetic factors or importance of the environment are not easily quantifiable. Even in their qualitative sense, they seem to be moving targets. For example, while earlier proposals by Chomsky (1981) suggested a complex innate linguistic knowledge in the form of *principles and parameters* theory (P&P, see Section 2.2.1), his later view seems to be reduced only to *recursion* (Hauser, Chomsky and Fitch, 2002). Even if we could state how much and what type of innate knowledge proves a certain point, our knowledge of language and how it is acquired is not sufficient to solve the debate: we know very little about the nature of our linguistic knowledge, and how we acquire it. This information should eventually come from neuroscience. However, we are a long way from a full characterization of neurological processes involved in language acquisition and language use.

Providing a detailed account of the debate is beyond the scope of this thesis.[4] Besides presenting a short overview of common arguments of the debate, the main point of this section is to argue that taking the nativist–empiricist debate as the main focus of language acquisition research is often counter-productive.

[3] There are a number of testable arguments as well, such as 'argument from poverty of stimulus' to which we will return to in Chapter 3.

[4] A popular reference on nativist side of the debate is Pinker (1994), and Sampson (1999) gives an accessible empiricist response to linguistic nativism. Most textbooks on language acquisition take clear sides on this debate. For a recent textbook that provides a balanced account of the debate along with the issues in language acquisition see Saxton (2010).

2.1.1 Difficulty of learning languages

The perception that learning a natural language is a difficult task is hardly controversial. However, the difficulty of learning languages as an argument in the nature–nurture debate requires more scrutiny than it typically receives. In this section I will briefly review a few aspects of natural languages that are assumed to be difficult to learn, and relate it to the debate. Some points which will be discussed in detail in later chapters will also briefly mentioned here. The discussion related to input and formal learnability theory will be left for Chapter 3, and the segmentation problem will be discussed in depth in later chapters.

The difficulties of learning a language start with this very first step: segmenting fluent speech into discrete units is a difficult task (see Chapter 4 for a detailed discussion). Despite its difficulties, children take first steps towards the solution of the segmentation problem as early as their first few months of life, and by their first birthdays they get very close to the solution (Jusczyk, 1999). The segmentation problem rarely makes its way into the nature–nurture debate. However, segmentation is a necessary step for identifying linguistic units like phonemes, syllables or words from a continuous stream of acoustic input. Learning to identify words, or lexical units, is the main focus of this thesis, and the problem will be discussed in detail in Chapter 4 and the chapters that follow.

Even if the segmentation problem is solved, learning words of a language alone is a challenging task. Words are arbitrary and ambiguous sound units. Nevertheless, children around 6 months of age start recognizing the words they hear frequently, such as their names (Bortfeld et al., 2005). With a large individual variation, children start producing their first words around their first birthday, but it is estimated that they understand much more (about 80 words, Fenson et al., 1994). It is a common assumption that sometime between ages 1;6[5] and 2;6 an explosive growth of lexicon, so called *vocabulary spurt*, starts (Bloom, 1976). Despite empirical evidence against certain forms of vocabulary spurt (Ganger and Brent, 2004), it is clear that children learn new words at an increasingly high rate. An estimate that is frequently cited in the literature for the rate of word learning in preschool children is 10 words per day (based on Carey, 1978). However, caution is needed for interpreting this number. Even if the number may be accurate for the complete process, reporting a single number can be misleading: a two-year-old's word learning speed is nowhere near ten words a day. The estimates in the literature suggest a slow start, about 1.6 words per day in the second year of life. The rate reaches to its peak, 12.1 words per day, between ages eight to 10 (Saxton, 2010, p.146 presents estimates of learning rate between ages one to 17). Estimation of lexicon size at age six varies between 10,000 (Bloom and Markson, 1998) to 14,000 (Clark, 1993; Templin, 1957) words. Estimated number of

[5] The age notation follows the standard age notation in language acquisition literature. The ages of children are indicated using three numbers '*year*;*month*.*day*', separated by semicolon and dot in this order. For example 1;3.10 means one year, three months and ten days.

words in the lexicon of an 18-year-old is around 60,000 (Aitchison, 1994). The figures reported here are based on the average values calculated for children learning English. Precise estimation of the vocabulary size is far from trivial (Miller, 1996, pp.134–137), and acquisition paths of children show a large individual variation. As a result, the estimates in the literature tend to show large variation as well. However, it is clear that lexical acquisition starts before the first year of life, and the rate of words learned increases until school years, with an overall rate of eight to ten words per day.

Even with conservative estimates, the ability to learn words that quickly is indeed impressive. However, since word learning also means learning their meaning and usage, the problem is even more difficult than storing sequences of phonemes. An apparent difficulty related to word learning is *referential uncertainty*. It is claimed that when a learner hears a word (or any other linguistic unit), finding the correct referent of the unit in the real world is intractable. The philosophical discussion of this problem can be traced back to Quine (1960, Chapter 2). Quine discusses a hypothetical problem where a linguist who is trying to learn an indigenous language hears the word ‘gavagai’ referring to a rabbit. The question is ‘how can the linguist conclude that the word means *rabbit*?’ He argues that it may as well mean ‘the tail of the rabbit’, ‘this particular rabbit’, ‘any mammal’, ‘color of the rabbit’, ‘tasty!’, ‘nice day’ or (infinitely) many other possibilities. The problem of associating words with their meanings is a popular subject in the linguistic literature, and it is addressed by a large number of researchers from a broad perspective (e.g., Bloom, 2000; Markman, 1989; Siskind, 1996; Tomasello, 2001; Xu and Tenenbaum, 2007).[6] Quine’s original discussion, and many appearances of the problem in the literature are rather informal. However, it is clear that in many circumstances, possible referents of a novel word are ambiguous. Despite this problem, people seem to learn words quickly. In most cases, only a few, or even a single exposure, are enough for people to learn meanings of newly-heard words.

The difficulty of word learning, particularly the problem of assigning meanings to words from the information available in the environment, is sometimes put forward as an argument for nativism. However, no concrete proposals exist for what sort of innate *linguistic* mechanisms may aid word learning. Words are, after all, arbitrary sound sequences, and specific to the particular language being acquired.

The main arena of the nature–nurture debate is learning syntax. Natural language sentences are not just random collection of words that are stringed together. To use a language properly, one needs to learn how to combine words together to form grammatical utterances. Certain assumptions about the nature of the utterances children hear during the acquisition process, *the input*, and negative learnability results from computational learning theory are frequently put together as an argument for linguistic nativism. The formal results from the computational learning theory and their impli-

[6] As well as the attention it received in the linguistic literature, the appearance of the problem in SpecGram (van der Sandt, 2005) is probably a good indication of the popularity of the problem in the linguistic literature.

cations on difficulty of learning languages are discussed in more detail in Chapter 3. For now, it suffices to note that these arguments are misguided because, (1) the input is more structured and richer than portrayed by these arguments; (2) the results from computational learning theory are based on restricted learning settings, such as a concept of learning requiring an ability to distinguish the language learned perfectly, and *formal* languages, neither of which is satisfied in the case of child language acquisition.

2.1.2 How quick is quick enough?

It is a common assumption that children learn languages very quickly. Most of the language acquisition studies cover only the first two or three years of life. The reason behind this is partially the fact that this period covers most of the interesting language acquisition phenomena. However, it is also widely assumed that by the age of three or four, children acquire most of the language. The claims go even further to assume that they show 'adult competence' by three or four (McGilvray, 2006). Even though the observation that a four-year-old child uses language effectively is hardly controversial, the stronger version of the claim that they show adult competence does not seem to hold. The facts about vocabulary learning presented in the previous section already point that most active period for acquiring new lexical items is a lot later than this period (between the ages eight to 10). The vocabulary learning aside, the language acquisition literature is full of examples of late-acquired linguistic phenomena. For example, it is well known that children acquiring Dutch show difficulties with interpretations of pronouns until age six (see, for example, Hendriks et al., 2007; van Rij et al., 2010). Similarly, children acquiring German do not seem to show adult competence in interpretation of case marking until age of seven (Dittmar et al., 2008). Even more dramatically, Omar (1973) reports that children acquiring Egyptian Arabic had difficulties with acquisition of noun plural at the age 15. In the light of this evidence, it is difficult to dismiss the importance of the later acquisition process, and (for example, as Crain and Pietroski, 2002, do) claim that the language acquisition process is 'a snap'.

Language acquisition seems to span a larger time period than most researchers commonly assume. However, the general agreement is that by the age of three or four, children's use of the languages resembles adult competence to a large extent. Now, we can return to the question of whether children learn languages in a short time or not. Claiming that a process takes a short time requires some reference amount of time it should take. In other words, how quick is enough for announcing victory for nativism, or how slow is enough for announcing victory for empiricism?

A possible path to take is to compare children with adults. However, this comparison is problematic for at least two reasons. First, children have rather limited cognitive capabilities compared to adults. Second, the learning settings are very different. Children acquire languages while communicating with adults and other children in their environment without explicit training. However, typically, adults learn languages using various training aids and with hard work. If we attempt to compare children and adults despite these differences, it is doubtful that we would find it impressive that an adult gains a four-year-old competence in a non-native language in four years. For example, in some universities, students who do not speak the language of education are required to reach an academic-level competence in a year or even a shorter time.[7]

It should be noted that neither argument puts adults in a more privileged position than children. There is no evidence that the methods used in second language acquisition are more effective than the child language acquisition setting. Possible interferences from the adult's first language aside, a second language learner rarely has the time and the motivation of a child trying to communicate with his/her environment. Furthermore, the limited capabilities of children may provide some 'maturational constraints' which in turn may help acquisition of languages (Krueger and Dayan, 2009; Newport, 1988, 1990, 1993).

Again, even if we reliably establish that children are faster or slower language learners, it does not necessarily answer the question of innateness. Age affects many cognitive tasks in complicated ways, and even if we can isolate effects of other cognitive functions, we return to a slightly modified version of the same question: How do fast*er* or slow*er* learning rates prove a certain view point?

A possible criticism for the above comparison between adult and child learning is that we are comparing first and second language acquisition processes. Comparing child language acquisition to delayed first language acquisition is, in principle, what we need to do for a fair comparison. And there have been claims based on delayed first language acquisition observed on so-called feral children.[8] However, these cases are far from normal in other aspects of the biological and cognitive development, and (fortunately) these cases are too few to serve for reliable conclusions.

2.1.3 Critical periods

The comparison between adult and child language learning brings us to another popular subject in the debate: *critical periods*. Existence of a critical period for language acquisition has been popularized by Lenneberg (1967). Critical period hypothesis states that successful language acquisition is only possible if the child is exposed to language within an early time window.[9] The critical periods are known to exist in other domains of biological development. A well known example of this kind is filial imprinting. Members of many bird species attach to a moving figure they observe shortly (typically in the first few days) after they were born and follow it as their parent. Similarly, in their Nobel prize winning work, Hubel and Wiesel (1970) found that cats can develop a normal vision only if they were exposed to visual stimuli

[7] It should be noted, however, that adult second language learners normally do not achieve native-like performance in some aspects of the language.

[8] Genie, who was deprived of normal human contact until age 13, is the most popular example documented in the literature (Curtiss, 1977, but see also, Rymer, 1994).

[9] According to Lenneberg (1967), before puberty. But more recent proposals suggest even earlier ages.

in the first 10–12 weeks of life. If a cat is deprived of vision during this period, it becomes permanently blind.

The critical periods are typically used as an argument for nativism: since critical periods are biologically determined, and if human language acquisition is also subject to critical periods, it must be biologically determined.

The problem with this argument is that it is far from being well established that the language acquisition is subject to a critical period. As discussed previously, like many other cognitive functions, language learning ability is linked to age. However, second language acquisition creates an interesting case: people can learn languages even at later ages. People not only are capable of learning languages during adulthood, limited use of the first language may even cause it to deteriorate and even loose its dominance, a process commonly called language attrition (Schmid, 2009). Unlike well-established cases of critical periods, the ability to learn languages seems to show a gradual deterioration, not a complete inability to learn after a certain age. The evidence from delayed learning cases, on the other hand, seems to be indecisive (for a thorough discussion see Saxton, 2010, chapter 3), and the data is interpreted differently depending on the inclinations of the researcher presenting it (Jones, 1995).

2.1.4 Summary

The nature–nurture debate is an exciting philosophical debate which has gained a central position in language acquisition literature. However, the empirical evidence put forward in favor of either theory in the language acquisition literature are far from being conclusive. Furthermore, most of these arguments are often not well defined, or very difficult (or sometimes impossible) to test concretely.

Taking positions based on fuzzy philosophical viewpoints may cause theories to be put forward and data to be interpreted in heavily biased ways. For example, the main motivation of the popular P&P theory of language acquisition is largely based on accepting the nativist viewpoint from the beginning, rather than on available data. Decades of research tried hard to support the theory, yet it has largely been abandoned by its inventors and most of its supporters (see Lappin and Shieber, 2007, for a discussion). Meanwhile, statistical approaches which were popular among structural linguists of the 1950's (e.g., Harris, 1955) had been neglected due to the dominant position of the nativist viewpoint until the 1990s. Saffran, Aslin and Newport (1996a) and subsequent research showed that children are good statistical learners and use statistical learning methods in various tasks in language acquisition, rekindling new interest in these methods.

Arguably, the debate can be fruitful as it may stimulate research and result in an active field. However, it also polarizes the field heavily, causing biased interpretation of scientific findings, making the field more susceptible to problems with the current scientific practice, such as *confirmation bias* (Nickerson, 1997) and *publication bias* (Dickersin, 1990). The confirmation bias is the tendency of people to favor the results that confirms their beliefs. The reflection to this psychological phenomenon in science occurs when scientists resist new discoveries or methods, selectively cite evidence that favors their presumptions. The publication bias is the tendency of publishing research with positive results. The studies that do not support the initial hypothesis tend to be neglected and stay unpublished.

The discussion above already touched on a number of these cases. For example, the case of the feral child known with the name Genie has been interpreted differently depending on who analyzed the case. Genie was kept in a closed space and deprived of normal human contact until she was discovered at the age of 13. A summary of mostly non-linguistic aspects of Genie's case can be found in Rymer (1994), and Curtiss (1977) is the most comprehensive overview of the case from a linguistic perspective. Starting with Curtiss (1977), the linguistic development of Genie has been used as a strong argument in favor of the critical periods in language development. However, Jones (1995) argues, based on the data presented in Curtiss (1977), that Genie's development was not as bad as it was portrayed by research with a nativist inclination. Furthermore, the linguistic data collected and presented in the literature seems too scarce to serve as concluding evidence. And last but not least, as Rymer (1994) clearly demonstrates, Genie's psychological development was far from normal. She not only had a traumatic start to her life, but the trauma continued in the foster homes after she was discovered. These aspects are never mentioned when Genie is brought up as an example that proves the existence of critical periods.

Another, less traumatic, case is the ongoing debate with learning regular and irregular forms, such as the past tense forms of verbs in English. Since the study of Rumelhart and McClelland (1986) there has been a constant debate as to whether the irregular forms are learned using a different mechanism than the regular forms.[10] The problem itself is interesting. However, the motivation in this debate is fueled by the bigger debate of innateness. Although there are no clear reasons for symbolic systems to prove a nativist standpoint (or statistical learning mechanisms to prove an empiricist one), since Chomsky's rejection of statistical methods (Chomsky, 1957), symbolic and rule-based methods of explaining linguistic phenomena have been in favor in nativist linguistic literature. The heated discussion caused this problem to be investigated well. However, the reason behind the missing consensus on how people learn these forms is not the lack of data. A closer look at the research on the subject shows that most researchers start with one of the conclusions and aim at supporting it. Arguably, a more neutral approach would allow us to learn more about the issue. Another unfortunate effect of this debate is due to the fact that it is commonly taken as the problem of learning morphology, causing a large number of interesting aspects of learning morphology to be overlooked behind the dominant interest in this particular problem.

[10] For an analysis of child language data from an alternative approach in this debate, so-called *single-route* vs. *dual-route* learning, see Marcus et al. (1992).

There are numerous other questions in the field that have been affected by the biases of the researchers studying it. Not all of the conclusions are contested by opponents or less biased researchers. However, a number of the cases where disagreement surfaced can be identified by reading Pinker (1994) and Sampson (1999) side by side.

Having said all this, I do not claim that the nativism–empiricism debate is irrelevant to language acquisition research. The long-lasting debate indeed shows that this question about human nature is an intriguing one, but it is also clear that it is far from being resolved yet (see Scholz and Pullum, 2006, for a relevant discussion). Like any other study of human cognition, the language acquisition research may also contribute to this debate. However, it is fruitless to take any a priori sides in this debate, or even a hasty one (as a 'working hypothesis'). Instead of this largely philosophical question, it is more productive to focus on specific questions and theories of the field.

2.2 Theories of language acquisition

Theorizing about nature plays an important role in our scientific inquiry. We typically build formal, testable theories of natural phenomena to explain, understand and predict the phenomena being modeled. Language acquisition is not an exception. Researchers put forward a number of theories with the aim of explaining the child acquisition data, providing more insight into the language acquisition process, and, hopefully, predicting yet unobserved, or unobservable aspects of language acquisition.

In this section I will review three influential theories, or rather frameworks, namely the *principles and parameters theory*, *connectionist networks*, and *usage-based theories* that are used in language acquisition research. A fourth approach, *statistical modeling*, will be discussed next.

Before describing these frameworks, a few notes are in order. First, in principle, the frameworks described here are not necessarily mutually exclusive, it is possible, and not uncommon to see models that cross-cut this classification in some ways. Some examples of these models will be presented in Section 2.2.4. Second, it is well known that all learning systems have to start with certain initial assumptions about the nature of the problem. In other words, there is no 'general-purpose' learning algorithm (see the 'no free lunch' theorem, Wolpert and MacReady, 1997). However, initial assumptions or knowledge do not always entail innate knowledge. For this reason, unless there is a clear theoretical commitment that a certain aspect is innate, borrowing the term from machine learning, I will refer to the initial assumptions of a model as *inductive bias*.

2.2.1 Parametric theories

The nativist conclusion that human languages are not learnable without rich innate linguistic knowledge led researchers to adopt theories that posit an innate UG. In these theories, the UG plays a central role by constraining the learning. The common path taken in these theories is reducing the acquisition process to adjusting a set of parameters. In parametric theories, the learner is assumed to (innately) know what the parameters are. The learning task is setting these parameters to one of the allowed values by observing relevant aspects of the input. Probably the most influential theory of this form is the *principles and parameters theory* (P&P, Chomsky, 1981) that I will describe here. Another well known theory of this form is *optimality theory* (OT, Prince and Smolensky, 1993/2004). P&P and OT differ in their theoretical backgrounds, the linguistic questions they are typically applied to, and the nature of the parameters used.[11] However, from a computational perspective, both theories view learning process as finding values for a number of linguistically motivated parameters.

The solution to the learnability problem proposed by the P&P is based on a set of universal *principles* and *parameters* that all possible languages share. The principles are considered to be shared by all human languages. The parameters are also universal, however, the particular values the parameters take define a particular language.

A frequently cited example of universal principle is that the rules of the grammar have *structure sensitivity*. For example, assuming that English interrogative sentences are formed from their declarative versions, the correct way to turn the declarative sentence *the dog that is in the corner is hungry* to an interrogative question is moving the second *is* to the front. That is, the correct question sentence is *is the dog that is in the corner hungry?* We move the second *is* instead of the first one, because we need to move the auxiliary in the main clause, and this can only be achieved by structure, e.g., *clause*, sensitive rules. Furthermore, it is claimed that this cannot be learned from the linguistic input, and hence, it must be an innate principle (Chomsky, 1965; Crain and Nakayama, 1987). Principles are not learned, and do not vary among different languages.

Like the principles, the specifications of the parameters are also assumed to be innate and universal. However, their values are set during the acquisition process. Parameters are set in one particular way when the child learns one language, and they set another way when another language is learned. The parameters proposed are almost exclusively binary. So, for N parameters it is possible to hypothesize 2^N grammars. A common example of parameters is *null-subject*, or *pro-drop*, parameter. This parameter defines whether it is allowed to skip the pronominal subjects in the language. This is true for languages like Italian, but not for English. So, children in an English speaking environment determine at some point that their language does not allow null-subjects and set the parameter to false, while children learning Italian set the parameter to true. Combination of many such parameters define different possible grammars. The language acquisition proceeds by setting these parameters based on the linguistic input received.

Among attempts to formally define P&P-based language acquisition procedures are Gibson and Wexler (1994) and Fodor (1998). Both procedures are rule based, they make parameter changes based on single examples. This allows the algorithms to

[11] OT stems from an empiricist tradition, and researchers working in OT framework do not typically share the nativist conclusion. The classification here signifies the similarities regarding parametric structure.

generalize quickly using fewer examples. However, they are sensitive to noise and it is not clear if they would recover from an incorrectly learned grammar (see also Niyogi, 2006, for a formal analysis.).

Yang (2002) presents one of the rare P&P-based models that combines statistical techniques with the classical P&P approach. Yang's *variational learner* learns a set of parameters by a statistical system inspired by evolutionary selection. The variational learner alters the weights of all grammars in the space of possible grammars based on input. The simple statistical augmentation compared to previous rule based approaches makes the variational learner more robust against the noise. However, the criticisms listed below are valid for Yang's (2002) variational learner as well.

From the perspective of language acquisition, the main motivation behind P&P approach is that it makes learning easier. One can obtain a large number of possible languages defined by a relatively small set of parameters. Then, the learning task is reduced to setting these small number of parameters. However, if P&P learning procedures are analyzed more carefully, one observes that this certain form of parametrization is neither necessary, nor sufficient for successful learning (see, Clark and Lappin, 2011a; Lappin and Shieber, 2007). However, even if we are convinced that a form of P&P approach makes learning problem easier, there are still a number of issues.

- Despite popularity of the P&P theory in the literature for decades, there is no explicit list of established parameters. Even the highly popular and widely accepted parameters cannot be applied to all languages reliably. For example, another example of a highly, if not most, cited parameter is the *head-direction* parameter that is set as *head-initial* for languages like English, and *head-final* for languages like Japanese. However, the place of Dutch and German is not clear with regard to this parameter. As pointed out by others, for example Newmeyer (2004, 2006) and Trask (2002), the attempts to come up with a list of parameters, e.g. Baker (2001), did not succeed. In Trask's words, '…"all grammars leak", but the parametric approach begins to look uncomfortably like a sieve'.
- Even if a set of parameters is found to explain the differences between languages, it is not clear how these parameters lead to an actual language processing system. For example, there are no explanations for how one can arrive at a generative grammar, e.g., in form of a phrase-structure grammar, from a set of parameter values. As a result, most work on language acquisition with the P&P framework focuses on showing that the learner choose one sort of grammar rather generally, and detailed models following P&P framework do not go beyond 'proof of concept'.
- Most of the P&P-style learning procedures learn from particular aspects of the languages, commonly called *triggers*. The explanation of why learners are sensitive to these particular aspects of the input is left unexplained.
- The parameters commonly listed in the literature generally depend on knowledge

that is likely to be learned. For example the *head-direction* parameter discussed above requires the abstract linguistic concept of *head*, for which even linguists are not in full agreement. Hence, the P&P approaches need to explain how this knowledge is obtained by the learners.

- P&P is typically applied to learning syntax. It is not clear how other aspects of language acquisition, such as word learning or segmentation can be learned by a P&P learner.

These problems have led P&P theory to lose popularity over the last decade. However, it is still rather influential particularly in the theoretical linguistics literature, and supported by many researchers (Boeckx, 2009; Niyogi, 2006; Yang, 2004, exemplifies some recent work on P&P).

2.2.2 Connectionist models

Connectionist systems have been the typical representatives of empiricist models of language acquisition. Unlike P&P-like models of language acquisition, connectionist models do not assume a specialized UG. The learning is achieved through general purpose learning mechanisms that are also useful in other domains of cognitive development.

Connectionist systems, or artificial neural networks (ANNs) are inspired by the functioning of biological neural networks. An individual neuron in an ANN receives a number of activations (numerical inputs). The weighted sum of the inputs is sent to a threshold function, which determines whether the neuron should ‘fire’ or not. The learning in an ANN is achieved by algorithms (such as backpropagation) that adjust the weights of the connections. In a way, one might regard the parameters of the connectionist systems to be the weights. It should be noted that this is a very different concept of parameter than P&P framework.

A typical ANN (backpropagation network) contains 3 layers: The *input layer* consists of neurons that receive the input. The input layer is fully connected to another set of neurons in the *hidden layer*, which in turn are connected to the *output layer*. Each neuron acts according to a simple threshold function depending on the inputs it receives and the weights of the input connections. Learning updates the weights of the connections between input layer and hidden layer, as well as hidden layer and output layer. Multiple hidden layers are possible, but rarely used in practice (however, see the discussion of simple recurrent networks below). The function of the hidden layer or layers can be conceptualized as forming an *internal representation* of the problem.

Probably the most common example of connectionist language acquisition system is English past tense learning system of Rumelhart and McClelland (1986). Rumelhart and McClelland (1986) presented a backpropagation network which learns past tense forms of English verbs. An interesting aspect of the system is that it shows some of the errors children acquiring English make. The model mimics the changes in the rate and the nature of the errors that children make during language acquisition.

One of the weaknesses of standard backpropagation networks is that they cannot generalize beyond simple associations. However, addition of a *recursive layer* allows ANNs to learn relations between inputs over a time period. A particular type of recursive network, called *simple recursive network* (SRN, Elman, 1990), has been popular in modeling various cognitive phenomena. SRNs will further be discussed in the context of modeling segmentation in Section 5.2.1.

Connectionist systems have been instrumental in modeling many cognitive phenomena, and diverse aspects of language acquisition such as, learning phonotactics (e.g., Stoianov and Nerbonne, 2000), speech segmentation (e.g., Christiansen et al., 1998), learning grammatical structure (Elman, 1991). A short but more comprehensive review can be found in Elman (2005).

However there are also a number of criticisms against the use of connectionist systems in modeling cognitive processes. The main points of criticism against connectionist systems as models of language acquisition (or other cognitive functions) are, (1) it is not easy to interpret what the networks learn, and (2) the amount of input needed to train these networks is too large for simulating some aspects of the language acquisition (e.g., word learning).

It should also be noted that contrary to common view, connectionist systems are not free of inductive bias. Most of the inductive bias of a connectionist system is encoded in the network structure and input/output representation. The weight updating systems are also subject to manipulation, and provide additional inductive bias.

2.2.3 Usage based theories

A non-nativist theory of language acquisition that has been gaining popularity for the last decade or so is the *usage-based theory of language acquisition* (Tomasello, 2003, 2009). The emphasis in usage-based theory of language acquisition is on its use in social and communicative context. Unlike nativist theories such as P&P that claim the major part of linguistic knowledge is innately specified, usage-based theory posits that the linguistic structure *emerges* from language use. According to usage-based theory, what children bring to the language acquisition process is their willingness to communicate, particularly, the ability to read intentions of others, and being able to find patterns in the linguistic input.

The usage-based approaches to language acquisition assume that children learn whole utterances as communicative units at the beginning. In time, the pattern-finding mechanism kicks in, and children start learning linguistic constructions which enables them to use language creatively. This aspect of usage-based theory blurs the borders of the lexicon and the grammar, which is another aspect that sets it apart from the nativist theories of language acquisition. Starting from fixed expressions, children develop increasingly general or abstract constructions by observing similar patterns in the input. Typically, it is assumed that children use distributional information and analogy as tools to achieve this, and the role of input frequency is emphasized.

Compared to other frameworks described in this section, the research in usage based theories have a higher tendency to rely on empirical studies. Even though there are several formal flavors of the theory (e.g., Croft, 2001; Goldberg, 2006), and explicit computational models inspired by the theory exist (e.g., Solan et al., 2005), the mainstream usage-based theory is less formally oriented compared to the frameworks previously reviewed. Since the usage-based theory is more empirically oriented, it tends to explain the observed empirical data better. However, the fact that the usage-based theory is not as formally oriented as others, it is difficult to verify its predictions formally, and it is less suitable for computational studies.

2.2.4 Statistical models

Last two decades witnessed a surge of models of language acquisition that do not necessarily follow any of the general theories discussed above. These new models have been influenced by the developments and methods in the fields of computational linguistics and machine learning, typically using statistical learning methods.

Related to the statistical learning methods in modeling language acquisition, there is another debate between symbolic and statistical approaches to learning. The nativist approaches to language acquisition traditionally downplay the role of statistics in language acquisition and use (e.g., Chomsky, 1957, p. 17). Researchers in the empiricist tradition, on the other hand, downplay the role of symbol-like, e.g., linguistic-specific, representations and domain specific learning methods. The connectionist systems summarized in Section 2.2.2 are essentially statistical learners. The representations in connectionist systems are distributed over the weights, and are not easily interpretable. Even though the combination of statistical learning with symbolic-like representations is possible, the polarization in the field led researchers to focus only on one or the other.

Before nativism gained a dominant status in linguistics, in combination with structured, linguistically informed representations, statistics was one of the tools of the structural linguists (e.g., Harris, 1951, 1955). The dominant status of nativist approach in linguistics caused statistical learning methods to be largely ignored until 1990s. After 1990s, many researchers began to investigate the problem of language acquisition using statistical methods (see Abney, 1996, for a thorough discussion of this shift).[12] Furthermore, unlike connectionist systems, these statistical approaches do not reject the use of symbolic/structured or domain-specific knowledge and mechanisms. The models presented in this thesis can be placed in this relatively new tradition. The learning strategies used in this thesis are general purpose statistical methods. However, language-specific mechanisms and representations are used where they help modeling the phenomenon of interest.

[12] This shift affected the nativist approaches to language acquisition as well. Statistical learning and statistical learning frameworks are used by some recent studies with clear statements of nativism (Niyogi, 2006; Yang, 2002).

2.3 Summary and discussion

Language acquisition is a long and complex process interacting with many other cognitive and social phenomena. As a result, defining and testing a precise theory of the complete process of language acquisition is not (yet) an attainable goal. In practice, we put forward theories and create models of parts of the language acquisition process. The general theories of language acquisition discussed in the previous section do not actually describe a complete formal model of the language acquisition process in detail. Rather, they define some general principles and convictions regarding the nature of the problem. The precise models of a particular aspect of language acquisition subscribing to one of these theories tend to follow these principles and convictions. The merits of the general theory are not testable directly, but we can examine how well they explain smaller, more concrete aspects of language acquisition.

A few examples of particular questions of language acquisition, and a few representative examples of research into these question are listed below.

- How do children segment fluent speech into lexical units? (Saffran et al., 1996a, and many others reviewed in Chapter 4)
- Does exposure to isolated words help word learning? (Brent and Siskind, 2001)
- Is Bayesian learning plausible for explaining word learning? (e.g., Xu and Tenenbaum, 2007)
- Is there a difference in learning regular and irregular aspects of the language? (Marcus et al., 1992, among others)
- Is it possible to learn morphology (e.g., Goldsmith, 2001), or syntax (Klein and Manning, 2004) in an unsupervised fashion?

The research cited in this list uses a diverse set of methods, including psycholinguistics experiments, corpus analysis, and computational modeling and simulation. Nevertheless, each of them try to answer a particular question by testing a specific and well-defined theory. It should also be noted that these questions are interesting by their own right, and the answers to these questions do not necessarily support or oppose any of the general theories of language acquisition discussed in Section 2.2. Even though it is common in the literature to suggest conclusions about the nature–nurture debate based on research done on a particular aspect of the language acquisition, these conclusions are, in principle, problematic. Any modeling attempt of a part of a long and complex process has to make some assumptions about the other parts of the process that are not modeled. The assumptions of inductive biases in these models do not warrant a nativist conclusion: the knowledge or mechanisms may have been learned in a previous step. On the other hand, even if a certain aspect of the language acquisition is possible with minimal assumptions that seem to warrant an empiricist conclusion, this does not disprove nativism since it will not show that the complete language acquisition process can be modeled with the same assumptions.

This chapter started with a review of the nature–nurture debate, a debate that has been very influential on the language acquisition literature for the last 50 years. After reviewing various arguments for each side, I argued (1) that despite early victories announced by the supporters of one side or the other, the debate is open, and unlikely to be resolved soon, and (2) even though it has had a central role in shaping the language acquisition research, it says little about the way children learn languages, and keeping this debate as the main focus of the research may even have counter-productive effects (Section 2.1.4). As a result, the modeling practice that will be followed in this thesis is agnostic about what is innate and what is learned, taking no side in the nature–nurture debate. Section 2.2 briefly reviewed four approaches to language acquisition, namely principles and parameters theory, connectionist systems, usage-based theory and the statistical modeling. In closing, a few specific questions about language acquisition and common methods of studying these questions were exemplified. The next chapter will discuss the computational approaches for modeling language acquisition.

3 Formal Models of Language Acquisition

> Essentially, all models are wrong, but some are useful.
>
> Box and Draper (1986, p. 424)

Modeling natural phenomena is one of the basic activities in science. We build models of a natural phenomenon to gain a better insight into the phenomenon being modeled, and predict aspects that we do not know about it. Models are used in a wide range of research in science, but they are also used in many disciplines that are not research oriented. Just to name a few well-known examples, we note Galilean model of solar system in astronomy; Bohr model of atom in physics; animal models in medicine that are used for studying human diseases on non-human animals; econometric models in economics; the formal models of atmospheric conditions used for weather forecasts; scaled physical models of bridges, cars, and other objects that are used frequently in engineering.

Similar to the models listed above, this thesis makes use of models of some aspects of language acquisition. The particular type of models that are interesting for modeling language acquisition are computational models. In this chapter I will identify two methods to study such models. The first method is the mathematical, or analytical, study of the formal models of learning. The second method is using computational simulations, which is the primary method that this thesis follows. Besides introducing both methods of the study, I will pay special attention to the interaction of these methods and the division of labor between the methods for studying models of language acquisition. Throughout the chapter some of the themes and discussions introduced in Chapter 2 will be revisited, and some that are left for this chapter will be discussed in detail.

The organization of this chapter is as follows. The next section will provide a brief overview of use of models in science, and place the models of language acquisition we are interested within these models. After an introduction to well-known frameworks of computational learning theory in Section 3.2, Section 3.3 will relate these frameworks and the later work in the field with the question of the learnability of natural languages. Section 3.4 provides an introduction to a particular classification of computational models that will be instrumental in interpreting the results of the models described in this thesis. Section 3.5 compares two methods of studying computational models, namely the method of formal analysis employed in computational learning theory, and the computational simulations. Here, I will describe the interactions of the models, and argue that in certain cases, computational simulations provide easier and possibly better answers. Section 3.6 provides a summary.

3.1 Formal models in science

As in many practical disciplines, models are indispensable tools in science. There is no clear recipe for building and using models to understand the phenomena of interest, and use of models in science still keeps the philosophers of science busy (Frigg and Hartmann, 2009). However, modeling is a well-attested method of studying a broad range of subjects in science.

An important fact often overlooked is that models are not equivalent to the phenomena they model. There are always some simplifying assumptions (for example, physicists frequently assume no friction). As long as the effects of these assumptions are taken into consideration, the model can be used to gain useful insights into the phenomena being modeled. If the relation between the model and the real phenomena is well understood, we can derive conclusions about the experiments that we do on the model instead of in the real world. Such experiments on models are particularly useful, if the experiment in real world is not feasible because of economical (for example, space shuttle launches) or ethical reasons (for example, testing effects of a nuclear plant meltdown). Similarly, if our knowledge about language acquisition allowed us to build models that we are confident about, we would be in a better position to answer some questions that we otherwise cannot test with direct experiments, for example, determining validity or nature of the critical periods discussed in Section 2.1.3.

We do not always need a highly tested model with a well understood relationship with the phenomena it models. Even with relatively naive models, the formal modeling effort may still provide further insight into the problem by forcing us to describe our assumptions and theories[1] about the phenomenon with scrutiny, by systematic testing of alternative models, and sometimes testing model's predictions with the help of real-world experiments.

In modeling language acquisition phenomena, we are particularly interested in computational models. In other words, we model a particular cognitive process by a computational process. Not surprisingly, this is one of the main tenets of cognitive sciences (see e.g., Miller, 2003). Once we have a formally defined model of the

[1] The distinction between the terms *theory* and *model* is in general unclear (see, for example, Frigg and Hartmann, 2009), and their use among in different disciplines vary. In linguistics, the term theory tends to refer to a relatively general set of principles with some underspecified elements. I will follow the same use in this thesis, while using the term *model* only for precisely spelled out representation of the phenomenon of interest.

phenomenon we are interested in, an obvious method of working with a formal model is to investigate the relevant questions by analytical methods. A mathematical justification (e.g., a formal proof) of a certain question about the model may help us transfer this knowledge to the modeled phenomena for making predictions or understanding it better. However, in many cases finding analytical solutions is difficult, either because of the complexity of the model, or aspects of the phenomenon that are difficult to formalize. A possible solution to these problems is to make some simplifying assumptions at the cost of reduced correspondence between the model and the phenomenon. An alternative approach is to use computational simulations. Computational methods can be used to obtain solutions to certain problems iteratively, where analytical solutions are not available. However, more importantly, computational simulations allow certain aspects of the phenomenon to be modeled using weaker assumptions. A relevant example we will discuss is the input to the language learner. An analytical solution would require strong idealizations, such as well-formed strings generated by a formal grammar with a probabilistic error rate. However, a computational simulation can take a less-idealized definition of the input: the transcription of utterances recorded during child–parent communication.

In formal study of language acquisition, the analytical results come from the field of *computational learning theory*. Early results from computational learning theory, in particular the seminal work of Gold (1967), have been influential in the language acquisition literature. On the other hand, in parallel with the developments in natural language processing and machine learning, computational simulations have been gaining popularity for the past two decades (see Brent, 1996; MacWhinney, 2010, for snapshots of the state of the art in computational simulations of language acquisition).

3.2 Computational learning theory

Computational learning theory, or *learning theory* for short, explores the limits of learnability. The field began with the seminal work by Gold (1967), in which language acquisition was the central motivation. Not surprisingly, the results from learnability theory played an important role on the study of language acquisition. These results have often been used for arguing against learnability of natural languages. Unfortunately, the results from the learning theory have often been misinterpreted and misused (see Clark and Lappin, 2011b, for elaboration.).

The typical application of the results from learning theory in the language acquisition literature is related to their implications for learning syntax. We assume that learning syntax of a certain language is inducing a *mental grammar* from available input. Since we do not know exactly how grammars are represented in the human brain, we use *formal grammars* that represent certain aspects of natural language syntax adequately. In addition, since children are capable of learning any natural language, we expect all possible mental grammars to share some features, forming a *class of grammars*. However, since our knowledge is far from characterizing the class of

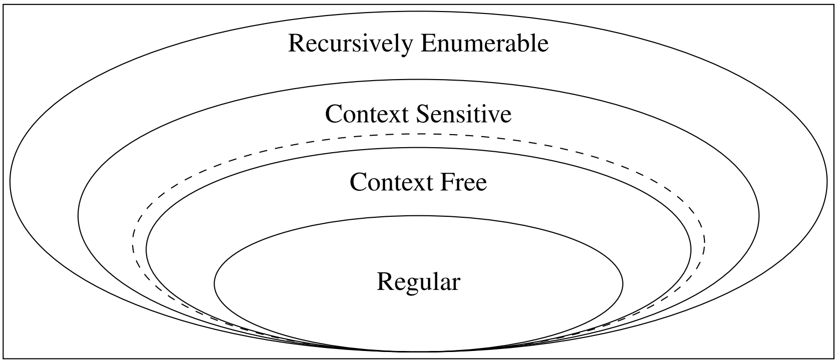


Figure 3.1: Chomsky hierarchy of language classes. The dashed ellipse represents so-called *mildly context-sensitive* languages, a subset of context-sensitive and a superset of context-free languages, which is believed to be adequate for representing natural languages. The classes are also known as *type-0* to *type-3* from the largest class to the smallest.

natural languages exactly, we turn to formal classes of grammars whose members are expressive enough to capture the syntax of all known natural languages.

This section presents the necessary formal apparatus for interpreting the results from learning theory. To make the discussion here accessible to a wider audience, it is intentionally kept informal—despite the fact the computational learning theory is a formal field of study. Formal and more comprehensive discussions of the learning theory can be found in a number of other sources (e.g., Clark and Lappin, 2011a,b; Jain et al., 1999; Kearns and Vazirani, 1994; Osherson et al., 1984). For the remainder of this section, I will give a brief informal review of two popular frameworks of learnability in relation to the problem of language acquisition. However, before starting the discussion of formal models, a short digression to a related concept, a hierarchy of language classes defined by Chomsky (1959a) is necessary.

3.2.1 Chomsky hierarchy of languages

The Chomsky hierarchy of languages is one of Chomsky's most important contributions to both linguistics and computer science (Chomsky, 1959a). This hierarchy defines a set of formal language classes. Figure 3.1 depicts this hierarchy. Each class in this hierarchy is the proper subset of the larger class. The larger classes are more descriptive, but their computational processing is more demanding. This particular classification of the language classes has some attractive formal properties, and each class corresponds to a certain type of abstract computational device that is capable of recognizing and generating the languages in the corresponding class. The details of these formal language classes and the abstract machines are not important for our purposes here. They are a well-established part of theories of formal languages and described in detail in textbooks such as Hopcroft et al. (2001) or Davis et al. (1994). Two aspects of this hierarchy are crucial for our discussion here. First, even the smallest class in this hierarchy, *regular languages*, includes an infinite number of finite and non-finite languages. That is, a regular language is a set of either a finite or infinite number of strings (e.g., sentences), and the class of regular languages has infinitely many of these languages. Second, in this hierarchy, the class that is adequate enough to represent all (known) natural languages is considered to be a subset of context-sensitive languages which is called *mildly context sensitive* languages (the dashed ellipse in Figure 3.1).[2]

Formalizing languages in this manner has proven to be fruitful both in linguistics and computer science, and this classification is central to formal study of learnability of languages. However, I want to close this brief description with two cautionary notes that we will return to in Section 3.3: (1) The Chomsky hierarchy is only one of the many ways of classifying formal languages. There are many other ways to define similar hierarchies on formal languages that cross-cut this classification. (2) The Chomsky hierarchy is a hierarchy of *formal* languages. Even though it has been an important tool in the study of language, none of these classes exactly match the class of natural languages.

3.2.2 Identification in the limit

The most popular results from learning theory in the language acquisition literature are from Gold (1967). The framework of learning introduced by Gold (1967) is called *identification in the limit* (IIL). Although many developments have been suggested since, the building blocks of this framework are still used in many studies in learning theory. Furthermore, this paper is the most cited learning theory work in the language acquisition literature.

In this framework, the learning task is viewed as identifying the correct language among a class of languages that the input language belongs to.

- The learner is presented with grammatically correct sentences from a target language he, she or it is supposed to identify.
- The learner knows the set (class) of languages that the target language is drawn from. It can test if a given string belongs to any of these languages or not. However, it does not know which language is the target one.
- The learner receives one input sentence at a time. The only restriction regarding the presentation of the input is that all grammatical sentences have to be presented

[2] Initially, context-free languages were considered to be adequate for representing natural languages. However, a small number of linguistic constructions cannot be represented using context-free languages. A typical example is the cross-serial dependencies found in Dutch (Bresnan et al., 1982) and Swiss-German (Shieber, 1985). Besides being able to represent these constructions, constructions requiring higher than context-free power are often used when they match better with linguistic intuitions.

at least once. Repetitions are allowed and the presentation of any grammatical sentence can be delayed indefinitely.

- After every input sentence, the learner re-evaluates his or her decision and selects a possibly different language from the class of languages.
- The learner needs to identify the target language exactly, and after it identifies the target language, no grammatical sentence should change its mind.
- The learner is expected to identify the target language with a finite amount of input sentences. However, there is no bound on the number of input sentences.

The most influential result from Gold (1967) is that *even the smallest class in the Chomsky hierarchy is not identifiable in the limit only from positive input*. More precisely, if the target class of languages includes all finite languages, and one non-finite language, the class is not identifiable in the limit only from positive input. This is the point in argumentation where all not-formally-oriented researchers get lost, and ready to accept the conclusion that this means natural languages are not learnable. I will return to the interpretation of this finding in the context of language acquisition in Section 3.3. Here, I will give an informal example, where the construction and consequences of this result can, hopefully, be understood better.

First thing to note about IIL is that it is not specific to language acquisition. Classes of grammars can be anything that describes a set of objects. Grammars, as a result, are nothing but sets of objects, not even necessarily sentences as we take it in linguistics. Observing that we formulate the following problem: in a boring cocktail party, a mathematician offers a linguist friend to play a game, and explains the rules: the mathematician picks a set of natural number sequences, something like ‘all sequences formed by even numbers’, and presents example sequences to the linguist whose task is to guess the set of the sequences that the example sequences are drawn. To win, the linguist needs to find the correct set of sequences at some point, and no matter which sequence from the correct set is presented, he should not change his mind. Despite not being keen on games on numbers, with some hope that the game may have some linguistic consequences, the linguists accepts the challenge. The mathematician presents following sequences,

- `7, 11, 13, 17`
- `5, 7, 11, 13`
- `13, 17, 19, 23`

What is the best bet that the linguist can make at each step? As most mathematicians would be tempted to suggest, is ‘ordered consecutive sequences of prime numbers’ a good hypothesis? The answer is no, this is too general based on the given evidence. It can, for example, be sequences of odd prime numbers (prime numbers except two), or ordered but not consecutive odd numbers, or just the set of sequences that were given so far. If the linguist prefers a conservative strategy, the last hypothesis is the right way to go. However, if the target set is the sequences of ordered consecutive prime numbers, then he will never guess it using the conservative approach. There are infinitely many finite subsets of the target set. On the other hand, if the linguist chooses a more general hypothesis, e.g., series of prime numbers, then there are infinitely many sets of sequences that are compatible with his hypothesis. If any of these is the target, then the linguist will have no reason to narrow down his choice, and he will never guess the target hypothesis. As a result, there is no guaranteed way for the linguist to guess correctly which set of numeric sequences is the correct set. He cannot identify, or learn, the underlying set of sequences the input comes from.

As this informal example demonstrates, the IIL framework does not model language acquisition narrowly, but rather much more general learning settings. The negative IIL result just demonstrated is due to the fact that the target class (the set of sequences of numbers) we choose had at least one infinite set (set of sequences formed by the ordered, consecutive prime numbers) and an infinite number of finite sets (particularly, all finite subsets of the infinite set). Since, all linguistically interesting classes of languages in the Chomsky hierarchy have this property, they are not identifiable in the limit.

Moreover, IIL has a number of other problems when applied to modeling language acquisition. The following list summarizes these problems. Some of which (the items one through three) had already been raised by Gold (1967).

1. The class of natural languages is likely to be much smaller than the language classes studied by Gold (1967).
2. The child may receive (indirect) negative evidence.
3. In IIL, the learner is expected to learn from any sequence of input strings, including the ones that are intentionally designed to trick the learner. The input children receive follows a much more restricted distribution.
4. The identification criterion is too restricted. All speakers of a certain language do not necessarily share the exact same grammar.
5. Accepting identification in the *limit* as success is unrealistic. Even if a class of grammars is identifiable in the limit, it may require an unrealistically large number of input sentences that cannot be observed in the time frame available for children to acquire the language.

The subsections that follow will discuss some of these problems, and review some of the solutions suggested in the literature.

3.2.3 Probably approximately correct learning

Probably approximately correct (PAC) learning (Valiant, 1984) is another popular learning framework in learning theory literature. The PAC framework differs from the IIL in three major ways. First, the *probably correct* learning states that learner is not required to learn from all possible input sequences. If learner fails to learn from sequences of inputs with a low probability, it does not count as a failure. Second, the *approximately correct* learning criterion relaxes the exact identification requirement in the IIL. In PAC learning, it is enough for the learner to converge to a language that approximates the target language with a small error. The last major difference of PAC learning from the IIL is that it requires learning to succeed with bounded input.

Compared to IIL, PAC learning is more suitable for many learning applications. Furthermore, it turns out that the PAC-learnability of a problem can be reduced to a combinatorial measure called *Vapnik-Chervonenkis dimension* (VC dimension, Vapnik and Chervonenkis, 1971). If the hypothesis space has a finite VC dimension, it is learnable in the PAC framework. This makes it easy to prove learnability results for certain applications. As a result, the PAC framework has been instrumental in developing and testing machine learning methods. As well as the PAC learning framework, detailed explanations of the VC dimension can be found in textbooks such as Kearns and Vazirani (1994). For our purposes here, it is important to note that VC dimension can be characterized only with labeled input, for example, in case of learning syntax, all sentences have to be labeled as grammatical or ungrammatical.

The improvements listed above make PAC learning more suitable for many practical learning scenarios compared to the IIL. However, it is still difficult to arrive at conclusions regarding child language acquisition based on models studied in this framework. I will briefly mention the shortcomings of the framework for modeling language acquisition. A comprehensive discussion of PAC learning framework from the perspective of the language acquisition problem can be found in Clark and Lappin (2011b, chapter 5).

The PAC framework allows learner to fail on some unlikely input sequences, but it does not restrict the input distribution. In language acquisition, one expects the distribution of input to have certain restrictions. With respect to the efficiency of learning, the PAC framework imposes bounds on the input required. However, these bounds are not restrictive enough for modeling language acquisition. The learner is expected to converge using input proportional to some polynomial function on the complexity of the class of concepts, e.g., grammars, to be learned. However, how this bound relates to the child language acquisition timeline is rather unclear, and in general, the polynomial bound on the input does not necessarily guarantee efficient learnability. I will return to these limitations in Section 3.3.

Even though I pointed out that the PAC framework is not perfectly suitable for modeling language acquisition, the question whether the classes of languages in the Chomsky hierarchy are learnable under PAC framework is an interesting question on its own right. The answer is no. This can be concluded easily from the fact that the class of all finite languages (a subset of regular languages) has infinite VC dimension (see, for example, Niyogi, 2006, p. 79 for a more formal statement of this result).

3.3 Learning theory and the learnability of natural languages

In Chapter 2, I discussed the arguments against the learnability of natural languages, and argued that most of these arguments are informal in nature, and difficult to decide. However, it is common to argue for a nativist standpoint based on the results from learning theory. A recent example from a highly influential article by Hauser, Chomsky and Fitch (2002) suggests that

> ... there are in principle infinitely many target systems (potential I-languages) consistent with the data of experience, and unless the search space and acquisition mechanisms are constrained, selection among them is impossible. A version of the problem has been formalized by Gold (100) and more recently and rigorously explored by Nowak and colleagues (72–75).

The argument, in more plain words, is that the results by Gold (1967) and subsequent research in learning theory support the argument that human languages are not learnable (unless the learner is constrained in certain ways).[3] This is just one of many examples of similar claims (Clark and Lappin, 2011b, present many more of the similar quotes from the literature). In this section, the question that I would like to return to is, ‘Do results from learning theory support one of the positions in the nature–nurture debate?’ Section 3.2.2 pointed out that as valuable as it is as an early work on learnability the IIL framework, Gold (1967), is not a good model of practical learning. Later work in learning, such as the PAC framework, improves some of the unrealistic assumptions of IIL, but it is not unproblematic in modeling language acquisition. Below, I will revisit these problems, and review some of the more relevant work from the learning theory literature.

The first problem is that IIL requires exact convergence to the target. In child language acquisition, this translates to the requirement that child learns exactly the same grammar his/her parents use. This begs the question of which parent, since it is no surprise that the child will generalize to a grammar somewhat different than the adults he/she communicates with. Furthermore, it is reasonable to expect that our grammars fluctuate even in adulthood. By analyzing Christmas messages of Queen Elisabeth II for a thirty year period, Harrington et al. (2000) demonstrates that even people whose language is taken as reference, are subject to this fluctuation. This problem is not peculiar to language acquisition. In many other learning tasks, approximately correct learning is what we are happy to accept, and the PAC framework acknowledges this. In PAC learning the learner is expected to make a small number of mistakes, and the error rate is required to drop as more input is provided.

[3] Incidentally, the research by Nowak and colleagues cited in this quote (Nowak and Komarova, 2001; Nowak et al., 2001, 2002, 2000) is concerned with the evolutionary conditions that would lead to a communication system with combinatorial syntax. These studies touch some of the issues from learning theory but they are not the best representative work from learning theory after Gold (1967).

A second problem is related to the efficiency of learning. The IIL framework does not impose bounds on the amount of input and time required for learning. This means that even if we have positive IIL results, if we care about learning in a limited time and limited exposure to input, we cannot conclude that the result is relevant to child language acquisition. The PAC framework requires learning to succeed with an amount of input-data proportional to the complexity of the class being learned. More precisely, the input is bounded by a polynomial function of the representation size of the target to be learned. This is also an improvement over IIL. However, it should be noted that the adequacy of this bound depends on the learning problem at hand. The complexity indicated by different polynomial functions cover a wide range on required input. For some problems any polynomial bound, indicated by high-order polynomials, may be adequate. However, in other cases, the amount of input required by a modest polynomial bound may not be available. Furthermore, the bound imposed by the PAC framework is on the input. It determines the *informational complexity* of the problem. Even if informational complexity is low, learning may be only possible with computational resources, such as memory or processing power, that are not available to children. As a result, the learnability results using either framework do not guarantee efficient learnability.

Another problem with both frameworks is the way they formalize the input. In IIL, the learner is expected to learn from any valid sequence of inputs, including the sequences designed to mislead the learner. Clark and Lappin (2011b) discuss a made-up example of a child whose only input is repetitions of 'shut up'. Surely, no one would be surprised if the child failed to learn to speak by the age of three.[4] The PAC framework does not require learning in such anomalous cases. However, it still requires learning to be distribution free. The learner is expected to learn from any distribution of grammatical input sentences. This includes arbitrary distributions, such as samples from sentences of length six or more, or samples of sentences from scientific literature. A model's assumptions about input are important in its correspondence with the real world learning experience. Both standard IIL and PAC frameworks, in this respect, fail to constrain input to a relevant set. The characterization of the input is important for the purposes of this dissertation, and I will revisit the problem in detail, and review some of the solutions offered in the literature in Section 3.3.2.

Besides some inherent problems regarding the standard learning frameworks, the above discussion points out another important factor for the formal modeling approach: the way we model the learning setting matters. To be able to define a learning problem formally, we at least need to define two aspects of the problem:

1. The object to be learned.
2. The learning environment, particularly the input provided to the learner.

[4] If we follow IIL strictly, though, we should not be worried. Three years may be too early, it is fine as long as he/she acquires the language before an age less than infinity.

To prove that languages are learnable under these assumptions, we need to define an algorithm with an acceptable time and space complexity that learns the target object with the available input. A negative result can be obtained proving that no such algorithm exists. The next two subsections discuss typical choices made in learning theory literature with respect to these aspects of the language acquisition.

3.3.1 The class of natural languages and learnability

While introducing the Chomsky hierarchy in Section 3.2.1, I have stressed that despite its usefulness in many areas of research, this hierarchy is most interesting because of its formal properties, rather than their relevance to human languages.

Even though the class of mildly context sensitive languages is considered to be adequate for representing syntax of human languages, this does not mean that it is the only adequate set. There are many other ways of defining formal language classes with this property. Grammar formalisms such as lexical functional grammar (LFG, Bresnan, 2001), head-driven phrase structure grammar (HPSG, Pollard and Sag, 1987), and combinatory categorial grammar (CCG, Steedman, 2000) define examples of such classes. These formalisms deal with most (if not all) known syntactic phenomena in natural languages in their own way, but they are not equivalent.[5]

Besides the fact that there are many adequate formalisms, not all languages in the class of mildly context sensitive languages, or even in regular languages, are possible natural languages. For example, by definition, all finite sets are members of regular languages. Just to give an example, we can define a finite language based on a cryptographic function over the words in the lexicon. By the virtue of being finite, this language is a regular language. Since both the IIL and PAC frameworks require all languages in the class to be learnable, we require learner to be capable of learning this type of languages as well. However, we do not have any reason to assume that a language defined as a long but finite set, whose elements are random sequences of words to be learnable by humans. Taking these facts into consideration, if we could properly formalize the class of human languages, their relationship with the classes in the Chomsky hierarchy would not be a proper inclusion relation. Rather, we expect a proper formalization of the class of human languages to cross cut the Chomsky hierarchy. Figure 3.2 depicts the place of the class of human languages in the Chomsky hierarchy.

The conclusion so far is that the mismatch between the classes of human languages and the classes in the Chomsky hierarchy makes the validity of arguments that depend on results obtained on Chomsky hierarchy, at best, doubtful.

The question is, then, whether there are other, linguistically interesting, classes of languages that are learnable. The answer to this question is positive; there are numerous

[5] Some of these language formalisms are overly expressive, they require power of Turing machines. However, the main point for our discussion is that these formalisms, and the variations of them, do not necessarily fit into the Chomsky hierarchy.

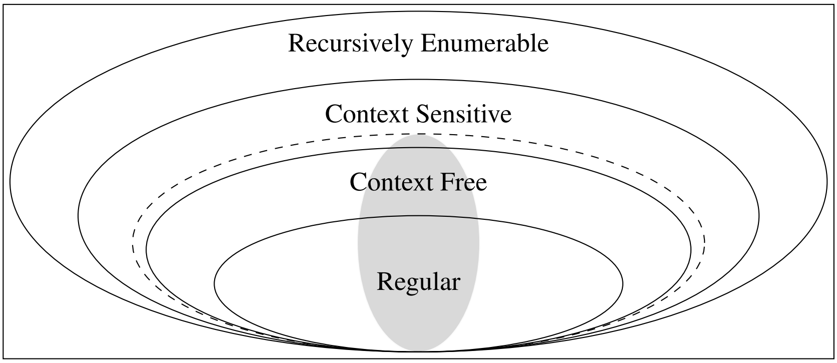


Figure 3.2: The Chomsky hierarchy presented in Figure 3.1 with an indication of the place of the human languages (the shaded area) in this hierarchy.

results in the learning theory literature with positive results on interesting classes of languages. For example, using a probabilistic version of IIL, Horning (1969) proved that the class of probabilistic context-free languages are learnable. The results obtained by Shinohara (1994) and Kanazawa (1996) are examples of positive IIL results by imposing rather reasonable constraints on more general grammar formalisms. For example, Kanazawa (1996) presents a positive IIL result for categorial grammars (a grammar formalism weakly equivalent to context-free grammars) if the number of categories assigned to a word has a finite bound. Another relevant strand of work by Angluin (1982, 1988a,b) presents positive results using modified versions of IIL, on certain subsets of regular and context-free languages. The language classes studied by Angluin in these works are not powerful enough to represent human languages. However, more recently, a similar line of research by Clark and Eyraud (2007) and Clark (2010) presented positive results for increasingly complex classes of languages.

The positive results obtained in some of these studies (for example, Horning, 1969; Kanazawa, 1996; Shinohara, 1994) are proofs about highly expressive sets of grammars. However, these proofs are proofs of learnability in principle, and they do not prove efficient learnability (see, for example, Costa Florêncio, 2003, for an analysis of Kanazawa's results). However, the line of research following Angluin (1982) on increasingly complex subsets of language classes on Chomsky hierarchy provides proofs of efficient learnability. Admittedly, the proofs are on rather restricted classes of grammars, and even the classes analyzed in the latest studies (e.g., Clark, 2010) are probably too small to capture all syntactic phenomena in natural languages. However, this approach to grammar learning, i.e., defining a restricted class of languages and finding algorithms that learn the class efficiently, provides insights regarding what is learnable, and which procedures are efficient at learning them.

Rather than a formal characterization of the language class, another common nativist argument is formulated as follows: if the set of possible human languages is infinite, then choosing the correct language among them is impossible (for example, the quote in page 31). Note that existence of infinitely many languages in a language class is not enough by itself for the conclusion that the language class is not learnable according to IIL. This has to be accompanied by another condition that some of the languages in the class are infinite. The claim that human languages contain infinitely many possible sentences serves this purpose.[6] Indeed, these two conditions are enough to conclude that the class is not learnable according to the IIL framework. However, it does not necessarily make the problem unlearnable in the PAC framework. Neither an infinite number of sentences, nor an infinite number of hypotheses entails an infinite VC dimension. However, as argued throughout this chapter, the learnability results obtained using standard frameworks are far from being conclusive. To be able to arrive at any conclusions, we need to model the language acquisition process in more detail.

3.3.2 The input to the language learner

In a formal model of learning, the characterization of the input affects our conclusions. Not surprisingly, the nativist claims of unlearnability generally come together with certain assumptions about inadequacy of the input. This argument, known as *argument from the poverty of stimulus* (APS), has been the main motivation for many nativist theories of language acquisition, and a source of active discussion in the nature–nurture debate. I will only provide a selective discussion of the APS here. A nativist introduction can be found in Cowie (2010); in their critique of the argument, Pullum and Scholz (2002) also summarize various versions of the APS claim found in the literature; and a book-length treatment of the subject can be found in Clark and Lappin (2011b).

The APS claims in the literature come in a number of forms. First, it is claimed that the language children hear is 'degenerate in quality' (Chomsky, 1965, p. 31), which makes the learning task more difficult. Empirical evidence suggests that this claim is wrong. Contrary to the claim, the child-directed speech seems to contain fewer errors, and adults seem to be adjusting the complexity of their utterances according to children's level of the language (Snow, 1972). Second, it is claimed that evidence required for learning certain type of grammatical constructions are not available in the input. Pullum and Scholz (2002) show through corpus analysis that the evidence that is claimed to be missing can be found in real language data if one looks closer. Furthermore, a large number of computational simulations show that with modest assumptions about the learner's nature, these grammatical constructions can be learned without the evidence claimed to be necessary (e.g., Clark and Eyraud, 2006; Yao et al., 2009). Finally, it is claimed that children do not receive *negative evidence*. This is based on the observation that children only hear grammatical utterances (with

[6] The assumption that natural languages are infinite is a widely held assumption in linguistic literature. However, it is not uncontested. See Pullum and Scholz (2010) for a critical treatment of this assumption.

some noise) in their language. Nobody gives them examples of ‘what not to say’. A possible source of negative evidence is the corrections they receive for the mistakes they make. However, it is argued that children receive no reliable corrections for the grammatical mistakes they make, and when they do, they do not seem to care about the corrections (Marcus, 1993; Pinker, 1989). This claim is not without controversy either (for example, Chouinard and Clark, 2003), however the assumption that children do not get negative evidence is more widely accepted compared to assumptions regarding degenerate and insufficient input.

The APS claims about the lack of negative evidence is used in connection with the results from Gold (1967) that even regular languages are not IIL only from positive evidence. With negative and positive input, powerful classes of languages become IIL. For example, the proof that the class of *recursive languages* (a superset of context sensitive languages) is IIL was presented in the same article by Gold (1967). The standard PAC framework assumes availability of both negative and positive evidence, and the learnability results from IIL framework cannot be carried over to the PAC framework easily. However, it is clear that availability of both positive and negative evidence facilitates learning. We may follow the common assumption that there is no direct negative evidence in the input to children, but, is there another source of information that may function as negative evidence? If we can identify such an information source by formal means, then it is an empirical question to verify its use in language acquisition. The following is a selection from such sources of information discussed in the literature.

3.3.2.1 Queries

A group of formal studies allows the learner to query additional facts about the language. I will briefly mention two well-known query methods in the learning theory literature (due to Angluin, 1987, 1988a). The first one, *equivalence queries* allows learner to verify his/her guess about the target language. After every input utterance, the learner guesses the language (by, e.g., writing down a set of grammar rules, or naming the language). The learner is allowed to query the validity of his/her guess (e.g. by asking ‘is it English?’). If the guess is the target language, it is confirmed. If it is wrong, a counter example that is in the target language, but not in learner’s hypothesis, is provided. The second one, *membership queries*, allows learner to generate a string and check its validity. The membership queries are interesting, since they model a more actively-communicating learner. It is possible to get efficient learnability results for larger classes of languages under these assumptions. However, both query methods are of limited value for modeling the child language acquisition setting.[7]

[7] Clark and Lappin (2011b, chapter 6) remark, however, that the membership queries can, at least in principle, be replaced by probabilistic data. This may mean that under the assumption that input to the child is probabilistic, learning theory results with membership queries might be representing child language acquisition better.

3.3.2.2 Probabilistic input

Language data is probabilistic. Even though precise characterization may be difficult, the linguistic units follow certain probability distributions. For example, the Zipfian distribution (Zipf, 1935/1965, see also Miller, 1996), is a time-tested distribution that models distributions of words in linguistic data quite well. The language children are exposed to during the acquisition process is not an exception. Furthermore, increasingly more studies in psycholinguistics show that children make use of statistical information in the input (e.g., Saffran et al., 1996a; Thompson and Newport, 2007).

The availability and awareness of probabilistic input may compensate for the need for negative evidence in the formal models described above. In other words probabilistic data may serve as *indirect negative evidence*. We have already seen that with probabilistic assumptions, the class of context free languages are IIL (Horning, 1969). In general, from a formal perspective, probabilistic data allows stronger learnability results (see Clark and Lappin, 2011b, chapter 6 for a comprehensive discussion).[8] Interpreting lack of evidence as negative evidence can also explain some of the language acquisition phenomena. For example, even though infants up to 6-month old seem to be sensitive to sound contrasts that do not exist in their language, they lose this sensitivity around a year of age (Werker and Tees, 1984). This is an indication that unavailability of evidence is used by children as a source for generalization.

It is clear that input, especially the probabilistic distributions of input, is modeled poorly by formal models in the literature. Most results in the literature are based on distribution-free learning, which is clearly unrealistic for the child language acquisition case. There has been some work on restricting the possible distributions (e.g., Clark and Thollard, 2004). However, it is difficult to choose justifiable distributions for complex aspects of language data, and the choices of distributions, or families of distributions, used in the literature tend to be arbitrary. I will return to this discussion in Section 3.5, and argue that computational simulations offer a better method of modeling the input for the child language acquisition: by using real-world linguistic data.

3.3.3 Are natural languages provably (un)learnable?

This section reviewed a number of formal studies on learnability of languages that relate to the question: based on results from learning theory, are natural languages provably learnable or unlearnable? The short answer is, there is no definitive proof for either position.

The usual nativist claim that natural languages are unlearnable based on the results of Gold (1967) is unwarranted. To be able to get to this conclusion we need to assume:

[8] In computer science, it is well known that probabilistic assumptions often make otherwise intractable computational tasks easier. Some problems that are intractable for Turing machines, for example, testing whether a number is prime or not, are solvable (with high probability) in polynomial time on *probabilistic Turing machines* (Solovay and Strassen, 1977).

1. All users of a language share the same grammar, and children are required to learn this grammar exactly.
2. The input and time available to the child are unlimited.
3. The target grammar to be learned is a grammar from the Chomsky hierarchy of grammars(such as mildly context sensitive grammars).
4. Except their grammaticality, there is no restriction on the input utterances children receive. Children should learn the language even with misleading input sequences as long as it is grammatical.

Some of these assumptions have been relaxed, and arguably have been drawn closer to child acquisition setting. However, the review I provided in this section points out that we are still a long way from a reasonably formal characterization of the child language acquisition process. The formalization of the last two, the class of natural languages and the input to the language learner, are particularly difficult. First, we do not know how languages are represented in human brain. Despite centuries of hard work, even for well-studied languages like English, descriptive grammars we come up with are far from complete. As Sapir (1921) famously puts it, 'all grammars leak'. That being said, a particular strand of work in learnability theory that seeks provably and efficiently learnable grammars (e.g., Angluin, 1982; Clark, 2010) may be very fruitful in finding good candidate grammar classes, or properties of these classes that are learnable. Other things being equal, i.e., if there is no empirical evidence against them, provably learnable representations should be favored for modeling language acquisition.

Second, formalizing linguistic input is another big challenge for formal models. The linguistic data tends to be complex and messy (Halevy et al., 2009), and it is difficult to develop simple and mathematically attractive models of all aspects of the linguistic data. Expressing language data with formulas is, indeed, difficult, but we are not helpless. We can model the input to the child using real examples of child-directed speech, and study these models using computational simulations.

3.4 What do computational models explain?

Like Chapter 2, this chapter may appear to be dominated by the nature–nurture debate. This is partially due to the central position this debate occupies in the language acquisition literature. But it is also because of the fact that some formal arguments, such as the argument from poverty of stimulus discussed in Section 3.3.2, are answered best with computational models. However, computational models can, in principle, answer more questions than the question of innateness. A selection of these questions has been listed in Section 2.3. In this section I will briefly describe a general classification of the models according to what type of questions they are designed to answer.

In modeling cognitive phenomena (and many other natural phenomena), we seek answers at different levels. The distinction made by Marr (1982) has been an influential way to classify the computational models according to the level at which they provide explanations. Marr (1982) suggested three levels at which an information processing system can be studied. First, the *computational level* seeks answers to the questions *what* and *why*, focusing on the more abstract explanation of the computational system. Second, the *algorithmic level* is concerned with the question of *how*, focusing on the procedures and input/output representations used by the system. The third level is the *implementation level*, which is concerned with how the system is realized physically. Marr (1982) presented explanations for a cash register at different levels as an example. At the *computational* level, an explanation of what the cash register does can be explained by addition. Notice that this explanation is independent of the explanation at the *algorithmic* level, e.g., whether the addition is carried out using binary or decimal digits. In turn, explanation at the algorithmic level is independent of the *implementation* level, for example, whether the hardware is realized using mechanical parts or electronic circuits.

In descriptions of the models in the literature, it is often not clear at which level the model is providing explanations. Most computational models of language acquisition fall into the computational or algorithmic level, even though connectionist systems are sometimes claimed to seek answers at the implementation level. The models that will be described in this thesis fit best into Marr's algorithmic level.

3.5 Computational simulations

The methods discussed in Section 3.2 and 3.3 are studies of language acquisition that stem from the field of computational learning theory. These studies typically define the computational system mathematically, and the questions of interest are investigated using a mathematical analysis of the model. A related method of study is to define a model formally, but to use computational simulations, instead of studying it analytically. Like computational learning theory, the early work on computational simulations started in 1960s (e.g., Olivier, 1968). However, there has been an increasing number of computational simulations of various aspects of language acquisition for the last two decades (Brent, 1996; MacWhinney, 2010). While formal analyses are better suited for finding provably working models, computational simulations are better suited for modeling aspects of language acquisition where it is difficult to find mathematical formalizations, such as the input to the child during language acquisition. A subset of the computational models that are related to the simulations reported in this thesis will be discussed in Chapter 5.

Both methods, mathematical analysis and computational simulations, require a careful description of the problem. Arguably, the analytical study of a model provides better insights into the problem being modeled. Computational simulations can treat certain parts of the model as a black box, and may not explain what the contents of the black box are, or how it relates to the problem (this is, for example, one of the criticisms against artificial neural networks). However, this can also work to the advantage of the modeler. For example, one can leave a certain aspect of the model, such as a certain parameter value, relatively underspecified, and let the simulation explore this aspect.

Furthermore, if some data the model uses is difficult to formalize mathematically, computational simulations allow modeling the data with a sample. As I argued in Section 3.3.3, this particular property of computational simulations is useful in modeling language acquisition phenomena. We cannot easily express the linguistic input to the children with mathematical formulas. However, we can easily take an appropriate sample for the problem at hand from the growing body of child-directed corpora, for example from CHILDES (MacWhinney and Snow, 1985). The results still depend on how representative the chosen sample is. However, in most cases the idealizations caused by samples taken from real-world data are not as restrictive as mathematical formulations of the data.

Besides the use of real-world data samples, the models designed for computational simulations can afford to experiment with more complex learning algorithms. Naturally, learning algorithms with formal proofs of convergence should be used when available. However, for many complicated problems, obtaining these proofs is difficult.[9] With the availability of cheap and powerful computation, it is easier, in most cases, to test the performance of the algorithms empirically using computational simulations.

Computational simulations of language acquisition are closely related to the natural language processing (NLP) applications that are designed for similar purposes. However, the NLP applications do not have to be compatible with the child language acquisition process. For example, it is typical in this field to use supervised learning methods that make use of informative annotations that are unrealistic for modeling language acquisition. Most learning models in the NLP literature are designed to solve an engineering problem without any interest in child language acquisition. However, solutions of similar problems in different fields often inform each other.

Even though most NLP work is not suitable as a model of human language acquisition, the unsupervised models in this field can, in fact, provide explanations to the language acquisition process at Marr's computational level described in the preceding section. An example relevant to the discussions in this chapter is the unsupervised learning of syntax. For example, Klein and Manning (2004, 2005) presented successful experiments with learning syntax in an unsupervised manner. Note that these models do not model how children learn syntactic rules. However, they show that using an unannotated corpus a certain level of success can be achieved in learning syntactic structure found in corpus. If we accept the level of success reported in these examples, this already indicates that the popular negative identification in the limit (IIL) results are not relevant when a sample of real-world language data is taken as input.[10]

[9] Formal proofs can also be discouraging since typically these proofs are concerned with worst-case scenarios. There are examples of computational tasks with complex worst-case running time, but rather good typical running times. Examples include well known sorting algorithm *quick sort* (Knuth, 1997, section 5.2.2) and parse recovery algorithm in a chart parser (Jurafsky and Martin, 2008, chapter 13).

[10] As a matter of fact, even the supervised models of learning syntax (e.g., Clark and Curran, 2003;

Similar results can be derived from computational models that specifically model aspects of human language acquisition. For example, Clark and Eyraud (2006) present a model that learns the formation of the English interrogative question (introduced in Section 2.2.1) from a set of sentences that are claimed to be insufficient for learning this phenomenon. In a similar study, as well as the English interrogative questions, Yao et al. (2009) presented a model that learns another syntactic phenomenon, English auxiliary order, that is frequently claimed to be unlearnable from the data available to children. Crucially, both studies assume relatively simple inductive biases. Although the question of whether the inductive biases of these models support empiricism or nativism is difficult (if not impossible) to answer, they demonstrate clearly that once the phenomenon of interest is modeled carefully, the conclusions from informal arguments may change drastically.

One last issue I would like to address here is the skepticism towards the utility of computational models. I believe the models from other fields listed at the introduction to this chapter present a clear case for modeling practice in general. We rarely doubt the utility of models, such as a mathematical model of the orbit of a planet, or an animal model for testing effects of a particular chemical on humans. The parallel modeling practice in language acquisition is computational models. Admittedly, we do not have computational models of language acquisition with similar predictive power. On the other hand, this is not a problem inherent to computational modeling practice. If our knowledge of language acquisition were as precise as our knowledge of astronomy, for example, then we would be able to build computational models with highly precise predictions. This does not mean that we should better stop the modeling practice until we have a level of knowledge adequate for a certain predictive accuracy. As discussed throughout this chapter, the predictions are not the only function of the modeling. Even predictions that arise from rather simple modeling practices can provide further insights, and raise useful research questions that may otherwise be overlooked. For example, a decision taken in computational modeling of segmentation by Brent (1999a), was to insert complete utterances into the lexicon if the input utterance cannot be segmented. Questioning the relevance of this modeling decision, Dahan and Brent (1999) tested this experimentally and found parallels between the model and the humans in this task.

The idealizations a computational model makes are another source of skepticism. Computational simulations generally make certain idealizations about certain aspects of the process or input to the language learner. However, it should be noted that idealizations are part of all other methods of study. For example, since use of artificial languages provide a better controlled experimental setting, many experimental studies use artificial languages to test human linguistic performance. On a related note, the experimental methods typically study a very specific question in a well-controlled

Collins, 1999) can be taken as an indication of failure of negative IIL results. For the IIL framework, as long as negative evidence is not supplied, supervised learning, does not change the conclusions. Learning from positive examples of a set of sentences, or positive examples of a set of sentences paired with a representation of their syntax does not make a difference.

setting. Computational simulations in the literature tend to model a relatively general phenomena (cf. the experimental study by Saffran et al. (1996a) and the computational models of segmentation reviewed in Chapter 5). As a result, studies in psycholinguistics tend to provide precise answers to specific questions, while most computational simulations model a more general question. This difference, rather than indicating a weakness of computational models, rather suggests complementary use of these two methods.

The aim of this section has been to reflect on the role of computational simulations as a method of study of language acquisition. I argued that the formal analysis and computational simulations are two complementary ways to study models of language acquisition. However, the simulation approach is at an advantage in modeling aspects of language acquisition that are particularly difficult to express using mathematical means. In general, despite their shortcomings, the computational models are a valuable tool in studying language acquisition. I will return to the issues raised in this section throughout the rest of this thesis.

3.6 Summary

Modeling is a common practice in science. We build models of natural phenomena to learn about the phenomena, as well as to make predictions about what we do not know about them. The computational models that will be presented in this thesis are examples of this practice.

In this chapter, I distinguished two different ways to study computational models of language acquisition. First, the formal, analytical studies of these models in computational learning theory, and second, the computational simulation practice that has become an indispensable method in modeling various cognitive phenomena. The main bulk of the chapter reviewed the first method of study along with its relation to the question of whether natural languages are learnable or not. Computational learning theory has advanced considerably since its inception by Gold (1967). However, with regard to the learnability question, arriving at any strong conclusion based on work on learning theory is not yet possible. On the other hand, the function of computational modeling is not only answering questions of learnability. These models, as they mature, may provide more insight into other interesting questions about language acquisition, and provide valuable predictions.

The study of computational models by formal analysis and computational simulations are complementary methods of study. However, as I argued in Section 3.5, the simulation approach is particularly useful when there are aspects of the phenomenon, such as input in language acquisition, that are difficult to formalize mathematically. This difference, however, should not be understood to mean that computational simulations are based on less-formal models. The difference emphasized in this chapter is about the method of study. Both methods of study require models to be specified formally. However, the choice of the method affects how certain idealizations, or approximations, can be made during the modeling practice. For example, noise in the input is likely to be modeled as a probabilistic bound if one choses to study the model analytically. On the other hand if simulations are to be used, one can explore a larger range of bounds in more detail, or if real-world data is used, model it with the noise in the input sample. Section 3.4 described an influential classification of computational models, (due to Marr, 1982), according to what type of questions they answer.

This chapter concludes our survey of the broader field of formal language acquisition research. The next chapter introduces the problem of segmentation, the problem which the modeling effort in this thesis will focus on.

4 Lexical Segmentation: An Introduction

> Music is a hidden arithmetic exercise of the soul, which does not know that it is counting.
>
> Gottfried Leibniz

Segmentation is crucial in language processing. In a large number of linguistic tasks one needs to segment a continuous stream into units such as words, morphemes, syllables and phonemes. The speech signal is continuous, and spoken language input does not include a single consistent marker similar to white spaces in most writing systems. However we recognize discrete units at multiple levels, e.g., phonemes, syllables, morphemes and words.

```
bljuuz                              pubuztsujjbjujbeuzjltsdbtuofui
epzpvuijoljutbljuuz                 oijtptjz
zpvdbouijoljutbljuuz                cublitipztbojjbubpfjljufzlv
jtuibubljuuz                        upjzijmululzpijfjzu
ifsfljuuzljuuztmffqjohljuuz         odhvsfzfxjigtuzquuld
mpplzpvdbotujdlzpvsgjohfsjouifsf    msphflluivgluczjbsxobjijfepufjsf
tujdlzpvsgjohfsjouibuipmfsjhiu      htjupmvszfuz
hppehjsm                            bjupvuoouuldmumhplouozsphfih
opxmfuttffxibutuifcbcztbzjoh        hoofut
```

Figure 4.1: Two input sequences representing unsegmented linguistic input.

Fluent language users identify the lexical units in speech so effortlessly that it is difficult to imagine that segmentation is a problem at all. The problem becomes apparent when one listens to an unfamiliar language, where identifying word-like units becomes close to impossible. Given that the speech signal does not include reliable indicators where one lexical unit ends and another begins, how do humans identify the lexical units, e.g. words?[1]

[1] Commonly, words are considered to be the lexical units. However, models that allow sub-word units,

To motivate the rest of the discussion and have an impression of the problem that children face, we will first work through an informal example. Consider the sequences in Figure 4.1.

The strings on the left side of Figure 4.1 are nine consecutive child directed utterances from the CHILDES database (MacWhinney and Snow, 1985). The transcriptions of the utterances are systematically (but simply!) modified to remove the advantage of the knowledge of the lexical units in English. The letters are garbled by mapping each letter systematically to a different one. All word boundaries, except the utterance boundaries, are removed from the transcriptions. This gives a first impression of the problem that a learner without knowledge of lexical units faces, e.g., an infant acquiring language. This example includes some additional information, such as distinction of letters (or phonemes in child's case), that an infant does not have at birth. On the other hand, the example also lacks some cues for boundaries, such as prosody and possible pauses, which are found in the linguistic input to children. The strings on the right side of Figure 4.1 are formed by using the same letters, however, sequences of letters and utterance boundaries are randomized.

Even though it is puzzling at first look, if we examine it carefully enough, we can find some regularities on the left side of Figure 4.1. First thing to note is that some substrings, such as `ljuuz`, `bljuuz` (which also occurs as an utterance by itself) and `epz` repeat multiple times. Another regularity we can observe is that some characters or character sequences consistently follow (or precede) others. For example, character sequences like `jo`, `ju` and `lj` are very frequent (with frequency of 8, 9 and 9, respectively), while some others are rare or not observed at all. 52 two-character sequences are observed only once, and only 91 of the 484 possible two-character sequences are observed. On the other hand, no matter how long we stare at the right side of Figure 4.1, we cannot observe similar regularities on the right since this sequence is formed by random concatenation of the same letters.[2]

Even in this small sample, we can find a number of regularities that are typical for natural languages, such as repeated strings, and principled sequences of basic units (e.g., letters). Furthermore, there are properties of the speech signal that are difficult to demonstrate on paper. The regularities demonstrated in this example are not all possible regularities that one can utilize to discover lexical units from natural language input. We will discuss others in detail, and return from time to time to this example.

The task for the language user, then, is to spot the lexical units in the continuous

i.e., morphemes, as well as multi-word units are better models of the human lexicon (see, for example, the discussion in Davis, 2006, p. 12, and the references therein). In this thesis the terms *word* and *lexical unit* are used interchangeably. It is explicitly indicated where the difference matters.

[2] Note that the sample on the right side is not completely arbitrary either. Although their location is randomized, the letters and the lengths of the utterances are chosen to be identical to the real language sample. Hence, for example, finding that the letter '*j*' is the most frequent letter on the right side just as in the left side is not surprising. For the benefit of those who have not yet discovered the cipher in Figure 4.1, the same sequences are repeated in Figure 4.2 without garbling the letters.

```
akitty                             otatysrtiiaitiadtyiksrcastneth
doyouthinkitsakitty                nhisosiy
youcanthinkitsakitty               btakhshoysaniiataoeikiteyku
isthatakitty                       toiyhiltktkyohieiyt
herekittykittysleepingkitty        ncgureyewihfstypttkc
lookyoucanstickyourfingerinthere   lrogekkthufktbyiarwnaihiedoteire
stickyourfingerinthatholeright     gsitoluryety
goodgirl                           aitoutnnttkcltlgokntnyrogehg
nowletsseewhatsthebabysaying       gnnets
```

Figure 4.2: The original sequences in Figure 4.1 where the letters are replaced with the succeeding letter according to ASCII code. The presentation is inspired by Cohen et al. (2007).

speech stream by making use of a number of noisy and sometimes conflicting regularities, or cues. For a competent speaker of a language, the task is somewhat easier: the (implicit) knowledge of the language, such as the words or possible phoneme sequences, is useful for segmentation. However, the task becomes more challenging for a learner who has only partial knowledge about the language to be learned or none at all.

The next section will provide a brief review of the field of word recognition. The studies reviewed in this section assume that the words are already known. Naturally, these studies are model of adult performance in this task. Section 4.2 will review the psycholinguistic literature most relevant to the computational models that will be developed throughout the rest of this thesis.

4.1 Word recognition

Even though the models developed in this study do not assume that learner starts the task of identifying words with a complete lexicon, the study of adult word recognition is a closely related field of research. Knowing words of the input language helps recognizing them in the continuous speech stream considerably. The informal example we have discussed at the beginning of this chapter already demonstrates that. Although it is almost impossible to identify the words in the left side of Figure 4.1, for speakers of English, the words in the left side of Figure 4.2 is easier to extract. However, the recognition of words in real world speech stream is more difficult than it seems in this example.

The difficulty of identifying words in continuous speech stems from two factors. First, the listener has to deal with a large number of acoustic hurdles, including noise and variability of speech signal due to speech rate, dialectal differences and differences in individual speakers' voice. Considering, in addition, that the word

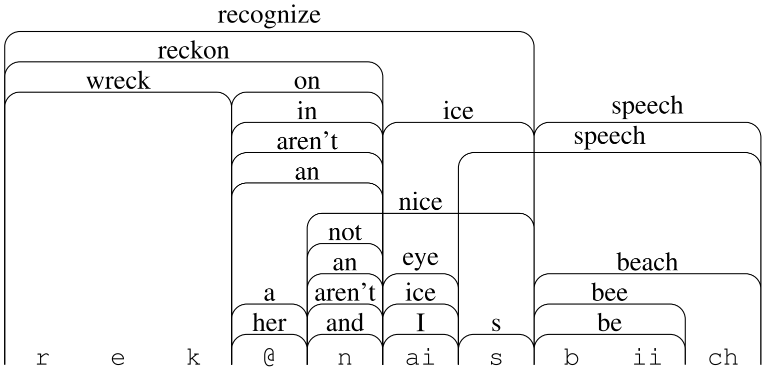


Figure 4.3: An automatic speech recognizer's attempt to segment the phrase *recognize speech*. Example re-produced from Shillcock (1995).

pronunciation (tokens) differ at least slightly from each other acoustically, even if they are heard in isolation, recognizing words in speech sound is a difficult task. Second, the input is generally compatible with multiple segmentations all supporting complete segmentation of the input utterance. Figure 4.3 demonstrates these difficulties by presenting segmentations offered by an automatic speech recognition system for the phrase *recognize speech*. This is a popular example in speech recognition literature where, among others, an alternative hypothesis during segmentation of the utterance *recognize speech* is *wreck a nice beach*.

A large body of literature exist on spoken word recognition (see Dahan and Magnuson, 2006; Davis, 2006, for comprehensive reviews). Most studies in the field are highly involved in the modularity debate. They focus on the manner in which low level perception is affected by higher level lexical knowledge. A number of phenomena demonstrate that word recognition is affected by other cognitive tasks or knowledge. In addition to the lexical knowledge, higher level linguistic processes—syntax, semantics and discourse— and non-linguistics context, e.g., related visual stimuli, affect the perception of the sounds, and recognition of the words. Some phenomena that show effects of higher level knowledge to perception of sounds are listed below.

- The word superiority effect: target phonemes are detected more quickly in words than in non-words (Rubin et al., 1976).
- The phoneme restoration effect: when certain phonemes are masked, by a cough or a buzz, listeners report hearing phonemes that are consistent with the lexical context (Warren, 1970).
- Ganong effect: an ambiguous sound in a phonetic continuum such as /t/ and /d/ is interpreted in the context that yield a lexical unit. When changed systematically from /d/ to /t/, the ambiguous sound is interpreted as /t/ earlier in the context

-ask, than in -ash (Ganong, 1980).

- Some phoneme changes (e.g., /bit/ and /pit/) detected by the listeners when a word is presented in isolation are undetected if the word is embedded in a phrasal or sentential context (Cole, 1973).
- Helpful visual context facilitates ambiguity resolution (Tanenhaus et al., 1995).

Even though the form and timing of the interaction is debated, it is uncontroversial that both high-level linguistic knowledge of the speaker, e.g., lexical units, and the low-level information from the acoustic signal interact in word recognition. All of the influential models of spoken word recognition, e.g., COHORT (Marslen-Wilson, 1987; Marslen-Wilson and Welsh, 1978), TRACE (McClelland and Elman, 1986), Shortlist (Norris, 1994), integrate both bottom-up and top-down sources of information. Differences exist in the way the recognition process incorporates the information, as well as the processing strategies. However, for our purposes here, these differences are irrelevant. All word recognition models base the identification of words in the input stream on the lexical knowledge. Words in the input stream are identified as the initial segments of the input are matched against already known words in the lexicon. These models provide varying degrees of success in predicting human performance. The weakness of word recognition models for our purposes is that they require a relatively complete lexicon, and they do not explain how infants start spotting unknown words in continuous speech stream without a lexicon.

Even though they do not offer a way to bootstrap the lexicon, the insights and mechanisms offered by these models are still relevant after the lexicon is populated by other means.

4.2 Lexical segmentation

While researchers studying word recognition were busy with the modularity debate, the focus of speech segmentation literature for the last two decades has been a debate between the role of prosodic and statistical cues for learning segmentation. Even though it is known that children are sensitive to a number of other cues, two cues have been studied extensively in psycholinguistic literature: *predictability* based on statistical relationships between consecutive sound segments, and *lexical stress*. The former is commonly called *statistics*, *statistical regularities* or *distributional statistics* in the literature.[3] The latter, lexical stress, is the most studied prosodic cue, although there are a number of other prosodic cues that are helpful for lexical segmentation.

In this section we will review some of the cues that are believed to be used by infants in constructing their initial lexicon. The cues are simply some aspects of linguistic input. They are typically low-level perceptual signals such as pauses. However, there

[3] Although it is very common to see these terms used as if they are equivalent to the predictability-based cues, statistics or distributional regularities that serve as predictability cues are just one of many other uses of statistics in language acquisition.

phoneme	**h**		**ɪ**		**y**		**z**		**k**		**w**		**ɪ**		**k**		**ə**		**r**	
SV		9		14		29		29		11		6		10		28		14		28

Figure 4.4: Example successor values for the phrase /hɪyzkwɪkər/ 'he is quicker' determined by Harris (1955).

is evidence that a broad range of cues that are derived in a longer time span, such as statistics over consecutive syllables, are also used in lexical segmentation. For a cue to be useful in extracting lexical units from continuous speech, we need to establish a number of facts:

1. The learners are sensitive to the cue.
2. The cue is useful for a learner as it is available in the input.
3. Existence of the cue in the input facilitates the segmentation task.

Although all the questions are addressed by the experimental studies to some extent, the majority of the studies in the developmental psycholinguistics are concerned with the first item above. Last two points provide a good case for computational simulations and corpus analysis.

4.2.1 Predictability and distributional regularities

Natural language utterances are not arbitrary strings of phonemes; instead they are produced by concatenating lexical units. Because of the way they are formed, natural language utterances exhibit certain statistical regularities. At least as early as Harris (1955), it was known that a simple property of natural language utterances can aid identifying the lexical units that form an unsegmented utterance:

> *Predictability within the units is high, predictability between the units is low.*

Harris operationalized this idea by introducing a measure called *successor variety* (SV). SV is simply the number of distinct phonemes that may follow a certain initial utterance segment. The higher the SV, the lower the predictability of the next phoneme, and the more likely it is to encounter a boundary after the segment. Figure 4.4 presents an example from Harris (1955), where successor values are given after each possible initial substring of the phrase /hɪyzkwɪkər/ 'he is quicker'. According to this strategy, the higher the successor value, the higher the chance of having a lexical boundary.

There was some early work on using this approach for natural language processing applications (e.g., Hafer and Weiss, 1974). However, for a long time, the idea was not investigated in developmental psycholinguistics as a possible source of information that children may use for segmentation. The influential study by Saffran, Aslin and Newport (1996a) showed that 8-month-old infants, indeed, are sensitive to this type of information and that they use it to extract word-like units from an artificial language stream without any other cues to word boundaries. Saffran et al. (1996a) used another measure of predictability which they called *transitional probability* (TP). The TP is defined over two successive units. It is simply the conditional probability of seeing a unit, e.g., a syllable, given the previous unit. Given two units l and r the transitional probability $\mathrm{TP}(l, r)$, sometimes denoted $\mathrm{TP}(l \rightarrow r)$, is defined as[4]

$$\mathrm{TP}(l, r) = P(r|l) = \frac{P(lr)}{P(l)} \approx \frac{\mathrm{frequency}(lr)}{\mathrm{frequency}(l)} \qquad (4.1)$$

For example, given the string '*bidakupadotigolabubidakugolabupadoti*' the TP values for pairs of syllables *bi–da* and *ku–pa* can be calculated as

$$\begin{aligned} \mathrm{TP}(bi, da) = P(da|bi) &= \frac{\mathrm{frequency}(bida)}{\mathrm{frequency}(da)} = \frac{2}{2} &= 1.0 \\ \mathrm{TP}(ku, pa) = P(pa|ku) &= \frac{\mathrm{frequency}(kupa)}{\mathrm{frequency}(ku)} = \frac{1}{2} &= 0.5 \end{aligned}$$

suggesting that it is more likely to have a boundary between *ku–pa* than *bi–da*.

Saffran et al. (1996a) used an artificial language stimuli constructed by concatenating 3-syllable artificial word-like units. The stimuli is formed such that the word-internal syllable transitions were completely predictable (TP = 1), while predictability was lower between the words (TP = 1/3). There were no acoustic cues indicating the word boundaries. After two minutes of familiarization, the infants reacted significantly differently to a test stream that was formed with the same words, compared to a speech stream that included same set of syllables with the same frequency of occurrence but was formed using concatenation of parts of words of the familiarization phase.

After Saffran et al. (1996a), a large number of studies confirmed that predictability based strategies are used by adults and children for learning different aspect of language (e.g., Aslin et al., 1998; Graf Estes et al., 2007; Newport and Aslin, 2004; Perruchet and Desaulty, 2008; Thiessen and Saffran, 2003; Thompson and Newport, 2007).

An important feature of the predictability based strategy is that it does not require any initial lexical knowledge, and it is completely language independent. A learner with the knowledge of the basic units, e.g., phonemes, can readily use a predictability based strategy to start extracting units from continuous speech.

The measures discussed above, *successor variety* and *transitional probabilities*, are not the only measures of (un)predictability. There are a number of other measures for quantifying the predictability, such as *entropy* and *mutual information*. We will return to alternative measures in Chapter 6.

[4] To avoid notational clutter, a somewhat sloppy probability notation is followed throughout this thesis. The notation $P(r|l)$ means the probability that the syllable in the next position is r, given the current syllable is l, $P(s_{t+1} = r|s_t = l)$. The symbols l and r will consistently be used for the sequences of the phonemes to the left and right of the position in question, respectively.

The strategy of positing word boundaries where there is an unexpected sequence of syllables or phonemes is one of the strategies that is in line with the evidence of the sort provided by Saffran et al. (1996a). Another strategy that is compatible with the same type of analysis is identifying sequences that frequently occur together in varying environments as possible words. Even though it is not studied extensively in the psycholinguistic literature, a large number of computational models of speech segmentation make use of this strategy (e.g., Brent, 1999a; Goldwater et al., 2009; Venkataraman, 2001, among others). The main driving force behind these models are sequences that co-occur frequently in varying contexts. It is strongly related to predictability based strategies. In their basic form, both strategies depend on the fact that the speaker generates the utterances using units from his or her lexicon. Both the unpredictability of syllable sequences and observing frequent sequences in differing context in the input are artifacts of this process. Most importantly, both strategies are language independent and do not require knowledge of any lexical units in advance.

4.2.2 Prosodic cues

It is known that even a few day old newborns show sensitivity to overall prosodic structure of the language they are exposed to before birth (Mehler et al., 1988; Moon et al., 1993; Ramus, 2002).

The term ‘prosody’ encompasses a relatively large set of acoustic phenomena that may be useful for segmentation, such as pitch contour that marks some sub-clausal boundaries and lengthening of final segments of a lexical unit. However, the only well-studied concrete prosodic segmentation strategy is based on *lexical stress*.

Since most content words in English follow a strong–weak (trochaic) stress pattern, it seems that adults (Cutler and Butterfield, 1992) and 7.5 month old children (Jusczyk, 1999; Jusczyk et al., 1999b) use a segmentation strategy that proposes word onsets before the strong syllables. The lexical stress is useful for segmentation of English as most content words in English are stressed on their initial syllable.[5] However, this strategy is not equally effective for all languages. The languages vary depending on where primary stress falls within words as well as consistency of location of the primary stress. The experimental evidence for use of stress in different languages are also mixed. On one hand, a similar trochaic pattern preference is found for Dutch (Vroomen et al., 1996), and Canadian-French speakers seem to make use of the weak-strong (iambic) stress pattern for segmentation (Polka and Sundara, 2003). On the other hand, infants learning (European) French and Spanish do not seem to make use of stress as a segmentation cue (Nazzi et al., 2006).

A more general version of this strategy, the so-called *metrical segmentation strategy* (MSS), is advocated particularly by Cutler and her colleagues as both a processing

[5] Cutler and Carter (1987) reports that 71% of all words, 90% of the content words in MRC Psycholinguistics database (Coltheart, 1981; Wilson, 1988) have strong initial syllables. One should note, however, that most frequent words, e.g., function words, are unstressed, and numbers vary depending on how one counts them (see Swingley, 2005, for a more careful analysis).

strategy for adults, and as a 'bootstrapping' method for children (Cutler, 1996; Cutler and Butterfield, 1990; Cutler and Carter, 1987; Cutler and Mehler, 1993). The MSS relies on the so-called rhythmic classes of languages (Pike, 1945). Despite contrary empirical evidence (Dauer, 1983; Roach, 1982), the intuition behind the rhythmic classes is that certain units are produced at approximately equal intervals. In *stress-timed* languages (e.g., English and Dutch) this unit appears to be the stressed syllables. In *syllable-timed* languages (e.g., French and Spanish) the syllable is assumed to be the rhythmic unit.[6] Then, MSS suggests that once the infants tune into their languages' rhythmic class, they use it for lexical segmentation. The lexical stress, and how it can help lexical segmentation is well founded. However, the status of an MSS-like strategy for syllable-timed languages is rather uncertain. To my knowledge, the studies advocating syllable-based segmentation do not go further than establishing the syllable's role as a perceptual unit which shows cross-linguistic differences (Cutler et al., 1986). Additionally, even though lexical stress has been used in explicit models of segmentation (e.g., Christiansen et al., 1998), there is no explicitly stated method that uses a general MSS like strategy based on syllables (see also the discussion on units on computational models of segmentation in Section 5.1).

The claim that the MSS can serve as a bootstrapping method to start extracting lexical units without aid of a lexicon (Cutler, 1996; Cutler and Mehler, 1993) relies on the fact that children are sensitive to prosody very early in life. However, it is again unclear how this can be done. Even if it is a very reliable indication to the word boundaries, the lexical stress patterns differ among languages. For example in two languages that are observed to have highly regular stress patterns, Finnish words are stressed on their first syllable while Polish words are stressed on the penultimate syllable. For a learner to figure out the stress pattern of the ambient language, they first need to learn a relatively large number of lexical items. There are suggestions that MSS can be bootstrapped from the lexical units learned either by using statistical predictability (Swingley, 2005; Thiessen and Saffran, 2004) or possibly by first learning the isolated words and short utterances (Johnson and Seidl, 2009).[7] In any case, the stress-based segmentation strategy as sole bootstrapping method is not a viable option.

The lexical stress seems to be the only viable rhythmic cue that is present in the input, and it is used at least by learners of some languages. The other rhythmic segmentation strategies may be useful for segmenting speech into basic perceptual units. However, these methods need explicit suggestions of how they can be used for extracting lexical units from continuous speech.

[6] Another category found in the literature is based on *mora* (a sub-syllabic unit, see Mazuka, 2007, for a recent discussion relevant to segmentation). The only language identified so far that is claimed to be mora-timed seems to be Japanese.

[7] Swingley (2005) presents an analysis of very short utterances in the Korman corpus (Korman, 1984) which largely undermines the second hypothesis. According to this particular analysis, the stress pattern of short utterances is dominated by strong-strong pattern by a large margin, and for some counts the reverse (weak-strong) pattern is also more frequent than the expected strong-weak pattern.

The majority of studies that investigate role of prosody on lexical segmentation have focused on rhythmic structure of the language. Another aspect of prosody that received relatively little attention is prosodic marking of sub-clausal units, such as *intonational phrase* and *phonological phrase*.[8] Since phrase boundaries are also word boundaries, they can be useful the same way pauses are useful for segmentation. Adults are found to be sensitive to intonational boundaries and they use this information in lexical segmentation (Shukla et al., 2007). It was also found that 6-month-old infants show sensitivity to prosodic phrase boundaries (Soderstrom et al., 2003). The way the phrase boundaries can aid lexical segmentation is clear: like utterance boundaries, they can constrain the hypothesized lexical units, and also give further hints for beginnings and ends of lexical units.

In summary, certain prosodic cues seem to play a role in segmentation. Concrete proposals exist for the effects of *lexical stress* and the sub-sentential units marked with prosody. However, proposals suggesting contribution of other prosodic cues (for example, syllable-based rhythmic structure of some languages) are rather unclear.

4.2.3 Phonotactics

In the segmentation literature *phonotactic cues* refer to the way sounds in a natural language are organized to form lexical units. For a given language, some sound patterns do not occur, some tend to occur more at the beginning of the words, some word internally and some at the end of words. For example, the sound sequence /θm/ does not occur in English words. If one observes this sequence in an unsegmented utterance such as *wi**thm**ilk*, it can serve as a cue that there is a word boundary between these two sounds. Commonly, these regularities are assumed to form *constraints*. That is, a certain sound sequence is either possible in the context of interest, or not. However, one can also expect more soft tendencies, where some sequences are more likely than others.

Combined with the findings that adults (Greenberg and Jenkins, 1964) and infants as young as 6-month-olds (Jusczyk et al., 1993) are sensitive to sound regularities that form words in their language, it is reasonable to assume that phonotactics is a possible source of information that can aid segmentation.

Jusczyk et al. (1993) showed that 9-month-old English learning infants distinguished between sound sequences of English and Dutch, two relatively similar languages. However, 6-month-olds did not show the same distinction, but distinguished more distinct sound patterns, e.g., sound patterns of Norwegian vs. English. Furthermore, Jusczyk et al. (1994) showed that 9-month-olds learning English were also sensitive to the frequency with which certain sound sequences occur in English. Thus,

[8] *Intonational phrase* refers to the segment of speech that occurs with single prosodic (i.e., pitch or rhythmic) contour. *Prosodic phrase* is typically a content word and its associated function words (Nespor and Vogel, 1986).

evidence suggests that the infants develop a sense of the sound sequences in their language early on, and that they do it in a graded manner.

Although it is common to view phonotactics as a set of hard constraints, phonotactics is essentially based on the statistics of sound sequences. Hence, it is closely related to the sensitivity that infants show to the distribution of sound sequences that signal predictability. However, there are two major differences. First, on one hand, the predictability cue does not require a repository of already identified words, but in order to learn phonotactic regularities one needs to identify some words first. On the other hand, once a set of words are available phonotactics can make generalizations based on positions in the word, while predictability cue does not. For example, the phonotactics of English would capture the fact that consonant cluster /kt/ would not occur word initially. On the other hand, this sequence is not likely to be useful for predictability-based statistics, since it occurs in other contexts relatively frequently, for example, in *talked* as in many other past tense forms of verbs ending with /k/. Second, phonotactics is a language specific cue. Even though /kt/, does not occur word-initially in English, it occurs in other languages such as Polish. There is some overlap in the way predictability and phonotactics may provide cues for segmentation. However, there are differences in what they predict, and in which conditions they function. As a result, contributions of phonotactics and predictability together are likely to be more effective than one of them alone.

4.2.4 Pauses and utterance boundaries

Intuitively, pauses are the most robust cues to word boundaries. However, fluent speech does not have consistent pauses between words. Even though frequency of occurrence is unclear, it is also known that pauses may occur within the words (Slis, 1970). Pauses alone are not sufficient to determine all word boundaries, nor are they completely reliable. Nevertheless, when pauses occur (for example at the utterance boundaries) they can be utilized for lexical segmentation (Christiansen and Allen, 1997). Further evidence suggests that they often occur between phrase boundaries (Wightman et al., 1992), which appears to be a characteristic of long child-directed utterances (Fernald et al., 1989).

Since the utterance or phrase boundaries are also word boundaries, pauses can be useful for segmentation in two ways: First, by restricting possible segmentations, i.e., eliminating candidate words that span over pauses. Second, by serving as indications of sound patterns that occur at the beginning and end of the words, hence, complementing phonotactics.

4.2.5 Words that occur in isolation and short utterances

A possible step towards building a lexicon is attending to the words that occur in isolation. The naive strategy of extracting words that are uttered in isolation is generally dismissed on the logical grounds that there is no reliable way to distinguish single-word utterances from multi-word utterances (Christophe et al., 1994). Further, it was found that words in isolation are not used consistently when mothers are asked to teach a new word to their children. Even though caregivers used other strategies (such as placing the new word at the utterance final position) to help the learners, they rarely used it in isolation (Aslin et al., 1996). Despite these criticisms, Brent and Siskind (2001) demonstrated that exposure to isolated words facilitates lexical development. Short utterances can be useful for segmentation in a number of ways. Because of the increased number of utterance boundaries, they contribute to form a better knowledge of phonotactics of the language. The number of alternative segmentations is also smaller for short utterances, which increases the chances of successful segmentation of utterances that contain unknown words.

4.2.6 Other cues for segmentation

Allophonic cues refer to the fact that some phonemes are realized differently (by different allophones) depending on lexical context. For example, in English stop consonants tend to be aspirated when they are word initial (Church, 1987). The phoneme /t/ in *toy* is likely to be uttered in an aspirated fashion, unlike the /t/ in *bottle*, or *cat*. Like the phonotactic cues, this acoustic variation depends on lexical context, and can be useful for segmentation. Indeed, infants around 10.5-months of age are able to use allophonic differences to extract words from speech (Jusczyk et al., 1999a). Jusczyk et al. (1999a) habituated infants to bi-syllable sequences like *nitrates* and *night rates*. During the testing phase, 10.5-month-olds showed a preference towards listening to the passages that contain the habituated words, but 9-month-olds did not show any preference.

Vowel harmony is another cue found to be used by speakers of languages where words follow some form of vowel harmony. Vowel harmony places restrictions on the vowel classes that can be found in a word. In such a language, a class mismatch between two consecutive syllables provides a cue for a word boundary between these syllables. Finnish adults (Suomi et al., 1997) and Turkish infants (van Kampen et al., 2008) have been shown to be sensitive to the vowel harmony of their language, and use it as an additional cue for segmentation.

Coarticulation refers to the phenomenon that consecutive phonemes overlap during articulation, and they are realized acoustically differently depending on their phonetic context. Since this is a source of variation that hearers need to cope with in order to identify the phonemes, it is normally a phenomenon that complicates the task of the speech recognition. However, coarticulation turns out to be useful for segmentation as the overlap between the consecutive phoneme pairs at the word boundaries is less than the overlap at the word internal consecutive phonemes (Fougeron and Keating, 1997). It is also observed that infants as young as 8-month-olds are sensitive to coarticulation (Johnson and Jusczyk, 2001) and it overweights other cues particularly in noisy listening conditions (Mattys, 2004).

The list of cues to speech segmentation reviewed here is by no means complete. Other cues varying from syntactic knowledge (Mattys et al., 2007), to visual environment (Hollich et al., 2005) seem to be playing roles in segmentation of speech signal by humans. The review in this section provides a background for the type of cues that are commonly used in computational modeling, including the model of segmentation presented later in this chapter.

4.3 Segmentation cues: combination and comparison

The experimental studies so far have established firmly that adults and children are sensitive to a number of cues, so that it is clear that combining information from multiples sources is advantageous.

- Using information from multiple sources provides redundancy, increasing the reliability in case some of the sources are noisy.
- By integrating information from multiple sources, it may be possible to derive conclusions that are not possible to derive from a single source. For example, using two eyes one can get reliable depth information through stereo vision.
- Using multiple sources of information also facilitates learning (Christiansen et al., 2005).

Cue combination and integration has been studied more in other areas of cognitive science such as depth perception and sensory-motor control (e.g., Ghahramani et al., 1997; Kording et al., 2007; Landy et al., 1995). However, most experimental studies in the segmentation literature are concerned with establishing an early or more prominent place for the cue of choice—particularly between *statistics* and *prosody*. Studies investigating the combination of these cues and mechanisms underlying the combination is relatively few (Mattys et al., 2005; Sanders and Neville, 2000; Shukla et al., 2007).

In the debate between prosody and statistics, the interest is whether one of the cues dominates the other in case of conflict (e.g., Johnson and Seidl, 2009; Thiessen and Saffran, 2003), or whether initial bootstrapping is provided by a certain cue (e.g., Thiessen and Saffran, 2007). Some of these studies seem to be fueled by the nature–nurture debate: Prosody is taken as the representative of innate, domain-specific knowledge, and statistics as an example of general purpose learning procedures (e.g., Gervain and Mehler, 2010). As with other instances of this debate, many conclusions on the subject are rather stretched. Indeed, various sorts of statistical learning seem to be domain general, e.g., statistics similar to the ones discussed in Section 4.2 used by humans in segmenting tone sequences (Saffran et al., 1999) and visual patterns (Fiser and Aslin, 2002; Kirkham et al., 2002). Statistical learning is also found to be not specific to humans. Other mammals are also sensitive to the statistical regularities in sound sequences (Hauser et al., 2001; Pons, 2006). Infants show sensitivity to the prosodic structure of the ambient language very early in life. However, the use of prosodic cues in segmentation requires learning. For example, the most studied cue, lexical stress, needs to be learned since it is language specific. Furthermore, it is likely that some statistics is involved in learning these patterns, since they are not inviolable rules, but (statistical) tendencies. For example, with the most favorable count, the dominant trochaic stress pattern of English is valid for at most 90% of the words.

Whether innate or learned, an important finding is that infants use a number of cues in combination for segmentation. It is yet to be established how these cues integrate and interact.

5 Computational Models of Segmentation

> Science is knowledge which we understand so well that we can teach it to a computer; and if we don't fully understand something, it is an art to deal with it.
>
> Donald E. Knuth

Segmenting a continuous stream into linguistically useful units is a computational problem. Figure 4.1 on page 45 presented an example that approximates the linguistic input we segment in everyday life. Faced with an input stream like the one given in Figure 4.1, at first sight it is surprising that children can extract anything from it at all. The studies reviewed in Chapter 4 suggests that there are certain features of the linguistic input, or cues, that children are sensitive to. Furthermore, some of these cues have been shown to be used by adults and children in the segmentation task.

Computational modeling and simulations provide a relatively easy way to investigate some of the questions regarding segmentation problems, particularly the availability of these cues in the input, and whether and how these cues may contribute to the solution of the segmentation problem. Not surprisingly, learning segmentation has been one of the most studied aspects of language acquisition from the computational perspective. Initial ideas go back to Harris (1955), and there have been explicit computational implementations as early as Olivier (1968). The fact that three out of six computational models presented in a recent special issue of the Journal of Child Language on computational models of language acquisition (MacWhinney, 2010) are models of learning segmentation (Blanchard et al., 2010; Monaghan and Christiansen, 2010; Rytting et al., 2010) is an indication that the solutions to the segmentation problem are far from being settled yet.

A review of the psychologically motivated computational models of segmentation up to 1999 can be found in Brent (1999b). Besides psychologically motivated models, there are also a number of segmentation tasks in natural language processing applications that we can also learn from. One of these tasks is segmenting written text for the languages that are written without white spaces or any other separators between the words, e.g., Chinese and Japanese. Another application arises in the problem of segmenting the words into morphemes for various levels of morphological analysis.

This chapter introduces the computational problem of learning segmentation with a review of the state of the art in computational modeling of segmentation, focusing more on psychologically motivated models. The criteria listed below for comparing computational models of segmentation will be considered during discussions of individual models of segmentation. These criteria below are similar to and partially overlap with the criteria suggested earlier in the literature (Batchelder, 2002; Brent, 1999b; Monaghan and Christiansen, 2010).

- *The input*: The way input is presented to a model is an important part of its specification. Particularly interesting aspects of the input include the basic units of the representation, whether it is naturally occurring speech or artificially generated input, and the aspects of the input that the model is sensitive to.
- *Processing strategy.* Two broad strategies of segmentation are found in the literature. Models that use the *prediction* strategy try to predict the boundaries directly, without extracting the lexical units explicitly. Models that use the *recognition* strategy try to extract frequently occurring substrings as lexical units from the input.
- *Processing characteristics.* In particular, whether the model needs a large number of utterances at once (*batch*), or if it, alternatively, processes each utterance in turn (*incrementally*). Brent (1999b) also distinguishes between incremental models and *on-line* models. He defines on-line models as models that do not wait until the utterance boundary to start extracting words. Further, he suggests that a *predictive-online* model is a better candidate for modeling human performance, as humans tend to guess the lexical units even before they completely hear them.
- *Search strategy*. A brute-force search through all possible segmentations is intractable and implausible (see Section 5.3 for a discussion) for realistic input. The search strategies vary in many ways. One aspect of interest is whether they start from the building blocks (e.g., phonemes) and combine them to lexical units (*synthetic*), or take the whole utterance and segment it into lexical units (*analytic*). Another aspect we are interested in is the way the search space is explored.
- *Performance.* We do not exactly know what the lexical units in our mental lexicons are and what the end product of the human segmentation process is. However, everything being equal, we would like our models to perform similarly to a hypothesized gold standard.
- *Computational complexity.* Since the human cognitive system has limited computational resources (such as memory and processing power), a computational model of human performance is expected to use plausibly limited resources.
- *External constraints*. Most models utilize some hard constraints that help the segmentation algorithm some way. A common example of such constraints is requiring all lexical items to have a vowel. Even though some of these constraints are indeed useful for segmenting the input, assuming the constraints still begs the

explanation of how these constraints are learned. Assuming these constraints are innate is a possible path, but, for such hard constraints to be innately specified they would have to be valid for all natural languages. Similarly some free parameters (the parameters that set by the model designer and not learned by the model) of the models also form constraints that require explanation.

- *Whether the model builds a lexicon or not*. Even though all segmentation models can be argued to have some sense of lexical units, the models that build an explicit lexicon are more attractive if we assume that the learner's interest is to assign meanings to the discovered lexical items.

5.1 The input

All the segmentation models we describe in this section are *unsupervised* models in the sense that they are not given the correct boundary locations in the input. The representation of the input varies among the models, but the aspect of the input common to all of these models is that their input is a set of unsegmented strings with possible noise. The models are not given negative examples labeled as such, nor are they provided by corrections if they make mistakes.[1]

For segmentation models that try to answer questions about language acquisition, it is natural to use the input that children receive during the acquisition of language. This makes child-directed speech and representations that are close to the real speech signal more attractive. However, some models developed on written text, or artificially generated text, may also demonstrate certain concepts better, or offer insights that can be transferred into more relevant models. Except for a few (e.g., Elman, 1990; Goldsmith, 2001; Perruchet and Vinter, 1998), most models discussed in this section use some encoding of child directed speech (CDS) for simulations.

The main body of work in infants segmentation in psycholinguistics considers the syllable as the base unit (e.g., Saffran et al., 1996a). However, a few exceptions aside (Perruchet and Vinter, 1998; Swingley, 2005), most computational models in the literature are developed and tested using phonemes as the basic unit of the input stream. At first, this may seem to be conflicting with the findings that the syllable is a salient unit of perception in early infancy (Bijeljac-Babic et al., 1993; Jusczyk et al., 1995). However, this does not necessarily pose a problem for the computational models that use phonemes. The syllable's role as a salient perceptual unit does not rule out the phoneme as another unit that infants are sensitive to. Even though the experimental evidence supporting the phoneme's role as (another) perceptual unit is not as abundant, there are studies which indicate that infants are sensitive to sub-syllabic differences as well. For example, Jusczyk and Thompson (1978) demonstrated that 2-month-old infants are sensitive to differences in place of articulation, and capable of distinguishing

[1] Note that the second assumption is not always true. Children may be correcting an initially wrong segmentation when they build the overall interpretation of the utterance, which is likely to be aided by their interaction with the environment.

sound sequences ***b**ada–**g**ada*, as well as the differences in non-initial phonemes like *da**b**a–da**g**a*.

Another point in defense of using phonemes (or any other basic unit) is the fact that the choice of phonemes does not necessarily change the nature of the computation. For most algorithms the principles are equally applicable to any basic unit of choice.

One last point to elaborate here is that assuming syllables as indivisible basic units has its problems in lexical segmentation as well. First, unless the input is syllabified in an unnatural way, word boundaries in natural speech do not always correspond to syllable boundaries. For example, the natural syllabification of the utterance *what's a kitty?* will likely have the first two syllables /wʌt/ and /zə/. Depending only on syllable sized units, a speaker who hears /wʌt/ and /zə/, and represent these words in the lexicon as /wʌtz/ (*what's*) and /ə/ (*a*) will fail to recognize it. Similarly, without a method to segment the stream into sub-syllabic units, the productive morphemes that do not have a vowel cannot be extracted as possible lexical items.

The input representations vary according to how they are represented. Phonemes are most commonly represented as individual symbols. However there are also a few (connectionist) models that represent each phoneme as a set of phonetic features (Aslin et al., 1996; Cairns et al., 1994; Christiansen et al., 1998). The phonetic feature-based representation becomes particularly useful for the models that guess word phonotactics from utterance boundaries. The feature representation allows models to exploit the information regarding similarities of individual phonemes. For example, a feature-based representation that includes information about vowel and consonant features for each phoneme can learn that consonant-vowel-consonant (CVC) sequences are more likely than CCC sequences. A symbolic representation that assigns arbitrary symbols to each phoneme would require more data to arrive at similar generalizations.

Another issue regarding the input is the representation of variability in the speech signal. Most work in computational modeling use transcribed child-directed speech which virtually contains no variability. In most cases, all words are transcribed using canonical phonetic or phonological forms. A number of models attempted to introduce variability. However, all such attempts create the variability in a artificial way. Cairns et al. (1994) introduced variability by flipping an unspecified number of random bits in their phonetic feature vector representation. In a more recent study, Monaghan and Christiansen (2010) processed orthographic transcriptions of child-directed speech with a speech synthesizer, which produced phonemic transcriptions of words that vary depending on the context. In another recent model, Rytting et al. (2010) run an automatic speech recognizer (ASR) on raw audio input. The ASR produced a vector of probabilities for each phoneme. If speech input is clear, the output would indicate a single phoneme with very high probability, and the rest of the phonemes would be assigned near-zero probabilities. However, if ASR is not able to decide the exact identity of the phoneme, it is likely that more than one phoneme would be assigned non-zero, possibly approximately equal, probabilities.

5.2 Processing strategy

As mentioned before, the computational models of learning segmentation can be divided into two broad groups according to the strategy they use for segmentation. Some of the models search for boundaries, where lexical units are identified as a side effect of discovering boundaries. I will call this strategy *prediction* (or *boundary-guessing*) strategy. Brent (1999b) distinguishes two sub-groups, one using predictability as the source of prediction, and the other using utterance boundaries. These two sources of information are frequently combined in models that guess boundaries. Regardless of the source of information they use, all models that guess boundaries will be classified under this group. The second group tries to recognize lexical units. This strategy will be called *recognition* strategy. A large subset of the models that belong to this second group will be called *language-modeling* strategy for the reasons that will be explained shortly.

5.2.1 Guessing boundaries

In its simplest form, e.g., as used by Harris (1955), this strategy is strongly related to strategies suggested by experimental studies like Saffran et al. (1996a). In natural language processing, similar methods have been used for morphological segmentation, i.e., segmenting words into their morphemes, (e.g., Al-Shalabi et al., 2005; Bordag, 2005; Hafer and Weiss, 1974; Stein and Potthast, 2008) and segmentation of words from texts that do not include white spaces (e.g., Ando and Lee, 2003). However, the use of this strategy in computational models of human segmentation is rather scarce. Except a number of connectionist systems that implicitly implement it, the only use of predictability based strategy that I am aware of is by Brent (1999a), where a simple *mutual information* based segmentation model has been presented as a baseline. If we include the models that guess boundaries using other means (Fleck, 2008; Monaghan and Christiansen, 2010), the count goes slightly up. However, the use of prediction strategy in explicit statistical models is relatively unexplored.

Starting from Elman's 1990 seminal work that introduced *simple recurrent networks* (SRN), the major representatives of the prediction strategy have been connectionist systems (Aslin et al., 1996; Cairns et al., 1994; Christiansen et al., 1998; Elman, 1990). Elman (1990) used a simple segmentation task as a demonstration of capabilities of SRNs. SRNs are standard feed-forward artificial neural networks except for a simple augmentation. The additional so-called context units keep a copy of the hidden layer in the previous step, which is fed back to the hidden layer as well as the current input. This allows SRNs to generalize from past as well as from current input. Typically, the task of an SRN is to predict the next input in the sequence. Elman used a 5-bit representation for the input. Each letter of the English alphabet was mapped to an arbitrary 5-bit binary string. The SRN was trained and tested on artificially generated English-like sentences. After training, the error rate (root mean square error calculated on output units) was lower when the SRN had to guess the next letter within the same word, but higher when it had to guess the first letter of the next word. Cairns et al. (1994) trained another type of recurrent network using a similar prediction task, but using CDS as input. In the Cairns et al. (1994) study input was represented as phonetic features of government phonology (Harris and Lindsey, 1995).

An advantage of boundary-guessing strategy is that it can include more sources of information in a natural way. This was demonstrated by another SRN model by Christiansen et al. (1998). In this model the network was trained using a prediction-based strategy. However, the input was marked for lexical stress, and an explicit utterance boundary unit was included in the input and the output. The task for the system was to predict the input, including the existence of an utterance boundary, which was turned on at the input layer only for the last phoneme of the utterance. The study showed that the network's prediction of an utterance boundary was higher on word boundaries compared to non-boundary locations.

Aslin et al. (1996) present a non-SRN connectionist model. In this study a standard feed-forward neural network which took a three-phoneme input window (each phoneme was coded as 18 binary phonetic features) at each time step. An additional input unit indicated the existence of an utterance boundary after the given three phonemes. There was only a single output unit that indicated an utterance boundary. The network was trained to find utterance boundaries. During testing, the output unit showed higher activation levels at word boundaries than at word-internal phonemes.

A recent model that uses explicit statistics for guessing boundaries was presented by Fleck (2008). This model learns the patterns that occur at the beginnings and ends of utterances to guess possible beginnings and ends of words, hence the word boundaries. In a sense, the model learns a form of phonotactics to guess word boundaries. However, the simulations presented in this study used large prefix and suffix lengths, probably allowing model to learn complete words and phrases. I will discuss this study further in Chapter 7.

The last study that is worth noting here is the study by Monaghan and Christiansen (2010). Unlike the models reviewed in this section so far, the model presented in this study uses previously learned words as the main source of information. However, unlike the models reviewed in the next section, it searches for boundaries in an incremental fashion. Further discussion of this model will be provided in Chapter 8.

In principle, the models using the boundary guessing strategy allow on-line processing of the input. That is, these models do not have to wait until the utterance boundary to posit a word boundary. In practice, this is somewhat more complicated. The actual threshold for deciding on a boundary tends to be decided globally (such as over the mean activation level of the connectionist models described here). These models typically require a relatively large amount of training data before they can achieve the levels of success reported in these studies. A truly unsupervised strategy is segmenting at the locations where there is a peak in unpredictability: an increase in unpredictability followed by a decrease. In principle, one may use peaks as boundary criterion in order to remedy the need for the threshold values. Use of peak-based boundary decision will be investigated in detail in Chapter 6.

It is also interesting that the only cue combination approach tried in the literature uses connectionist models. Although the recognition-based models discussed below can be argued to incorporate information from different sources, the possibilities of integration of new cues are limited barring fundamental changes to the model architecture. On the other hand, the models of the sort reviewed in this section provide a natural way of combining multiple cues.

An apparent weakness of most of the models discussed in this section is that they do not build an explicit lexicon, which means that they cannot use the information from the already extracted lexical units. In augmenting the boundary-guessing models with a lexicon, the models with explicit representations, in comparison to the connectionist models, seem to be at an advantage. Even though modifications to connectionist models to include information from a lexicon or other higher level information sources are possible, in practice these models seem to work best with lower level perceptual cues.

5.2.2 Recognition strategy

Except for a few models discussed above, the majority of the computational models of segmentation in the literature use a recognition strategy. These models try to identify re-occurring strings in the input as candidate words. Even though theoretical motivations are different, common to all these approaches is to define an objective function that indicates what a 'good word' is. Then, the best segmentation is defined as one that favors the use of words with higher likelihood.

A common approach is defining a representation scheme in line with the *minimum description length* principle (MDL, Rissanen, 1978). Then, the aim is to find the best representation that minimizes the 'code length' of both the lexicon and the input corpus. The method has been applied to segmentation of transcribed child directed speech (Brent and Cartwright, 1996), as well as many examples in natural language processing (e.g., Goldsmith, 2001, 2006, in morphological segmentation and analysis). The MDL based models are appealing as they are considered to be a formulation of well known *Occam's razor* in that they prefer simpler solutions. They also have strong ties to data compression, since finding the minimum representation scheme will compress the data. In practice, MDL based approaches work best with batch systems, and in many ways they are similar to the probabilistic approaches I will describe in detail below.

Another related approach in recent models of segmentation is to define a probabilistic generative model that is hypothesized to have generated the data. Then, finding the most probable segmentation under this model tends to give rise to a good segmentation of the linguistic data. In principle, the model can be arbitrarily detailed, capturing all relevant aspects of the system being modeled. Once the generative model is defined, one can assign the probabilities to the segmentations of a given sequence.

As in Bayesian models by Brent (1999a) and Goldwater et al. (2009), some models define the generative model explicitly and find the (approximate) highest probability segmentation of the complete corpus. In doing so, one simply assigns probabilities to segmentations (sequences of words) under the generative probabilistic model. A number of models, on the other hand, assign probabilities to segmentations in a similar way, but without defining an explicit generative probabilistic model (e.g., Batchelder, 2002; Venkataraman, 2001). Below, I will go through a formulation of a simple segmentation model that assigns probabilities to segmentations of a given string. The models in the literature vary in the way they model different aspects of the task. However, the use of so called *language models* (Jurafsky and Martin, 2008, chapter 4) for sequences of words and phonemes forms the skeleton of many successful segmentation models that follow a recognition strategy.

The model defined by the Equation 5.1 and Equation 5.2 below demonstrates this modeling practice. Assuming a segmentation s is composed of the lexical units $w_1 \ldots w_n$, the probability of the segmentation is calculated as,

$$P(s) = \prod_{i=1}^{n} P(w_i) \tag{5.1}$$

The probabilities of individual lexical units are calculated using

$$P(w) = \begin{cases} (1-\alpha)f(w) & \text{if } w \text{ is known} \\ \alpha \prod_{i=1}^{m} P(a_i) & \text{if } w \text{ is unknown} \end{cases} \tag{5.2}$$

where $f(w)$ is the empirical probability, or relative frequency of w, the sequence $a_1 \ldots a_m$ is the sequence of phonemes that form an unknown lexical unit and $P(a_i)$ is typically estimated using relative frequency of the phoneme a_i.

This formulation suggests that if the proposed lexical unit is already known, the probability of each possible lexical unit $P(w_i)$ is estimated using maximum likelihood estimation. More elaborate models tend to estimate this probability using higher level n-grams, taking the effects of sentential context into consideration. Higher level n-grams change the way the individual word probabilities are calculated, but they are still conceptually the same as the basic model defined above. In case the proposed lexical unit is unknown, the model falls back to the phonemes that form the lexical unit. Except for elaborate methods of phonotactics which typically employ phoneme n-grams, the formulation here is identical to a large number of models in the literature. The coefficient α in Equation 5.2 determines the probability of seeing a new lexical unit. Most models choose a way to decrease α as the model commits to more lexical units, slowing down the model's preference towards new units as more lexical units are seen.

The main body of this formulation gives rise to two tendencies, simply because multiplying more probabilities (real numbers in range $[0, 1]$) will result in smaller numbers. First, Equation 5.1 sets a preference towards segmentations composed of smaller number of lexical units. If left uncontested, this probability assignment would prefer whole utterances as lexical items, causing extreme undersegmentation. On the other hand, Equation 5.2 prefers shorter lexical items. Everything else being equal, this formula prefers segmentations where each word is a single phoneme, causing extreme oversegmentation. These two preferences, in effect, work similarly to MDL-based approaches. The undersegmentation tendency imposed by Equation 5.1 has the same effect as the preference towards the shortest encoding of the corpus, while the oversegmentation tendency imposed by Equation 5.2 has the same effect as the preference towards shortest encoding of the lexicon.

In general, this type of modeling makes use of two n-gram *language model*s. One of the language models is used to model the sequences of words that form an utterance, and the other is used to model the phonemes that form a word. For this reason, I will refer to this type of segmentation strategy as *language-modeling strategy* (LM). Despite the fact that language models have been useful in many natural language applications, there are a number of shortcomings of these models. In this work, we are particularly concerned with the problems of the n-gram models that one cannot incorporate arbitrary features of the input, and the search strategy typically used in these models is implausible for modeling human performance. A reference implementation of a LM-based segmentation model will be presented in Section 5.5.

5.3 The search space

Given a sequence (e.g., of phonemes), finding the best segmentation can be characterized as finding the best segmentation among all possible segmentations. The brute-force solution to this problem is to enumerate all possible segmentations of the string, compare them according to a criterion, and pick the best scoring segmentation. However, this brute-force approach faces a serious problem: the number of possible segmentations for a given sequence composed of n units is 2^{n-1}.

Searching through all 2^{n-1} possible segmentations is not feasible except for short strings. To give an idea of what a brute-force hypothesis generation method is dealing with, Table 5.1 lists number of possible segmentations for utterance lengths of 1, 10, 20, 30, 40, 50, and 100,000 (approximate number of phonemes in the reference corpus used in this study). A brute-force approach needs to consider over 5.6×10^{14} segmentations for a 50-phoneme long utterance. Assuming that we can process (e.g., calculate a score and compare to the best so far) a million segmentations per second, processing a 50-phoneme utterance would take over 17 years. Furthermore, adding a single phoneme doubles this time. Since batch segmentation algorithms search through all possible segmentations of a complete corpus instead of a single utterance at a time, they have a more difficult problem. Clearly, a brute-force comparison of all possible segmentations of a given utterance, or corpus, is both computationally infeasible and cognitively unappealing.

Since the brute-force search through all segmentations is computationally intrac-

length (n)	substrings ($\frac{n(n+1)}{2}$)	segmentations (2^{n-1})
1	1	1
10	55	512
20	210	524,288
30	465	536,870,912
40	820	549,755,813,888
50	1275	562,949,953,421,312
100,000	5,000,050,000	$\approx 5.0 \times 10^{30102}$

Table 5.1: Number of substrings and possible segmentations of an utterance of length n.

table, either algorithms that reduce the time complexity with some additional memory usage, or approximate search algorithms that do not guarantee the globally optimal solution are commonly used in the field.

5.3.1 Exact search using dynamic programming

Even though all possible segmentations are exponential in the size of the string, all possible continuous substrings found in a string of size n amount to $\frac{n(n+1)}{2}$: For a string of size n, there is only one possible substring with length n, the string itself. For example, for the string *akitty*, only possible substring of length n is *akitty*. There are two substrings of size $n-1$, {*akitt*, *kitty*}, three substrings of size $n-2$, {*akit*, *kitt*, *itty*}, four possible substrings of size $n-3$, {*aki*, *kit*, *itt*, *tty*}, and so on. In general, for a string of size n there are $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ possible substrings. This number grows much more slowly than the number of segmentations, and gives some indication that dynamic programming algorithms (algorithms that avoid re-calculating the same values at the expense of some additional memory use) can be useful. This search strategy has been used in a number of incremental models in the literature (e.g., Brent, 1999a; Venkataraman, 2001). We will briefly describe the first one.

Brent's algorithm finds the best segmentation of the utterance length n in time proportional to n^2, by using a memory proportional to $2n$. The algorithm assigns a probability to all initial substrings of the utterance given past input and the model assumptions. As in Equation 5.1, the probability of a string is the product of the probabilities of the lexical units that form it. The dynamic programming algorithm first calculates these probabilities for each possible substring starting from the beginning of the utterance. The algorithm stores the best probability value and starting point of the last word for best segmentation of each substring. Using the information stored in the previous step, the algorithm finds the best last word, accepts it as part of the segmentation, and moves to the phoneme before the start of this word and repeats the same exercise until the beginning of the utterance is reached. For example, upon seeing the string *akitty*, Brent's algorithm assigns probability values to all initial substrings: `a`, `ak`, `aki`, `akit`, `akitt`, `akitty`. For each substring, the best segmentation is found by considering all possible binary splits. The probability of a substring is the product of the best segmentation probability of the initial part and the lexical probability of the final part.

The algorithm demonstrates the use of dynamic programming to find exact solutions without exponential time complexity. However, this comes with a number of additional limitations. First, the particular algorithm described here is possible because of the assumption that the words in an utterance are independent. It is possible to lift or weaken the independence assumption by introducing additional complexity. However, the attempts so far (e.g., bigram and trigram models by Venkataraman, 2001) did not show substantial improvements. This type of algorithms are also limited to finding the single best segmentation. Similarly, the algorithm can be modified to find n-best segmentation by introducing additional time and memory complexity.

Another complication arises due to the learning setting. The algorithm finds the best segmentation according to the prior knowledge of the learner. Since prior knowledge of the learner is incomplete, details of the search procedure affects the performance of the learner. For example, Brent's search procedure performs best when words at the end of the utterance are known (assuming known words on average get higher probabilities than unknown words). This is probably an arbitrary choice, but incidentally it conflicts with the findings of Aslin (1993) that caregivers tend to place the words that they want to teach their children at the end of the utterances.

The use of dynamic programming to reduce the exponential time complexity of the search problem to polynomial time complexity is attractive for the computational process. However, one expects humans to reduce the search space, rather than comparing all possible segmentations in a more efficient way. For example, comparing all splits of the utterance and its substrings is likely to be a more laborious task than the task humans take during speech segmentation. The hypothesis space is likely to be restricted by the prior knowledge of the learner, and most segmentation hypotheses are likely to be not considered at all.

5.3.2 Approximate search procedures

The batch models benefit from the fact that they have more data at once to decide for a segmentation that is consistent with the complete input. However, the size of the hypothesis space is even larger for the batch algorithms that try to find a single best segmentation of the whole corpus. For incremental segmentation models, dynamic programming allows exact search in a reasonable computational complexity with the expense of additional memory. However, for batch segmentation models, time and memory complexity becomes intractable for the dynamic programming algorithms as well. The last line of Table 5.1 gives an impression of the both time and space complexity of a batch algorithm on a relatively small corpus. Hence, particularly the batch algorithms need to resort to approximate search procedures.

One way to do this is to use heuristics to explore a more relevant part of the complete search space (e.g., Brent and Cartwright, 1996). Generic search methods such as sampling techniques used in machine learning are another way of finding good segmentation solutions without exploring the complete search space (e.g., Goldwater, 2006; Goldwater et al., 2009).

The insights offered by the batch algorithms are particularly suitable for answering *what* questions, giving relatively little insight into *how* questions, i.e., what information is in principle available, but not how a learner exploits it. This limitation is due to the fact that the batch models do not fit the observation that human language processing and acquisition is incremental. Humans do not even wait until the end of the current utterance to guess a word, or word boundary. Assuming that a large amount of corpus is necessary before taking any segmentation decisions is certainly not in accordance with what we know about human segmentation performance. Furthermore, neither the heuristics, nor the general methods of finding approximate solutions based on sampling provide further insights into what sort of hypothesis space humans might be exploring.

Having stated the shortcomings of batch models as psychologically plausible models of human performance, it is interesting to see that they can search through a space of linguistically relevant segmentation hypotheses using search procedures that have no notion of linguistics. The solution is possible because of explicitly stated assumptions about the generative model that produces the utterances and the structure of the input.

Searching through only a restricted part of the search space is common to batch segmentation models. However, it is also used frequently by some of the incremental models to restrict the search space using certain heuristics (or constraints).

The models that build up from primitives using *synthetic* algorithms limit the hypothesis space by considering sequences that have occurred in the input before (e.g. Batchelder, 2002; de Marcken, 1996; Olivier, 1968; Perruchet and Vinter, 1998). At the beginning, these algorithms initialize the lexicon to contain only the primitive units, e.g., phonemes. The algorithms, then, consider only those (sub)strings that can be obtained by binary combination of the current lexical items. This way, the lexicon, and in effect the previous input, constrains the hypothesis space to be explored.

The lexically driven hypothesis space exploration in synthetic models seems plausible, yet, the arbitrary and completely mechanical binary combination of lexical units as possible new lexical units again produces a large number of hypothesis to consider, and its psychological plausibility is questionable. Furthermore, the method conflicts with the evidence that adults and children tend to start with larger blocks (such as complete utterances) as lexical items and break them into smaller units when further evidence supports it (Bannard and Matthews, 2008; Dahan and Brent, 1999; MacWhinney, 1982; Tomasello, 2000).

5.3.3 Search for boundaries

The search strategies discussed in previous subsections consider the models that try to find the best segmentation given an unsegmented utterance. However, some models, particularly ones that follow the prediction strategy (e.g., Christiansen et al., 1998; Elman, 1990; Monaghan and Christiansen, 2010) do not directly search for the best segmentation of the given utterance. In essence, these models find an approximation to the best segmentation of the input utterance. However, they do not view the segmentation problem as finding the best segmentation of an utterance. Instead, they search for boundaries in an on-line fashion, without considering multiple alternative segmentations of the input utterance.

The obvious benefit of these models is their computational simplicity. They try to identify the boundaries without explicitly identifying lexical units. Since most of these models do not build and make use of a lexicon, these models tend to perform worse than the other models (see Table 5.2 for performance comparison). In this thesis, I will present a model that follows a similar search strategy, while performing close to the state of the art models using search strategies described before.

5.4 Performance and evaluation

As with other models of the acquisition of natural languages, we know rather little about our target, the human lexicon. However, everything else being equal, we would prefer the models that perform well against a theoretical *gold standard*.

Even though some of the recent studies in the field made a number of advances that allow easier comparison, comparing performances of segmentation models still faces a number of difficulties.

The first difficulty is the lack of a standardized gold-standard corpus. A large number of different corpora have been used by different studies. The candidate standard corpus seems to be a version of the corpus collected by Bernstein Ratner (1987), and phonologically transcribed and processed by Brent and Cartwright (1996) (see Appendix A for additional information on this corpus). This corpus will be used for the evaluation of the models developed in this study.

The second difficulty arises because of the lack of a common measure of performance, particularly in earlier studies. As in natural language processing and machine learning, reporting a set of performance scores traditionally used in the information retrieval research, *precision* (P) and *recall* (R), is becoming the norm in computational models of cognition as well. These performance scores are based on four counts indicating success or failure of the outcome of the model compared to a gold standard.

- *True positives* (TP) is the number of items of interest correctly identified. For example, if the model is guessing word boundaries, the number of correct boundaries the model found is the TP. This quantity is also called *hits*.
- *False negatives* (FN) is the number of items that the model failed to identify. For

example, the word boundaries that the model could not find. This is also called *misses*.

- *False positives* (FP) is the number of items that the model incorrectly suggested. For the cases of boundaries, the number of times the model suggested a boundary in incorrect locations. It is also called *false alarms*.
- *True negatives* (TN) is the number of cases where the model was correct in not identifying a relevant item, e.g., a boundary. This quantity is not used in calculation of precision and recall, however, it will be relevant for our later discussion.

The precision and recall scores are calculated using the formulas:

$$\begin{aligned} \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \end{aligned}$$

Precision can be seen as a measure of exactness, and it is sometimes called *accuracy* in the cognitive science literature.[2] Recall is a measure of *completeness*, and sometimes called so in cognitive science literature. In informal terms, high precision means that the model has found only correct items, but many relevant items might have been missed. High recall, on the other hand, means that the model has not missed anything, but it may have suggested many irrelevant items. To have a balanced indication, a derived measure, *f_1-score*, is used, which is the harmonic mean of precision and recall.

$$\text{f}_1\text{-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

The subscript ‘1’ indicates that the measure gives equal weights for precision and recall. In its more generic original formulation, f_α-score gives higher weight to recall for higher values of α, and lower values give higher weight to precision (van Rijsbergen, 1979).[3] Since we do not have a reason to prefer precision or recall over the other, all f-scores presented in this thesis are f_1-scores, the subscript is dropped in the remainder of the thesis, and abbreviation ‘F’ is used for denoting f_1-score.

As in recent studies of computational segmentation, in this thesis three different types of precision and recall values are distinguished.

[2] Unfortunately, accuracy is ambiguous in cognitive science literature. The more common definition of accuracy in many branches of science is different than precision. The widely used definition of accuracy for our case is, $\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$.

[3] With regard to f-score, an unfortunate typo is in common use in the computational segmentation literature. A number of papers use the term f_0-score, or F_0, instead of f_1-score. Although it is unlikely to cause any confusion, according to the original definition f_0-score exists, and it is equivalent to precision.

- *Boundary* precision (BP) and boundary recall (BR) calculations consider the boundaries that match the gold standard segmentation as a *true positive*, where the mistakenly proposed boundaries that do not exist in gold standard are considered *false positives* and the boundaries that are in gold standard, but not spotted by the model, are considered *false negatives*. Since utterance boundaries are clearly marked, not to give credit to the segmentation models for stating the obvious, the utterance boundaries are not included in calculation of the boundary scores.[4] The f-score calculated using BP and BR will be denoted BF.
- *Token*, or word, precision (WP) and token recall (WR) scores require both boundaries of a word to be found to count positively in TP. Likewise, the words that are suggested by the model, but not in the gold standard, are FPs. The words that the model could not segment correctly are FNs. The token scores are naturally lower than the boundary scores. Similarly the f-score calculated from WP and WR will be denoted WF.
- *Type*, or lexicon, precision (LP), type recall (LR) and type f-score (LF) are similar to token scores, however, the comparisons are done over the word types the model proposed and word types in the gold standard. These scores are typically less than the token scores. If a model does a good job only at segmenting high-frequency words (e.g., function words), type scores will be much lower than the token scores, but if the model is good at segmenting low frequency words as well, lexical scores will be closer to the token scores. In case the model is particularly bad at segmenting high-frequency words, but good at segmenting low-frequency words, the type scores can be higher than the token scores.

All segmentation models we are interested use unsupervised learning methods in the sense that the algorithms do not have access to information regarding real boundary locations. As a result, it is common practice to present the results on a single data set without training–test data separation. However, in some cases the same practice has been followed mistakenly. Even if the model is unsupervised, it collects information from the corpus during the learning process. Hence, if the algorithm makes use of multiple passes over the data, the information collected in previous passes affects, and likely improves, the scores obtained. The simulations used in this study do not make use of multiple passes over the data, or use any other source of information without explicit notice.

To give a rough idea of how good the models in the literature perform, Table 5.2 presents a few models and associated measures that are found in the studies presenting them. In case there were multiple models reported in a study, the model with the highest lexicon f-score is presented in Table 5.2. Four of the models (Blanchard et al., 2010; Brent, 1999a; Goldwater et al., 2009; Venkataraman, 2001) in the table were chosen since they are easily comparable because of their use of the same corpus. These models also use relatively similar strategies and input representations. Two additional

[4] This is not always clear in the results reported in the literature.

		boundary			word			lexicon		
model	corpus	P	R	F	P	R	F	P	R	F
Christiansen et al. (1998)	Korman	70.2	73.7	71.9	42.7	44.9	43.8	–	–	–
Brent (1999a)	BR	80.3	84.3	82.3	67.0	69.4	68.2	53.6	51.3	52.4
Venkataraman (2001)	BR	81.7	82.5	82.1	68.1	68.6	68.3	54.5	57.0	55.7
Goldwater et al. (2009)	BR	90.3	80.8	85.2	75.2	69.6	72.3	63.5	55.2	59.1
Blanchard et al. (2010)	BR	81.4	82.5	81.9	65.8	66.4	66.1	57.2	55.4	56.3
Rytting et al. (2010)	B&S	54.1	64.8	59.0	21.1	25.3	23.0	13.7	34.3	19.6

Table 5.2: Comparison of performances of some computational models. 'BR' in the corpus column is the Bernstein Ratner (1987) corpus processed by Brent and Cartwright (1996), the Korman corpus (Korman, 1984) is another relatively popular corpus in computational segmentation literature. B&S is the corpus collected by Brent and Siskind (2001). If there were multiple models reported in a study, the model with the highest lexicon f-score is presented.

models (Christiansen et al., 1998; Rytting et al., 2010) that use a different strategy and different input representation are also presented. Christiansen et al. (1998) exemplifies a connectionist (multiple cue combination) model that uses phonetic feature vectors as input. The models presented in Rytting et al. (2010) are based on Christiansen et al. (1998), but they introduce variation in the input by processing acoustic input using an automatic speech recognizer. It should be noted, however, that the performance numbers may not be directly comparable due to the use of a different corpus.

The use of different corpora is not the only reason for difficulty of comparison. For example, results from Christiansen et al. (1998) and Rytting et al. (2010) presented in Table 5.2 are obtained after a certain amount of training. The rest of the studies report the scores calculated without an initial training period. There is another, more hidden, difficulty in comparison of incremental and batch models. The batch models process the complete corpus, typically many times, before they produce any output. Hence, all output from a batch model reflects the final state of the model. However, the incremental models, start with little or no knowledge and learn during the course of segmentation. During the initial phases, the models are expected to make mistakes. As they learn, the performance improves. Hence, even if the final state of an incremental model performs better than a batch model, a simple comparison of the values in Table 5.2 will not indicate this. Some studies present the progression of the incremental models as more input is provided. However, the progression of performance improvement, or the performance at the final phases of the learning process, is not consistently reported in the literature. For this reason it was not included in Table 5.2. Performance reports of the models described in this thesis will include indications of the progression of the model performance throughout the learning process.

Precision, recall and f-score are the standard measures that are well understood and have proven to be useful in the literature. However, it is often more insightful to study where the system fails. For this reasons, I will describe two error measures relevant to segmentation, and report these measures along with the precision, recall and f-score values for the models developed in this study.

A segmentation error can be due to one of two reasons. First, the model may fail to detect a boundary, causing *undersegmentation*. Second, the model may insert a boundary where there is none, causing *oversegmentation*. The simple counts of oversegmentation and undersegmentation errors change depending on the size of the corpus. Hence, they are not comparable across the simulations that run on different corpora. Furthermore, in a typical corpora, there are more word-internal positions than boundaries. As a result, there are more chances to make an oversegmentation error compared to an undersegmentation error. To overcome these difficulties we will use the following error measures for oversegmentation and undersegmentation respectively:

$$E_o = \frac{FP}{FP + TN}$$

$$E_u = \frac{FN}{FN + TP}$$

where TP, FP, TN and FN are true positives, false positives, true negatives, and false negatives respectively. If a single measure is desirable, similar to the definition of f-score, one could also define a combined error measure, e.g., harmonic mean of E_o and E_u. In case where there is no particular reason to prefer reducing a certain type of error, such a measure may be used for model selection. However, since reporting both measures is more informative, and the combined measures can be calculated from the two measures trivially, the combined measure will not be reported in this thesis.

In plain words, E_o is the ratio of the false boundaries inserted by the model divided by the total number of word internal positions in the corpus. Similarly, E_u is the ratio of boundaries missed to the total number of boundaries.

The two error measures described above are related to precision and recall, but the quantities cannot be derived from each other directly. Undersegmentation will reduce true positives which, in turn, reduce both precision and recall. Oversegmentation, on the other hand, will cause false positives to increase, which will affect precision adversely, but will not have an effect on recall. As a result, good recall and bad precision are a typical sign of oversegmentation, and bad precision and bad recall are likely to be due to undersegmentation. So, the error measures are reflected to some extent in precision and recall, but it will be useful to examine them directly as well.

A last note about all the performance scores discussed in this section is that they take values between zero and one. However, as in table Table 5.2, it is common to present values in percentages. This way, the space available can be used more efficiently for significant digits. In this thesis, all values in the tables are percentages, and the values in the graphs are absolute scores (between zero and one).

5.5 Two reference models

Ideally, the performance of a model of the human cognitive capacity should be evaluated based on its match with the human performance. From this perspective we should prefer models that segment as children do—including the incorrect segmentations of children. However, we currently lack the theoretical understanding, the data, and the tools to do this in a realistic way. In any case, everything else being equal, we prefer models that perform better at the task in question. This is reasonable, since language learners eventually segment quite well. To be able to evaluate our models, we need references that we can compare our model's performance to. A trivial way to show that a model does something relevant to the task at hand is to compare it with the model that makes random choices. A second method is to compare the model with a state of the art alternative. In this section, I will define two such models that will serve as a reference for the models that are developed in this study.

5.5.1 A random segmentation model

A trivial random model can be defined as one which makes a random boundary decision for each possible boundary location. For a boundary guessing algorithm, performing consistently better than this model would already indicate that the algorithm is finding something relevant for the solution of the segmentation problem. However, it is customary (since Brent and Cartwright, 1996) in speech segmentation literature to set the bar a little bit higher. The typical random baseline used in computational segmentation literature inserts boundaries with the probability of boundaries in the actual corpus. In other words, it inserts as many boundaries as in the gold-standard segmentation, however, at random locations. Throughout this thesis, performance scores obtained by this particular random model (RM) will be presented as a baseline reference. Note that the RM knows an important fact about the language that no other unsupervised models of segmentation know: the average length of words (estimated from the corpus studied). Although expected error rates E_o and E_u and boundary scores are easy to calculate for the RM, the direct calculation of the word and lexicon scores is not trivial. Table 5.3 presents all performance scores discussed in Section 5.4 for both random procedures.

Since the RM model inserts boundaries at random, its performance is varied. This variation is expected to be small for a large enough corpus. However, for additional reassurance, the results reported for RM baseline are obtained by averaging of 50 runs over the relevant corpus.

5.5.2 A reference model using language-modeling strategy

Section 5.2.2 outlined a commonly used model of segmentation. Differing theoretical and practical motivations aside, most successful computational models assign probabilities to possible segmentations in a way closely resembling the Equations 5.1 and 5.2 (repeated here as Equations 5.3 and 5.4 respectively for convenience).

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
random	27.4	50.0	35.4	8.6	13.6	10.5	7.4	38.1	12.4	50.0	50.0
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0

Table 5.3: Performance scores of two random segmentation strategies. The scores in the first row are obtained by a random algorithm that decides for boundaries with probability 0.5. The RM algorithm, as described, inserts boundaries with the probability of observing boundaries in the reference BR corpus. The scores presented are average of 50 runs, standard deviations for all scores were less than 0.01.

$$P(s) = \prod_{i=1}^{n} P(w_i) \tag{5.3}$$

$$P(w) = \begin{cases} (1-\alpha)f(w) & \text{if } w \text{ is known} \\ \alpha \prod_{i=1}^{m} P(a_i) & \text{if } w \text{ is unknown} \end{cases} \tag{5.4}$$

where w_i is the i^{th} word in the sequence (utterance or corpus), a_i is the i^{th} sound in the word, and α is the only parameter of the model (will further be discussed below).

For the incremental model defined here, a word is ‘known’, if it was used in a previous segmentation, otherwise it is unknown (for a batch model these definitions depend on the definition of the generative model). The model accepts the whole utterances as a word if the utterance do not contain any known words. The reason for this can be seen easily by an example. Assume that the input is a two phoneme utterance /ab/, where probabilities of the phonemes are P_a and P_b respectively. The probability of the two-word utterance /a b/ is $\alpha P_a \alpha P_b$, and the probability of single-word utterance /ab/ is $\alpha P_a P_b$. Since α is a value between zero and one, the second probability will be higher, and the model will take the complete utterance /ab/ as a word. This result can easily be extended to longer sequences of phonemes. In general, this model will never segment an unknown sequence of phonemes.

The major alternations to the models include use of larger word context, e.g., bigrams and trigrams, to calculate the known probabilities (e.g. Goldwater et al., 2009) or using more elaborate models of phonotactics (Blanchard et al., 2010). However, as can be seen in Table 5.2, the overall performances of the models are similar. The performance differences, when observed, are also likely to be due to processing and search strategies as well as the way the scores are calculated.

In this modeling setup, α can be interpreted as the probability of seeing a novel word. If α is large, the novel words get higher probability. If α is small, known words

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
Brent (1999a)	80.3	84.3	82.3	67.0	69.4	68.2	53.6	51.3	52.4	–	–
Venkataraman (2001)	81.7	82.5	82.1	68.1	68.6	68.3	54.5	57.0	55.7	–	–
Goldwater et al. (2009)	90.3	80.8	85.2	75.2	69.6	72.3	63.5	55.2	59.1	–	–
Blanchard et al. (2010)	81.4	82.5	81.9	65.8	66.4	66.1	57.2	55.4	56.3	–	–
$LM_{0.1}$	87.2	66.4	75.4	66.2	55.0	60.1	34.5	66.9	45.6	3.7	33.6
$LM_{0.3}$	85.1	79.8	82.3	71.8	68.6	70.2	45.8	63.4	53.2	5.3	20.2
$LM_{0.5}$	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3
$LM_{0.7}$	82.1	84.7	83.3	70.1	71.6	70.8	52.6	57.3	54.9	7.0	15.3
$LM_{0.9}$	78.2	85.9	81.9	64.9	69.4	67.1	48.2	47.6	47.9	9.0	14.1

Table 5.4: Performance scores of the baseline model with varying α in comparison to the other models using the similar strategy. All scores are obtained on the BR corpus.

are more preferable. This probability can be estimated from type/token ratio (i.e., ratio of the number of novel words seen so far to the number all words seen so far). Some models in the literature (e.g., Venkataraman, 2001) use this intuition to remove the free parameter α. Even though a parameter-free model is indeed more desirable, as will be presented shortly, the relationship between the value of α and segmentation performance is not trivial.

The model defined by Equations 5.3 and 5.4 is used as a reference model for comparison with the other models developed in this study. The generic model with free parameter α (Equation 5.4) will be called the LM_α.

The corpus used for testing the models in this thesis is the corpus used by many recent studies. This corpus was collected by Bernstein Ratner (1987) and processed by Brent and Cartwright (1996). Following the convention in the literature the corpus will be called the *BR corpus*. Some details about the corpus are presented in Appendix A. Table 5.4 presents the performance of the LM_α model on the BR corpus for changing α values in comparison with the other studies that used the same corpus (the results of other models are repetitions of the results presented in Table 5.2).

For most values of α, the performance of the LM_α model is competitive with the recent models in the literature, performing better at some scores.[5] Surprisingly, for a wide range of alpha values the differences in performance are rather small. To illustrate the effect of α on the performance, the performance scores and the error rates of the model on the BR corpus for α values in the range $[0, 1]$ are presented in Figure 5.1. The performance graphs also confirm that for a wide range of α values, the change in α values affect the performance rather slightly. The error graphs provide a better interpretation. The increasing α values decrease the oversegmentation errors and increase the undersegmentation errors with a slower rate. This particular trend is because of the fact that with higher values of α the model gives more weight to novel

[5] Note that the variants of other models are selected based on their LF.

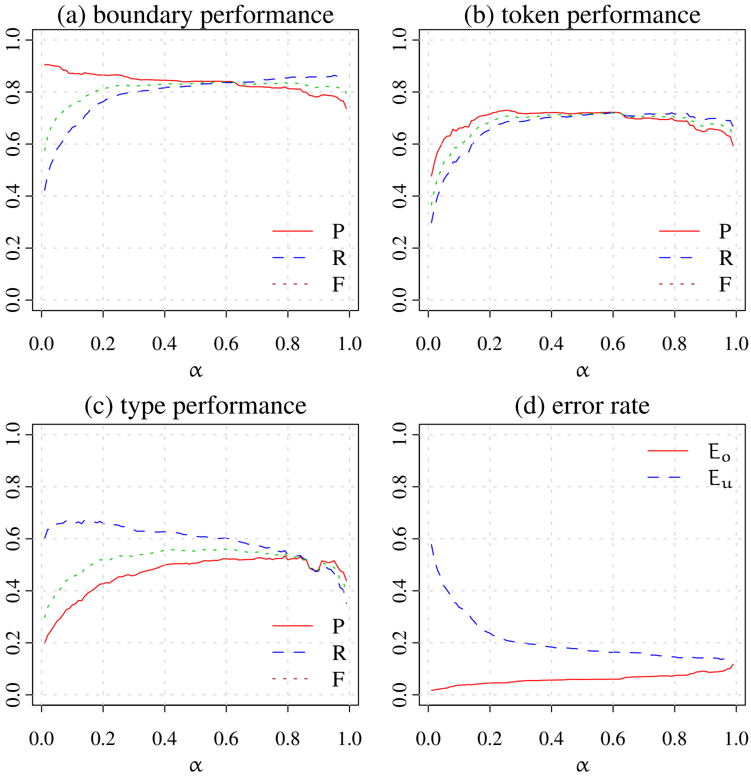


Figure 5.1: Precision (P) recall (R) and f-score (F) values (for boundaries, word tokens and word types) values and oversegmentation (E_o) and undersegmentation (E_u) errors of the LM_α on the BR corpus plotted against changing values of α.

words, and Equation 5.4 assigns higher weights to shorter words. By increasing α, model is encouraged to segment more, and this is clearly visible in the decrease of undersegmentation errors. On the other hand, since Equation 5.3 prefers fewer words, the model still does not make a large number of oversegmentation errors.

For the rest of this thesis, the LM_α model with $\alpha = 0.5$ will be used for comparing various results obtained (subscript will be dropped, and the model is simply denoted 'LM'). The LM shares the basic structure of state-of-the-art segmentation models, and it achieves competitive results with other segmentation models on the known benchmark corpus. As a result, it serves as a good reference model.

As an added benefit of reimplementing the reference model, Table 5.4 also reports the error scores described in Section 5.4. Furthermore, it also enables us to investigate an incremental model's performance over time. Notice that the best performing model in Table 5.2 is the batch Bayesian model presented by Goldwater et al. (2009). Besides the modeling practice used, there are two more reasons why this model can perform better than an incremental model. First, since it has access to complete data, in principle, it can arrive at generalizations that are consistent with the complete corpus. Second, the performance of the incremental models in the same table includes the initial output of the learning process where errors are expected. A batch model, on the other hand, outputs its results after the learning process is completed. To demonstrate the performance of the LM with the increasing input, Figure 5.2 presents the performance scores plotted for each 500-utterance block during the learning process. The first value of each score in this graph is calculated using first 500 utterances, the second value is calculated on 501th utterance to 1000th, and each successive score is calculated on the next 500-utterance block. Since the corpus contains 9790 utterances, the last scores in this graphs are calculated using the last 290 utterances. As expected, the performance scores increase and errors drop as the learning progresses. It also seems that the learning is fast, since, after the third or fourth block, the scores stabilize. At the end of the last phase of the learning from this corpus, the performance scores of the LM are substantially better than the performance scores calculated on the output of the model during the complete learning process (BF=89.0%, WF=80.6%, LF=74.0%, E_o=4.4%, E_u=11.1%). And in fact, these performance scores are also higher than the performance scores reported in Goldwater et al. (2009).

A possible objection to reporting the performance scores for last 290 utterances is that the scores can be a result of idiosyncrasies of this particular small sample. Figure 5.2 shows that despite slight fluctuations, the scores obtained for earlier blocks of 500 utterances are also similar, and Section 9.4 will provide further assurances that the results are not due to idiosyncrasies caused by chance effects.

The LM and the related models set a high standard of performance to achieve. Instead of proposing a different system, improving the LM, as in few other studies, could be another strategy to follow. However, there are a few shortcomings of this modeling strategy, as a model of human performance.

First, as the review of the relevant psycholinguistic literature suggests, there is a relatively large number of cues used by humans for the segmentation tasks. On the other hand, the LM-like modeling practice evaluates the segmentations of an utterance based only on probabilities of the lexical units used in the segmentation. The probability of the words are in turn estimated from their frequency of occurrence in the input, if the word is known. If the word is novel, the probability estimation makes use of the frequency of the phonemes. In other words, for these models the probability of a segmentation is based on (relative) frequency. The improvements of the standard model by introducing word context, e.g., using bigram counts instead of word counts, changes only the units whose frequency is calculated. Similarly, better modeling of phonotactics is generally done by using letter bi- or trigrams. In summary: the LM and its variations use frequency of lexical units or phonemes as the sole quantity to

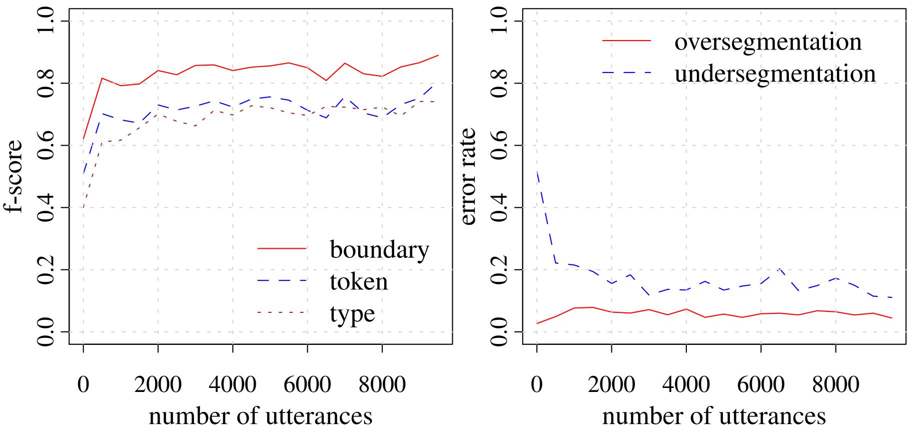


Figure 5.2: (a) Boundary, word token and word type f-scores and (b) oversegmentation and undersegmentation rates of the LM on the BR corpus for successive blocks of 500 utterances each.

estimate the probability of a segmentation. Even though some cues discussed in the literature can be fitted in the phonotactics part of the standard model (Equation 5.4), it is difficult to integrate most of the cues into this type of model.

A second problem arises because of the way probabilities are calculated by Equation 5.3 and Equation 5.4. These equations prefer very short utterances and very short words. Assigning lower probabilities to longer sequences seems to have some merit. Indeed, lexical units formed by longer sequences of phonemes, and utterances formed by longer sequences of lexical units become more and more improbable as the length of the sequence in the relevant units increases. However, the prediction of these models at the other end of the scale is wrong. These equations assign considerably higher frequencies to one-unit sequences than two-unit sequences. In other words, the equations assign high probabilities to single-word utterances, and single-phoneme words.

Table 5.5 lists a few examples of probabilities of utterances (assuming all words used in the utterances are known) and words (assuming that they are novel). The calculation was carried out on the gold standard segmentation of the BR corpus. As can be seen in Table 5.5, the length of the utterance or word is the main determining factor for the LM. A relatively rare utterance /`yu`/ ‘you’ gets a very high probability by virtue of being a high frequency single word. For the same reason, the word /`In`/ ‘in’, which never occurs as an utterance in the corpus, fares better than many highly frequent utterances, such as /`WAts D&t`/ ‘what’s that’ and /`WAt du yu want`/ ‘what do

utterance probabilities				word probabilities				
utterance	freq	rank	p	word	freq	rank	p_c	p_ℓ
yu	4	165	0.05	t	0	NA	0.09	0.07
In	0	NA	0.01	6	895	3	0.04	0.02
WAts D&t	208	2	0.0004	yu	1704	1	0.001	0.0002
bItwin	0	NA	.00003	Z	0	NA	.00002	0.0001
WAt du yu want	33	21	0.0000001	WAts	569	9	0.000002	0.0000003

Table 5.5: Example probabilities assigned to utterances by Equation 5.3 and to words by Equation 5.4. For word probabilities, p_c is calculated using relative frequencies of the phonemes in the complete corpus, and p_ℓ is calculated using relative frequencies of the phonemes in the unique words, i.e., lexicon.

you want'. Similarly, the word /bItwin/ 'between', which occurs only once in this corpus, is assigned a probability higher than 21st most frequent utterance. Similar observations can be made for high frequency non-word /t/ (as in /6bQt/ 'about'), and low frequency non-word /Z/ (as in /yuZw6li/ 'usually') and words with varying frequency /6/ 'a', /yu/ 'you' and /WAts/ 'what's'. The shorter sequences of phonemes, even when they are not words, get higher scores than the real words formed by longer sequences of phonemes. In both cases this formulation results in a strong bias towards short sequences.

As well as providing a more cognitively plausible method of modeling learning segmentation, explicit cue combination model that will be described next aims to improve the segmentation performance by solving these problems.

5.6 Summary and discussion

This chapter provided a general overview of the state of the art in computational segmentation. First, along with common modeling practices in computational models of segmentation, a review of the previous models in the literature was presented. Second, the issue of evaluation is discussed. In this discussion, common problems in interpreting performances of the segmentation models are pointed out, and two new error measures are suggested. Finally, Section 5.5 introduced a model that follows the state of the art strategy in the literature, as well as a random baseline. These models will be used as references in evaluation of the performance of the models described in the rest of this thesis. Although the reference model described here performs quite well, the aim of this study is developing a model that combines multiple cues using an incremental segmentation algorithm.

It has been well established that children, as well as adults, are sensitive to multiple, overlapping and noisy cues in the speech input. Furthermore, they use these cues in discovering lexical units in continuous speech. As in experimental studies, computational mechanisms of combinations of multiple cues for speech segmentation are under-studied. Most computational models of segmentation focus on one particular cue.

One of the reasons for this lack of interest in multiple cue combination is practical. The computational models in the literature almost exclusively focus on information that can be extracted from transcribed speech, e.g., distributional regularities or utterance boundaries. This is mostly due to the fact that we do not have corpora rich enough to include reliable acoustic cues, or good standardized representation schemes for acoustic input. Even if researchers attempt to use acoustic cues, they resort to methods of synthesizing them using canonical forms in dictionaries (Christiansen et al., 1998), or using automatic speech recognition systems (Rytting et al., 2010). The synthesized cues used in these studies may not exactly match the real-world speech input. However, the models that combine multiple cues still provide insights into how these cues can be combined, and their predictions can be tested experimentally.

Another reason for not integrating multiple cues is related to the strategy used in most segmentation models in the literature. The typical modeling practice based on n-gram language modeling presented in Section 5.5 cannot be extended easily to incorporate multiple cues. Even though one can argue that these models combine phonotactics and distributional regularities, the mechanism of combination is highly restricted, and phonotactics come into play only for unknown words, as part of a back-off mechanism. The same problem has been observed in natural language processing, where a similar modeling practice for probabilistic context free grammars (PCFGs) has been found to be difficult to modify to incorporate arbitrary aspects, or features, of the input.

In the computational modeling of multiple cues, the only computational models that emphasize cue combination are incarnations of the same connectionist system developed by Christiansen et al. and his colleagues (Allen and Christiansen, 1996; Christiansen et al., 1998, 2005; Rytting et al., 2010). Although these models clearly demonstrate that the combination of cues is useful, more explicit models of cue combination would provide better insights into the phenomenon. For example, even though these systems integrate the cues, it is difficult to assess the effects of the cue conflicts, or relative importance of certain cues directly. Starting with the next chapter, a computational model of segmentation that attempts to combine arbitrary cues using explicit statistics will be presented in a number of incremental steps.

6 Segmentation using Predictability Statistics

> Two quite opposite qualities equally bias our minds—habits and novelty.
>
> Jean de la Bruyere

Predicting things to come is a natural activity of the human brain. We have predictions, or expectations about things such as what the weather is going to be like tomorrow, what will happen in the next episode of our favorite TV show, who else would come to the birthday party we plan to go, how safe is it to walk at night in our favorite city, how many more cups of coffee may be enough to work thorough the night, or how long it will take to finish a dissertation chapter. The predictions are not only there for high-level cognitive functions exemplified here either. Many sensorimotor functions depend on predicting the state of the world at next time step. For example, our success in sports depends on our predictions, as well as what we perceive. For most of these matters, if we are not asked, we would not even know we are making prediction about them. Of course the predictions are not always accurate, and the mechanisms behind human predictions are interesting for psychology in general. The point for the discussion here is that it is a fundamental part of our cognitive processes. During the time that we are conscious, the 'internal prediction machine' never stops, predicting the next state of the environment at many levels. An interesting aspect of the human cognition is not only how we set expectations about the next step on a task, but how we react when expectations fail and we are surprised. We remember and we learn most from surprising events. It seems that prediction is an important aspect of how human cognition works, and when it fails, it has further consequences on the cognitive system.

Returning from these informal common-sense statements about prediction and surprisal to more concrete facts related to research in this study, we know from the research reviewed in Chapter 4 that predictability has an important function for speech segmentation as well. Section 4.2.1 introduced a general observation about the speech stream that aids lexical segmentation: 'predictability within the lexical units is high, predictability between the lexical units is low'. This strategy has been found to be useful for computational models of segmentation (e.g., Christiansen et al., 1998; Cohen et al., 2007; Elman, 1990; Hafer and Weiss, 1974; Harris, 1955). Moreover, it is known that a similar strategy is used by infants for segmenting continuous speech (Saffran et al., 1996a). In the upcoming sections in this chapter, I will lay out a computational model of segmentation that builds on this strategy. The next section investigates various formal measures of predictability and compares their effectiveness using statistical analysis of child-directed speech. Section 6.2 will introduce an unsupervised strategy that combines multiple measures to segment a continuous stream and the simulations carried out using this strategy.

6.1 Measures of predictability for segmentation

It is clear from the psycholinguistics literature that predictability is used by humans for the task of segmentation. Particularly, it seems when consecutive units do not predict each other, even 8-month-olds tend to assume that there is a word boundary (Saffran et al., 1996a). To formally express the notion of predictability, Section 4.2.1 introduced two measures, *transitional probability* and *successor variety*. Besides these two measures, this section will formally introduce two other measures, *pointwise mutual information* and *boundary entropy*, and present an analysis of child-directed speech that investigates usefulness of these measures as indications of word boundaries.

Before analyzing the measures listed above, I will first review a relevant study by Hockema (2006), which presents an indication of word boundaries. This indication is not suitable for unsupervised learning, but nevertheless the study uncovers an interesting property of speech sequences relevant to this research.

6.1.1 Boundary probability

Hockema (2006) analyzed a large corpus of child-directed speech according to a measure he called *conditional boundary probability*, which is defined as the probability of observing a word boundary given a phoneme pair lr:

$$P_{wb}(l, r) = P(boundary|lr)$$

He transcribed all child-directed utterances in the American English section of the CHILDES that were available at the time using the Carnegie Mellon Pronouncing Dictionary (Carnegie Mellon University, 1998). For each possible phoneme pair lr, he estimated $P_{wb}(l, r)$, and plotted the histogram of these probability values. The result showed that the distribution is strongly bimodal. Phoneme pairs show a high tendency to occur either word-internally or at word boundaries.

Figure 6.1 presents graphs produced by the same procedure on the BR corpus. The differences between data used for producing these graphs and Figure 2 in Hockema (2006) are in the size of the corpus and the number of phonemes used for transcribing the corpus. Hockema's data consisted of 8,078,540 phoneme pairs transcribed using a 39-phoneme alphabet. In contrast, the analysis used here is based on 86,019 phoneme

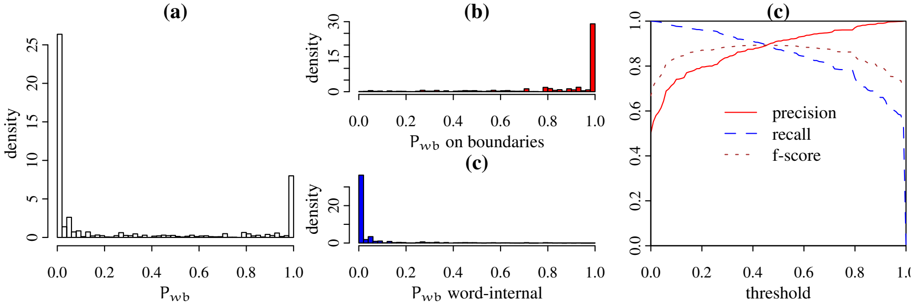


Figure 6.1: (a) Histogram of $P_{wb}(l, r)$ values. (b–c) Histograms of P_{wb} for all pairs that occur at word boundaries (b), and word-internal positions (c). (d) Precision, recall and f-score values for against changing threshold.

pairs that are transcribed with a 50-phoneme alphabet. Despite the differences, the same trends hold: the distribution of phoneme pairs is strongly bimodal.

The presentation of the data is also different. All histograms presented in this section are like Hockema's *normalized* histograms: they count the relevant values as many times as the corresponding phoneme pair occurs in the corpus. As a result, compared to a histogram that is based on phoneme-pair types, these histograms are better representations of the distributions that a child would hear.

Figure 6.1a presents the histogram of $P_{wb}(l, r)$ for all phoneme pairs that were observed in the corpus. Large portions of the probability mass are lumped together either at the very first bin where the probability of a word boundary is zero or close to zero, or on the opposite end of the scale, where the probability of a word boundary is one or close to one. This clearly shows that there is a tendency for some phoneme pairs to appear only word internally, and some others to appear on word boundaries. Figure 6.1b–c presents the two separate histograms of the same quantity. In Figure 6.1b only the probabilities of phoneme pairs that straddle word boundaries are shown, while in Figure 6.1c only word-internal phoneme pairs are considered. These histograms clearly show that bimodality of the measure is indeed due to the differences between the phoneme pairs that occur at the word boundaries and the word internal positions. Figure 6.1d presents the segmentation performance of a simple segmentation algorithm that segments between phoneme pairs for which $P_{wb}(l, r)$ is greater than a threshold value. The graph presents precision, recall and f-score for varying threshold values. The results indicate that a very high level performance is attainable for a large range of threshold values. For example, for a threshold of 0.5, we get 91.2% precision, 87.3% recall which amounts to an f-score of 89.2%. These figures seem to be somewhat lower (86.5% precision, 76.0% recall) for Hockema's larger CDS data set.

Considering that this segmentation performance can be achieved only using statistics over phoneme pairs, it is an impressive result. However, there are two major problems with this analysis. First, the learner has no access to the information needed (the knowledge of word boundaries) to build this distribution. As a result, even though it uncovers a nice regularity about the data, it is of little direct use for an unsupervised segmentation algorithm. Second, since the method is based only on phoneme pairs, there is no way of distinguishing occurrences of a phoneme pair that occurs both word-internally, and at word boundaries. This becomes particularly problematic for some of the frequent phoneme pairs. For example, /s I/,[1] occurs 153 times on a word boundary, such as in *what's **it***, and 163 times word internally, such as in ***sit***, in the BR corpus. The method suggests that either all occurrences of the word *sit* and the words including the phoneme pair /s I/ will be oversegmented, or phrases like *what's it* will be undersegmented.

The analysis provided above indicates that given a correctly segmented corpus, one can come up with relatively accurate segmentation based on the likelihood that a phoneme pair occurs at word boundaries. Even though this is not immediately useful to a learner without access to an already segmented corpus, similar results may be obtained based on various measures of predictability that do not require a segmented corpus. The rest of this section provides similar analyses for such measures.

6.1.2 Transitional probability

As a measure of predictability, most studies in the psycholinguistic literature use *conditional probability*—or *transitional probability* (TP), as it is known in this field (e.g., Saffran et al., 1996a). The conditional probability of syllable r given we already observed syllable l is defined in Equation 4.1. It is repeated here for convenience:

$$\mathrm{TP}(l, r) = P(r|l) = \frac{P(lr)}{P(l)} \approx \frac{\mathrm{frequency}(lr)}{\mathrm{frequency}(l)} \tag{6.1}$$

Instead of syllables, throughout this chapter, l and r will refer to phonemes (and later, sequences of phonemes).

Intuitively, if the phoneme pair lr is highly probable, it is likely that they are part of a word for two reasons. First, words repeat, and that makes parts of the words repeat as well. Second, since words are not formed randomly, certain sequences are more likely to be within words. These observations indicate that the *joint probability*, $P(lr)$, is a useful measure. However, if l is very frequent, the reason that lr is also frequent may often be just by chance. For example, since the phoneme /i/ is rather frequent in English, the sequence /i I/ occurs frequently even though it rarely occurs within

[1] The symbols used for phonemes in these examples and for the rest of these thesis follow the conventions used by Brent and Cartwright (1996) in transcribing the BR corpus. The transcription system is described in Appendix A.

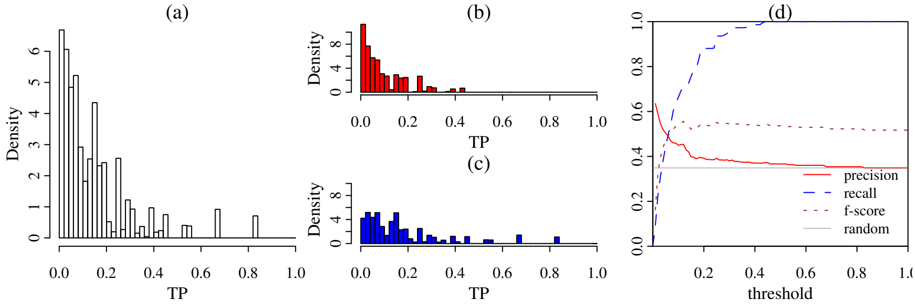


Figure 6.2: (a) Distribution of transitional probabilities. (b–c) Distribution of TP for boundaries and word-internal positions, respectively. (d) Performance of algorithms that segment at locations where P(r|l) is lower than a threshold value. The solid gray line in (d) represents precision, recall and f-score of a pseudo-random segmentation method which inserts as many boundaries as in the gold-standard segmentation.

words. On the other hand even though the phoneme sequence /WI/ occurs exclusively within words in the BR corpus, probability estimate of P(/iI/) is 3.67 times P(/WI/). As Equation 6.1 suggests, conditional probability is high if joint probability is high. The division by P(l) in the definition of TP, reduces this 'chance effect' to some extent. For the same example, even though it is still higher, TP(/i/, /I/) is only 1.71 times TP(/W/, /I/).

Figure 6.2a presents distribution of conditional probability values. Unfortunately, there is no clear indication of a bimodal distribution. If we plot histograms of the conditional probabilities at boundaries and word-internal positions separately (Figure 6.2b–c), we can see that the distributions are somewhat different. As expected, the probability mass for boundaries is found more towards the lower end of the distribution. However, even though the distribution of word-internal conditional probabilities is more spread towards the higher values, there is still a large number of word-internal positions with low conditional probabilities. Figure 6.2d presents the performance scores for a strategy that segments at the locations where conditional probability is lower than a threshold. The gray line in this graph presents the performance of a segmentation strategy where boundaries are inserted randomly with the constraint that the number of boundaries inserted is the same as the number of boundaries in the gold-standard segmentation. The random segmentation model (the RM) is explained in Section 5.5. Since precision and recall scores of the random segmentation are the same, the f-score is also the same. As a result they appear as a single line in Figure 6.2d. It should be noted that even though the boundaries are chosen at random, this particular segmentation strategy is a rather informed baseline: it knows the number of boundaries.

This analysis indicates that even though it is not as impressive as the measure suggested by Hockema (2006), a naive segmentation strategy based on TP consistently performs better than random. Furthermore, this measure is more suitable for unsupervised methods, since calculation of conditional probabilities does not require the knowledge of word boundaries.

Using threshold values for unsupervised segmentation is problematic because it requires a non-trivial way to set a threshold value without knowing which value is a good option. This problem and possible solutions will be discussed further in Section 6.2 where explicit unsupervised algorithms for segmentation will be described. The analysis provided in Figure 6.2d serves as an indication that this measure is useful, and allows us to compare it with the others.

Using the conditional probability measure as presented here has two other weaknesses. First, like the P_{wb} measure discussed above, TP calculated on only two consecutive phonemes cannot handle effects of larger sequences of phonemes or non-adjacent phonemes. This is not an intrinsic property of the measure, and use of larger phoneme context will be discussed in Section 6.1.6. Second, as it is also discussed in Brent (1999a), the conditional probability is asymmetric, $P(l|r)$ is not the same as $P(r|l)$, and $P(l|r)$ can also provide useful information for segmentation. The utility of the backward version of the measure will be discussed in Section 6.1.7.

6.1.3 Pointwise mutual information

Pointwise (or specific) mutual information is an information theoretic measure of association between two random variables. It is used in many natural language processing tasks, and its use in segmentation, albeit rare, is not exceptional (e.g., Brent, 1999a; Swingley, 2005). Pointwise mutual information (MI) is defined as,[2]

$$\mathrm{MI}(l, r) = \log_2 \frac{P(l, r)}{P(l)P(r)}$$

Neglecting the logarithm for now, in this definition, the joint probability is divided by $P(l) \times P(r)$. As a result, the high association one would get by chance for highly frequent phonemes is reduced just as in the case of TP. Unlike TP, the MI score is affected by frequencies of both phonemes, and it is symmetrical. The logarithm defines the unit of the measure. The binary logarithm (base two) is commonly used in information theory, and the resulting unit is called *bit*.

There has been some work on computational modeling of segmentation which used MI (Brent, 1999a; Swingley, 2005). However, it is virtually unmentioned in the psycholinguistic literature.

[2] *Mutual information* is a related but different information theoretic measure. However, in this thesis, following the related work in computational models of segmentation, the term mutual information and the abbreviation MI always refers to pointwise mutual information between two consecutive sequences.

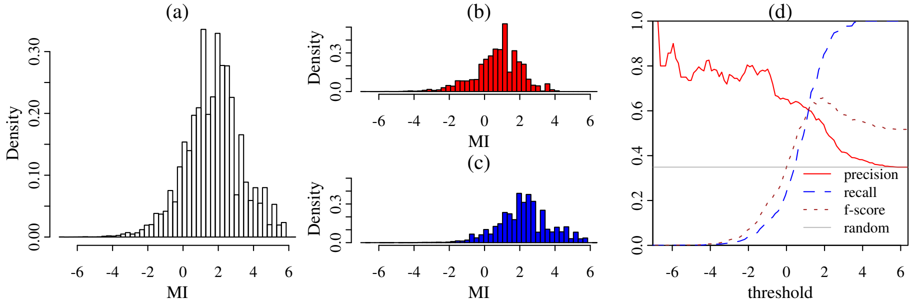


Figure 6.3: (a) Distribution of MI. (b–c) Distribution of MI for boundaries and word-internal positions, respectively. (d) Performance of algorithms that segment at locations where MI is lower than a threshold value. The solid gray line in (d) represents precision, recall and f-score of a pseudo-random segmentation method that inserts as many boundaries as in the gold-standard segmentation.

Figure 6.3 presents the same analysis for MI that Figure 6.2 presents for the TP. The first difference to note is that the shape of the graph is different from TP. This is because of the fact that the probability values are estimated from frequencies of phonemes and phoneme bigrams. Like many other frequency distributions in linguistic units, distribution of probability values, such as TP, follows an exponential trend. On the other hand, MI is the logarithm of a combination of probability values,[3] and the logarithm function transforms the exponential-like distribution into a roughly normal distribution. In addition, the difference between the distributions of MI values for boundary and non-boundary phoneme pairs seems to be slightly better separated. This is also evident from the differences of performance graphs in Figure 6.2d and Figure 6.3d. F-score for TP barely exceeds 50%, while f-score for MI is well over 60% for some threshold values. Before providing a more detailed comparison, two more measures will be introduced.

6.1.4 Successor variety

Among the measures we consider in this chapter, the *successor variety* (SV) (Harris, 1955) is probably the earliest measure suggested for lexical segmentation. Section 4.2.1 has already introduced the SV measure informally. More formally, SV can be defined as

[3] It should be noted that the quantity $\frac{\mathtt{P(l,r)}}{\mathtt{P(l)P(r)}}$ is not a probability. For positively correlated phonemes this value is greater than one, and MI score is positive.

phoneme	**h**		**i**		**z**		**k**		**w**		**I**		**k**		**R**	
SV_{BR}		16		13		22		2		0		0		0		0

Figure 6.4: Successor variety values calculated from BR corpus for the same utterance presented in Figure 4.4. Note that the phonemic transcriptions are different.

$$SV(l) = \sum_{r \in A} c(l, r)$$

where,

$$c(l, r) = \begin{cases} 1 & \text{if substring } lr \text{ occurs in the corpus} \\ 0 & \text{otherwise} \end{cases}$$

and A is the list of phonemes (the alphabet).

Unlike the measures discussed previously, SV is only a function of the initial sequence, l. In Harris (1955), this sequence is the sequence from the beginning of the utterance to the position to be evaluated. Figure 6.4 presents the successor values for the utterance /`hizkwIkR`/ ‘he is quicker’. The SV values for the same utterance determined by Harris (1955) were given in Figure 4.4. The SV value after the word *he’s* is the highest, and a reasonable algorithm based on SV would segment this utterance correctly. However, Figure 4.4 also points to a problem. As the initial sequence gets longer, the likelihood that it has never occurred before in the input increases. As a result, even for child-directed speech, which is characteristically repetitive, the SV values drop to 0 and become useless after a short initial sequence. A segmentation algorithm based on the SV values calculated as in Figure 4.4 is likely to fail to find boundaries after a few initial boundaries. There are ways to solve this problem, but even in its simple form, SV has been popular in morphological segmentation literature (e.g., Al-Shalabi et al., 2005; Bordag, 2005, 2007; Déjean, 1998; Demberg, 2007; Goldsmith, 2006; Hafer and Weiss, 1974; Stein and Potthast, 2008). Morphological segmentation is the task of segmenting words into morphemes, it is useful in many natural language processing tasks ranging from stemming to machine translation of agglutinative languages. Since words are more repetitive than utterances, the measure works better for morphological segmentation. However, the measure may benefit from some improvements in this task as well (Çöltekin, 2010).

To adapt the SV measure to the segmentation of utterances into lexical units, the discussion here is based on calculations made using a varying size phoneme context. It is not very useful to use SV as a segmentation measure calculated using a single-phoneme context. For example, in BR corpus, SV after the phoneme /`W`/ is 7 and SV after the phoneme /`t`/ is 46. A threshold value between these numbers will always segment after the phoneme /`t`/ and will never segment after /`W`/. However, to provide a comparison with the other measures, Figure 6.5 presents an analysis of SV values where boundaries are classified using the SV value of a single preceding phoneme.

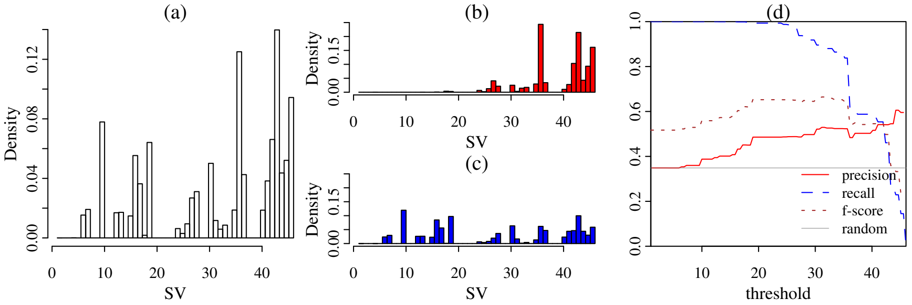


Figure 6.5: (a) Distribution of SV. (b–c) Distribution of SV for boundaries and word-internal positions, respectively. (d) Performance of algorithms that segment at locations where SV is higher than a threshold value. The solid gray line in (d) represents precision, recall and f-score of a pseudo-random segmentation method that inserts as many boundaries as in the gold-standard segmentation.

Nevertheless, Figure 6.5 indicates that, even in this form, the measure performs as well as others.

Some improvements to make SV-like measures more effective will be discussed in Section 6.1.6 and 6.1.7. The next section finalizes the discussion of individual predictability measures with a similar but theoretically more attractive and better studied measure.

6.1.5 Boundary entropy

Entropy (also called *Shannon entropy* when there is a need to distinguish from entropy in thermodynamics) is the information-theoretic measure of average uncertainty.[4] Entropy is also known as *average surprisal*, where surprisal ($-\log P(l)$) is another information theoretic measure suggested by Shannon (1948). As a result, it is one of the natural choices for measuring (un)predictability. However, in psychologically motivated models of segmentation the entropy is rarely mentioned. Use of the entropy is common in segmentation of written text, particularly for languages like Chinese and Japanese which are the typical examples of languages that use writing systems without a word boundary marker (e.g., Huang and Powers, 2003; Kempe, 1999; Zhikov et al., 2010). As far as I can determine, Cohen et al. (2007) is the only study of entropy-based segmentation motivated by human (or human-like) performance.

[4] The inventor of the measure, Claude Shannon initially named the quantity ‘uncertainty’, but based on suggestion of John von Neumann, another pioneer of the field, he named it entropy (Tribus and McIrvine, 1971).

The measure that will be used in this chapter, *boundary entropy* (H) defined as,[5]

$$H(l) = -\sum_{r \in A} P(r|l) \log_2 (P(r|l)) \tag{6.2}$$

where the sum ranges over all phonemes in the alphabet, A. Given the sequence l, this formula gives a measure of how much uncertainty still exist. As in MI, the binary (base 2) logarithm makes the unit of the measure the *bit*. In more intuitive terms, this quantity measures how many yes/no questions are necessary on average to predict the next phoneme.

Even though it may not be clear at first sight, the entropy measure has strong similarities with the SV. Both measure promiscuity of l. That is, if l combines with many different phonemes, than both SV and entropy are high. The difference is that entropy is sensitive to the token frequencies of the sequences, while SV only considers types. The difference may be easier to grasp with an example: Assume we have a corpus consisting of three words *xa*, *xb* and *xc*, and we are interested in unpredictability after *x*. Obviously SV is three, and calculating entropy using Equation 6.2 we find that entropy is 1.56 bits. If we had a corpus where *xa* occurred twice while the other two words in our previous corpus occurred once, that would not make any difference for the SV, it is still three. However, since the knowledge that *a* is a more probable phoneme after *x* reduces uncertainty, the new value for entropy (1.5 bits) reflects this.

Like SV, calculating entropy values conditioned on a single phoneme is not a good strategy. However, for the sake of completeness, Figure 6.6 presents the analysis presented for other measures for boundary entropy.

6.1.6 Effects of phoneme context

It is plausible to assume that humans do use a predictability strategy based on a larger phoneme context. Many studies in psycholinguistics showed that humans are sensitive to transitions of the syllable, which is typically a multi-phoneme unit. Furthermore, at least at some level, adults seem to be sensitive to expectations about longer and even discontinuous sequences of syllables (Dilley and McAuley, 2008). On the other hand, almost all computational models of segmentation use predictability measures calculated only on consecutive phonemes. For example, although Brent (1999a) notes that calculating TP and MI values on single-phoneme context does not reflect their full utility, he nevertheless calculates these values on the basis of single-phoneme context. Here, I will extend the analysis carried in the previous subsections and discuss the effect of calculating predictability measures on larger sequences of initial phonemes.

[5] Boundary entropy defined here is similar to but different from a well known entropy measure, conditional entropy, which is defined as $-\sum_{r \in A} P(r,l) \log_2 (P(r|l))$. In preliminary experiments conducted, the results obtained for both measures in segmentation task were similar. The boundary entropy is adopted here since it was used in previous research for segmentation (e.g., Hafer and Weiss, 1974).

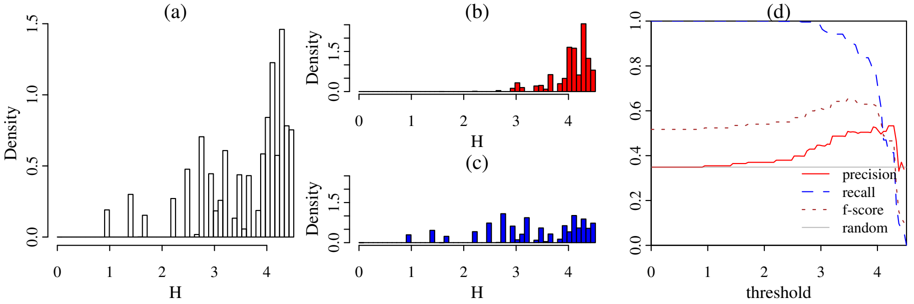


Figure 6.6: (a) Distribution of entropy. (b–c) Distribution of entropy for boundaries and word-internal positions, respectively. (d) Performance of algorithms that segment at locations where entropy is higher than a threshold value. The solid gray line in (d) represents precision, recall and f-score of a pseudo-random segmentation method that insert as many boundaries as in the gold-standard segmentation.

Figure 6.7 presents a set of graphs that visualize the effect of increasing the length of preceding phoneme context, l to two and three. The figure also provides a direct comparison of the predictability measures discussed so far. In this figure, the first three columns display the distribution of the measures with changing phoneme context size between one and three. The last column compares the performance of segmentation algorithms using a single measure with varying phoneme context size. Performance comparison is presented using precision/recall graphs. The horizontal axes of these graphs are precision values, and vertical axes are recall values. The perfect segmentation corresponds to upper right corner where both precision and recall are one. Otherwise, the closer the curve to the upper left corner, the better the performance is. In other words, a large *area under curve* is indication of a measure that performs well over a range of threshold values. First four rows, separated by dotted lines, correspond to the measures: TP, MI, SV, H respectively. Each row contains two rows of histograms, top ones depicting the distribution of the measure at boundary locations and bottom one depicting the distribution of the measure at word-internal positions. The fifth row presents precision/recall graphs comparing measures that use the same context length.

Figure 6.7 demonstrates that increasing the context size increases the separation between the distributions of boundary and non-boundary locations. This is particularly visible for context size two, and measures TP, SV and H. The separation is not that clear for MI, and for context size of three. Additionally, the increase of phoneme context from two to three does not seem to have a dramatic effect on the performance. However, there is a general trend of increase with the context size. This trend is clearly visible from the area under the precision/recall curves. Especially the precision/recall

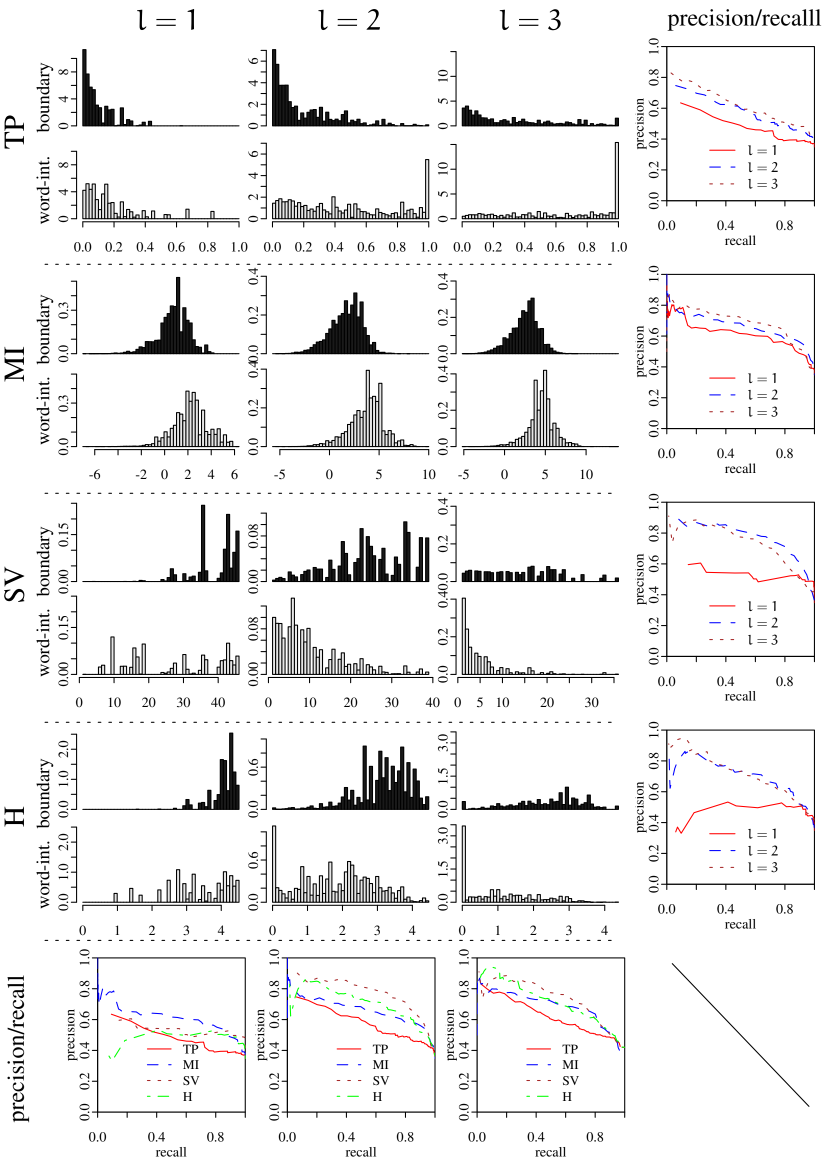


Figure 6.7: The effect of context on predictability measures. First three columns in the first four rows (rows are separated by dotted line) present distributions of measure values for varying context size. Last column presents the precision/recall graphs for each context size. The last row presents the precision/recall values for each context for all measures.

curves at the bottom row of Figure 6.7 demonstrate this clearly. The area under the curves increases in these graphs from left to right (by increasing phoneme context).

Figure 6.7 shows that increasing the phoneme context for all predictability measures affects how well they predict the word boundaries, making the measure more useful. However, an interesting question to ask is whether they give the same information or not. For example, does calculating TP conditioned on previous two phonemes give us all the information we get from calculating it by conditioning on a single previous phoneme? The question is important, because if different context sizes provide different information, than instead of using the higher context size, one can use both to achieve a better performance compared to the performance achieved by using the better of them. Using multiple context sizes is appealing, also because the unpredictability of word boundaries is due to their being dependent on different linguistic units, such as words, syllables and phonemes. Changing the phoneme context size may capture regularities that exist because of different linguistic units. The relation between the phoneme context size and the linguistic units, of course, is not clear-cut. However, for example, it is likely that a context size of two or three captures more about relationships between syllables, while context size of one mostly captures the relationships between single pairs of phonemes. If we expect regularities at both levels, then we expect combination of different context sizes to be helpful.

Section 6.2 will investigate the effect of varying context size on an unsupervised segmentation algorithm. Here, I will provide some evidence that different context sizes provide different information. The evidence comes from the fact that if two sources of information contribute independently in favor of a certain conclusion, their correlation is expected to be lower when we know the conclusion is correct. They correlate in the first place, because they measure the same quantity. However, given the conclusion, they should not be correlated if they make errors independently. If they are not completely independent, but still provide some independent information, we expect the correlation to be lower when the conclusion is known. Returning to the segmentation problem, if two context sizes, say one and two, for the same measure provide independent information regarding word boundaries, we expect their correlation after we know there is a boundary to be lower than their correlation independent of the word boundaries.

Table 6.1 presents correlation coefficients for context sizes between one and three for all measures, for all possible boundary positions, and only for word boundaries. With some variability of the magnitude of the change depending on the measure, the correlations at word boundaries are lower than the correlations for the overall corpus. The results indeed indicate that the measures calculated using each phoneme-context size provide some information about the word boundaries that the other context-length options do not provide. This result (based on cases of genuine boundaries) gives some indication that the use of statistics between phoneme sequences with varying lengths may be useful for the segmentation task. The use of information from multiple measures calculated using varying length will be investigated empirically in Section 6.2.

	1	2	3
1	1.00	0.58	0.40
2		1.00	0.74
3			1.00

(a) TP all

	1	2	3
1	1.00	0.67	0.52
2		1.00	0.80
3			1.00

(b) MI all

	1	2	3
1	1.00	0.63	0.44
2		1.00	0.74
3			1.00

(c) SV all

	1	2	3
1	1.00	0.61	0.43
2		1.00	0.74
3			1.00

(d) H all

	1	2	3
1	1.00	0.55	0.33
2		1.00	0.64
3			1.00

(e) TP boundaries

	1	2	3
1	1.00	0.64	0.46
2		1.00	0.74
3			1.00

(f) MI Boundaries

	1	2	3
1	1.00	0.35	0.27
2		1.00	0.58
3			1.00

(g) SV boundaries

	1	2	3
1	1.00	0.38	0.20
2		1.00	0.56
3			1.00

(h) H boundaries

Table 6.1: Correlation coefficients for different phoneme context sizes for each measures. The top row gives the correlation coefficients over all boundary locations. The bottom row presents the correlation coefficients calculated only at boundary positions. The correlation coefficients are calculated after a log-transforming TP, SV and H, since log-transform makes these distributions roughly normal.

6.1.7 Predicting the past

Except MI, all three predictability measures discussed in this section are asymmetric. They take an initial sequence of phonemes, and measure the predictability of the next phoneme. Moreover, the SV and entropy measures do that without actually seeing the next phoneme. It is clear that the reverse quantities that measure the predictability of the previous phoneme given the current phoneme or phoneme sequence provide some additional information. Taking TP as an example, we know that $\mathsf{P}(\mathsf{l}|\mathsf{r}) \neq \mathsf{P}(\mathsf{r}|\mathsf{l})$. If they are both useful for segmentation, using both measures is, in principle, better than using only one of them.

This section will show empirically that the reverse versions of the measures discussed so far are also good measures for segmentation. However, for a truly online-predictive system, predicting past events based on current may seem odd. The justification of using reverse predictability measures for segmentation comes from two sources. First, intuitively, it seems that what we hear at a particular moment changes our interpretation of past input, especially if the previous interpretation is uncertain in some way. It is not unusual that when reading some text or listening to someone, things we read or heard start making sense only after we hear or read more. The second, more concrete evidence is from developmental psycholinguistics. Pelucchi et al. (2009) showed that 8-month-old infants (the same age as the infants in Saffran et al. (1996a) study) were able to track statistical regularities that are only possible to detect if they were sensitive to some reverse predictability measure between the successive syllables. Pelucchi et al. (2009) carefully selected words from a natural but unfamiliar language with sequences of syllables that differed only in their 'backward' transitional probabilities. Results were similar to Saffran et al. (1996a), confirming that infants do use backward predictability.

Since the direction does not make sense for MI,[6] only the reverse versions of TP, SV and H will be analyzed in this section. Reverse measures will be indicated by a subscript 'r' here. The reverse of TP and SV are sometimes abbreviated as BTP (backwards TP) and PV (for predecessor variety) in the literature. It is easy to deduce the definitions of reverse measures from the forward counterparts. The definitions are provided here for the sake of completeness.

$$\mathrm{TP}_r(l, r) = P(l|r) = \frac{P(lr)}{P(r)} \approx \frac{\mathit{frequency}(lr)}{\mathit{frequency}(r)} \tag{6.3}$$

$$\mathrm{SV}_r(r) = \sum_{l \in A} c(l, r) \tag{6.4}$$

where,

$$c(l, r) = \begin{cases} 1 & \text{if substring } lr \text{ occurs in the corpus} \\ 0 & \text{otherwise} \end{cases}$$

and A is the set of phonemes (the alphabet).

$$\mathrm{H}_r(r) = -\sum_{l \in A} P(l|r)\, log_2 P(l|r) \tag{6.5}$$

As can be seen in Figure 6.8, the reverse measures seem to achieve similar segmentation performances as their forward counterparts. From their mathematical formulation, it is clear that the forward and reverse versions of the measures are not equal to each other. $P(l|r) \neq P(r|l)$, and SV_r and H_r calculations do not even share the strings that they are calculated on with their forward counterparts. Like the analysis for varying phoneme-context length in Section 6.1.6, we can also check if correlation between forward and reverse version of these measures provide independent information. Since both are useful for detecting boundaries, they will naturally be correlated. However, if they provide some independent information, we would expect the correlation of the measures for the boundary locations to be lower than the correlation for the complete corpus. Indeed, the correlation coefficients for TP, SV and H and corresponding reverse measures on the BR corpus are 0.62, 0.15 and 0.21 respectively. And when calculated only on boundary locations, the same measures are 0.52, -0.01 and -0.06. The question as how to combine the forward and backward information efficiently still remains, to which we will return in Section 6.2.

[6] This is not strictly true if phoneme sequences of unequal length are used for l and r. However, for ease of comparison this section only considers measures calculated on single phonemes.

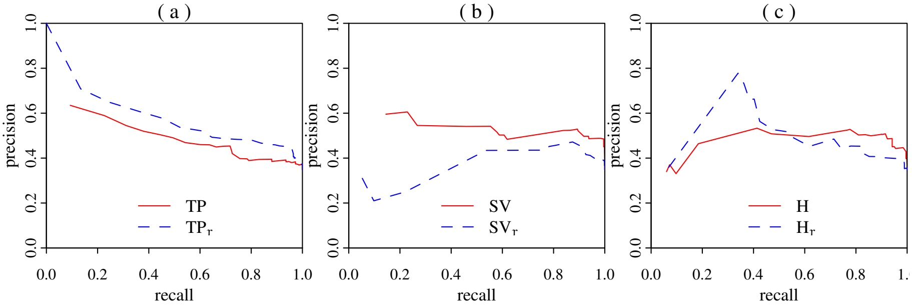


Figure 6.8: The precision/recall curves comparing the forward and reverse predictability measures: (a) TP and TP$_\mathtt{r}$, (b) SV and SV$_\mathtt{r}$, (c) H and H$_\mathtt{r}$.

6.1.8 Predictability measures: summary and discussion

So far, this chapter discussed four predictability measures: transitional probability, mutual information, successor variety and entropy. The reason all these measures work is that in an unsegmented speech stream, predictability inside the lexical units is high and predictability at the lexical unit boundaries is low. Our analysis is based on two consecutive sequences of phonemes $\mathtt{l}$ and $\mathtt{r}$. In informal terms, TP measures how likely it is to observe $\mathtt{r}$ after $\mathtt{l}$ is observed. If TP is high, we expect to be within a unit, if TP is low it indicates a possible boundary. MI measures whether $\mathtt{l}$ and $\mathtt{r}$ are highly associated or not. Again if MI is high, we expect $\mathtt{lr}$ to be a word-internal sequence, otherwise at a boundary. The other two measures, SV and H, are measures of unpredictability (surprise). Hence, high values of SV and H indicate word boundaries. Another difference of these measures is that they are functions of only $\mathtt{l}$. Informally, they try to answer the question 'how much do I (not) know about $\mathtt{r}$ after observing $\mathtt{l}$?'. The difference between these two measures is in their sensitivity to the distribution of the sequences that follow $\mathtt{l}$. Entropy is affected by the frequency of these sequences, while SV is oblivious to it.

All measures discussed in this section so far have some overlap in what they measure, but they are not the same. Most psycholinguistic studies consider TP as the measure of predictability, but the results from these experimental studies are compatible with all four. For example, given a sequence similar to the stimuli presented to the infants in Saffran et al. (1996b) and subsequent studies, Table 6.2 presents the values of all measures discussed so far for two syllable pairs. One of the syllable pairs /`bi-da`/ is part of one of the artificial words /`bidaku`/ that form this sequence, while the other /`ku-pa`/ is not. Table 6.2 shows that, as expected, all measures indicate a higher chance for a word boundary between /`bi-da`/ compared to /`ku-pa`/. It would be interesting to see experimental results that would be compatible with only one of the

	TP	MI	SV	H	TP_r	SV_r	H_r
/bi-da/	1.0	3.4	1.0	0.0	1.0	1.0	0.0
/ku-pa/	0.5	2.4	2.0	1.0	0.5	2.0	1.0

Table 6.2: The predictability scores for syllable sequences /bi-da/ and /ku-pa/, given the sequence /bidakupadotigolabubidakugolabupadoti/ is observed. Note that for the TP and the MI lower values, and for the SV and the H higher values indicate word boundaries.

	TP	MI	SV	H
TP	1.00	0.77	-0.45	-0.40
MI		1.00	-0.51	-0.43
SV			1.00	0.76
H				1.00

(a) All phoneme pairs.

	TP	MI	SV	H
TP	1.00	0.77	-0.10	-0.13
MI		1.00	-0.13	-0.09
SV			1.00	0.82
H				1.00

(b) Boundaries.

Table 6.3: Correlation coefficients of predictability measures for all phonemes in BR corpus (a) and for the phoneme pairs that straddle a word boundary. The coefficients are calculated after log-transforming the TP, SV and H values.

measures but not the others. However, it is a difficult task to design such an experiment.

The analysis in this section showed that all the measures discussed here do something relevant to segmentation, all scoring consistently over a random (but non-trivial) baseline. The performance analysis done by plotting precision/recall curves or by plotting precision, recall and f-scores gives an indication of the potential of a particular measure. The way they are used in an actual learning algorithm in combination with other information may result in different performance. Here, I will provide another way to look at the similarities and differences of these measures before switching to explicit models of segmentation with concrete algorithms. Table 6.3 presents the correlation coefficients for all (forward) measures calculated on the BR corpus.

Table 6.3a confirms that all four measures are correlated. However, TP and MI are more strongly correlated with each other compared to their correlations with SV and H. Similarly SV and H are more strongly correlated with each other. Hence, the four measures fall into two groups: TP and MI in one, and SV and H in another. The correlations between former and the latter group of measures is negative, since the former two measure predictability and the latter two measure unpredictability. Table 6.3b gives the correlation coefficients of the measures where a boundary is observed. This also reveals an interesting relationship between these groups of measures. Given boundaries, the correlations between the groups drop substantially, while correlations within the groups do not change much. This is an indication that the measures within the same group are highly dependent, while being relatively (conditionally) independent of the measures in the other group. Similar to the analysis provided for varying phoneme context size in Section 6.1.6, this is an indication that a learning algorithm that combines measures from different groups will gain additional information, while an algorithm that uses measures of the same sort will not.

This section provided an analysis of four measures of (un)predictability for their use in lexical segmentation. All of them measure something relevant to segmentation as they all perform better than a random segmentation baseline. The analysis also showed that the use of additional context improves their performance, and it is useful to consider the reverse of the asymmetric measures. Further analysis showed the similarities and differences between these measures. Next section will layout unsupervised computational models of learning segmentation that build on these measures.

6.2 A predictability based segmentation model

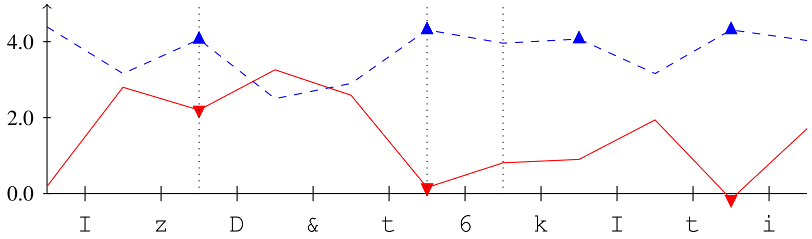
Existing predictability-based computational models of segmentation typically use a single measure of predictability calculated on single phoneme (and rarely syllable) contexts. However, the analysis of child-directed utterances in Section 6.1 indicates that the four measures discussed (transitional probability, mutual information, successor value and boundary entropy) are useful indicators of word boundaries. This analysis has also shown that even though these measures are similar in many ways, they measure different aspects of the input. As a result, the combination of these measures should help finding boundaries more than each measure alone. Another aspect discussed during this analysis is the effect of the phoneme context, which is also shown to affect the performance of the measures. According to the analysis, increasing the number of phonemes that the measures are calculated on, and combining measures calculated on varying context size is expected to increase the performance. Section 6.1 presented the effectiveness of each measure using a simple threshold based algorithm, leaving the development of an unsupervised algorithm that combines information from multiple sources for later. This section aims to fulfill this promise by developing an unsupervised algorithm for learning lexical units from continuous speech in a number of incremental steps.

6.2.1 Peaks in unpredictability

Besides the non-trivial problem of choosing a threshold, the segmentation algorithms based on thresholds do not exploit the relation between predictability and lexical units fully. Deciding for a boundary when an unpredictability measure exceeds a threshold (or equivalently a predictability measure is less than a threshold) is in line with the idea that predictability is low between the lexical units. However, the thresholds do not directly utilize the fact that predictability is high within the units. In the following pages a completely unsupervised strategy that explicitly attends to high predictability within the units and low predictability between the units will be discussed. That is, this strategy posits a boundary if an unpredictability measure at

		I		z		D		&		t		6		k		I		t		i	
TP	0.05		0.17		**0.15**		0.29		0.39		**0.03**		0.06		0.08		0.25		**0.03**		0.26
MI	0.19		2.80		**2.20**		3.26		2.59		**0.16**		0.81		0.90		1.94		**-0.15**		1.71
SV	38		19		**43**		10		30		**46**		36		**43**		19		**46**		42
H	4.39		3.16		**4.06**		2.50		2.90		**4.31**		3.96		**4.07**		3.16		**4.31**		4.03
TP$_r$	0.09		0.29		**0.11**		0.31		0.18		0.07		**0.05**		0.06		0.16		**0.06**		0.09
SV$_r$	42		34		38		40		**41**		40		40		**42**		41		33		39
H$_r$	4.47		3.57		**3.80**		3.37		**3.87**		3.50		4.06		**4.47**		3.87		4.12		4.41

(a) Example predictability scores



(b) Graphical representation of MI and H values.

Figure 6.9: Predictability measures for example utterance /IzD&t6kIti/ 'is that a kitty'. (a) presents all predictability measures discussed in this chapter calculated on the BR corpus using single-phoneme context. The values where unpredictability peaks are marked with boldface. (b) represents a graphical representation of the MI (solid line) and the H (dashed line) values for the example utterance. Dotted vertical lines mark expected boundary locations, and the triangles mark the positions where the measures indicate a boundary according to peak criterion. Note that 'valleys' rather than peaks are indications boundaries for MI.

the position is greater than the measure before and after the position. Following the previous research (e.g., Hafer and Weiss, 1974; Harris, 1955), I will call the strategy *peak-based* predictability strategy. However, it should be stressed that the term peak is valid for only unpredictability measures, such as SV and H. For predictability measures such as TP and MI, we look for 'valleys' rather than peaks. As well as reflecting the intuition 'high predictability within the words, low predictability between words', the peak based segmentation strategy is also completely unsupervised: we do not need to tune any parameters, or use any labeled data where word boundaries are segmented.

Figure 6.9a presents values for all the measures discussed in this Chapter for each possible boundary position in the utterance /IzD&t6kIti/ 'is that a kitty'. The measures calculated for the beginning and the end of the utterance are useful for discovering peaks at neighboring positions, but, for a segmentation algorithm, there is no point in trying to discover boundaries at these locations. The values where the peak strategy suggests a boundary for each measure are indicated with boldface. Figure 6.9b represents the values for MI and H graphically.

The measures presented in Figure 6.9 are calculated using single phoneme contexts. That is, the sequence l and when required the sequence r are taken to be single phonemes. As a result, performance of a peak based segmentation algorithm is bound to be adversely affected by the short context length. Since SV, SV_r, H, and H_r are functions of only l or only r, their performance is particularly low, for the example in Figure 6.9. However, unlike the threshold strategy which gives the same decision before or after a certain phoneme, the peak strategy considers the surrounding values as well. As a result, even with short context used for calculating the measure, the segmentation decision is affected by a larger surrounding context.

Even though the benefits of peak strategy for discovering boundaries are clear, there are a few weaknesses to note here. First, the peak-based boundary decision is rather conservative. It requires both sides of the boundary candidate to have the right kind of slope. Even a very sharp increase on one side will be discarded unless it is followed by a fall. Considering that most of the measures we discussed here are asymmetric, and their indications are stronger in one direction than the other, the problem certainly deserves some attention. This problem becomes more serious for single-phoneme words. Since peak based algorithm never makes two boundary decisions on a row, it never detects single-phoneme words. This problem will be revisited in Section 6.2.4. A second problem that I will leave relatively unexplored in this study is the fact that peaks do not take into account how steep the slopes are. Intuitively, the sharper the slope the higher the expected boundary indication. However, the peak-based boundary decisions used here ignore this fact.

The peak-based segmentation method that is demonstrated informally in Figure 6.9 can easily be implemented as an unsupervised segmentation algorithm for each measure alone. A possible realization of the peak-based segmentation is described in Algorithm 6.1. For all measures, the algorithm essentially follows the same steps. The predictability measures for each phoneme position in the utterance are calculated using the definitions given in Section 6.1. Unlike the values presented in Figure 6.9, the calculation of measures is not done using the complete corpus. The frequencies of phonemes and phoneme pairs are updated in an incremental fashion, using only the corpus seen so far. The beginnings and ends of the utterances are treated as special phonemes for the calculation of the measures, otherwise the utterance boundaries are not used as separate cues.

Table 6.4 presents the results obtained on the BR corpus for each predictability measure by Algorithm 6.1 in comparison to the random baseline (RM) and the reference recognition algorithm (LM) described in Section 5.5. As described in Section 5.5 the random baseline is not completely random. It knows about an important fact about the language: probability of word boundaries, which is not available to any of the models presented in this thesis.

Results in Table 6.4 indicate clearly that the performance of the peak-based prediction strategy as used here is far behind the LM. However, the results also show that for all of measures, the algorithm performs consistently better than random. As it will be discussed next, this is all we need to know about these measures for now.

Using peaks in unpredictability, Algorithm 6.1 exemplifies a completely unsupervised method of segmentation. However, two other problems raised in Section 6.1,

Algorithm 6.1: A peak-based segmentation algorithm.

```
Input: A sequence of utterances without word boundaries
Output: The sequence of utterances with boundaries
foreach utterance $u$ in the input do
    foreach phoneme position $i$ in $u$ do
        Update frequencies of phoneme$_i$, phoneme$_{i+1}$ and phoneme-pair$_{i,i+1}$;
        $P_i \leftarrow$ predictability value between $i$ and $i + 1$;
        if $P_{i-2} > P_{i-1}$ and $P_{i-1} < P_i$ then
            insert a boundary between phoneme$_{i-1}$ and phoneme$_i$;
        end
    end
    output the segmented utterance ;
end
```

	boundary			word			lexicon			error	
measure	P	R	F	P	R	F	P	R	F	E_o	E_u
TP	57.6	68.9	62.7	42.8	48.7	45.6	15.0	37.2	21.3	19.2	31.1
MI	66.3	74.1	70.0	52.2	56.6	54.3	18.5	42.5	25.8	14.3	25.9
SV	49.3	53.4	51.3	34.3	36.3	35.3	12.3	38.1	18.5	20.7	46.6
H	51.3	56.5	53.8	38.1	40.8	39.4	13.8	38.8	20.4	20.3	43.5
TP_r	53.3	67.5	59.6	36.3	43.1	39.4	14.4	35.5	20.5	22.4	32.5
SV_r	36.7	40.0	38.3	22.7	24.1	23.4	8.4	32.3	13.3	26.0	60.0
H_r	43.5	49.6	46.3	28.9	31.7	30.2	10.1	33.7	15.6	24.4	50.4
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 6.4: Boundary/word/lexicon Precision/recall/f-score values and oversegmentation and undersegmentation error for the peak-based segmentation algorithm on the BR corpus. RM represents a pseudo-random segmentation that inserts a word boundary with the probability of word boundaries in the gold-standard segmentation. The LM is the recognition-based reference model. Both models are described in Section 5.5. The performance and error scores are described in Section 5.4.

combination of measures and making use of larger phoneme context, are still left unanswered. The next subsection will offer solutions to these problems, starting with the former.

6.2.2 Combining multiple measures and varying phoneme context

The discussion so far supports the expectation that using multiple measures and varying context size may be beneficial for segmentation performance. Using multiple measures is expected to be better than a single one, since, even though they have a lot in common, each measure seems to be measuring some aspects of the input that the others do not. It was also shown in Section 6.1 that the phoneme context size makes a difference in the performance of all measures. Furthermore, combining the measures calculated on varying phoneme context size was also conjectured to be useful. Here, Algorithm 6.1 will be extended to handle multiple sources of information coming from multiple measures calculated on varying phoneme-context length.

In its essence, the peak-based segmentation method presented in Algorithm 6.1 is a binary classifier. It classifies each possible boundary position in an utterance as boundary or non-boundary. Using different measures results in multiple classifiers that do the same task. Viewing the problem as combining a number of classifiers for achieving a better performance than each individual classifier is a relatively well studied problem in the machine learning literature, where the sets of classifiers are known as *ensembles* or *committees* (e.g., Bishop, 2006, chapter 14). For an effective combination, the classifiers should be accurate and diverse (Hansen and Salamon, 1990). Accuracy refers to the requirement that the individual classifiers perform better than random. Diversity is taken as the requirement that, to some extent, the classifiers are independent. Most combination methods in machine learning, such as *bagging* and *boosting*, are typically suitable for supervised classifiers. However, the field offers a set of practical and theoretical tools for the problem at hand. Here a simple and well-known method, *majority voting*, will be used for combining the multiple measures for segmentation.

As well as machine learning applications, majority voting is also a common (and arguably effective) method in everyday social and political life. As a result it has been well studied, and known to work well especially if the accuracy and the diversity requirements are met. A theoretical justification of majority voting is given by well-known 'Condorcet's jury theorem' which dates back to late 18th century (Boland, 1989). Provided that each member's decision is better than random, and the votes are cast independently, the Condorcet's jury theorem states that the probability that a jury arrives at the correct decision by majority vote monotonically approaches to one as the number of members is increased. Informally, this states that in the long run the decision of a large number of less competent individuals is better than the decision of a single individual with the greatest competence. In practice, even though the votes are almost never independent (especially in the social scene) majority voting is still an effective way of combining outcomes of multiple classifiers (see Narasimhamurthy, 2005, for a recent review and the discussion of effectiveness of the method).

Algorithm 6.2: The majority voting algorithm for multiple measures and multiple context size. The function $\mathtt{m}()$ at line 9 calculates the predictability score (hence, unpredictability measures are multiplied by -1) according to measure $\mathtt{m}$ on given sequences of phonemes. If required n-gram is not available, the algorithm backs off to the n-gram with the highest available rank.

```
   Input: A sequence of utterances without word boundaries and the maximum
          context size $M$
   Output: The sequence of utterances with boundaries
1  foreach utterance $u$ in the input do
2      for $n = 1 \ldots M + 1$ do
3          update $n$-gram frequencies for the $n$-grams in $u$;
4      end
5      foreach phoneme position $i$ in $u$ do
6          $votecount \leftarrow 0$;
7          foreach measure $m$ do
8              foreach context size $n = 1 \ldots M$ do
9                  $P_i \leftarrow m(\text{n-gram ending at i-1}, \text{phoneme}_i)$ ;
10                 if $P_{i-2} > P_{i-1}$ and $P_{i-1} < P_i$ then
11                     $votecount \leftarrow votecount + 1$ ;
12                 else
13                     $votecount \leftarrow votecount - 1$ ;
14                 end
15             end
16         end
17         if $votecount > 0$ then
18             insert a boundary between $\text{phoneme}_{i-1}$ and $\text{phoneme}_i$;
19         end
20     end
21     output the segmented utterance ;
22 end
```

The majority voting provides a simple way to incorporate the information from multiple and (somewhat) independent measures and the information provided by calculating these measures on varying context size. Instead of calculating a single value for a measure and for a given context size, we can calculate multiple values for multiple measures with multiple context sizes. Each *measure–context size* pair forms a voter. If there is a peak in unpredictability according to this pair, we get a boundary vote. If the majority of the voters vote for a boundary for a possible segmentation

	boundary			word			lexicon			error	
max. context	P	R	F	P	R	F	P	R	F	E_o	E_u
1	73.6	68.2	70.8	57.3	54.3	55.8	16.7	49.5	24.9	9.2	31.8
2	86.6	72.9	79.2	70.7	62.8	66.6	22.0	58.6	32.0	4.3	27.1
3	89.7	77.5	83.1	75.6	68.3	71.8	27.7	63.3	38.6	3.4	22.5
4	93.4	73.6	82.3	76.4	65.0	70.2	26.1	63.1	36.9	2.0	26.4
5	94.1	72.2	81.7	76.3	63.7	69.4	26.2	64.0	37.2	1.7	27.8
6	94.9	66.1	77.9	73.4	57.7	64.6	22.8	61.7	33.3	1.4	33.9
7	95.1	63.5	76.2	72.4	55.5	62.8	21.4	60.1	31.5	1.2	36.5
8	95.4	58.7	72.7	70.3	51.2	59.2	19.6	58.3	29.3	1.1	41.3
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 6.5: Performance error scores for peak-based majority voting algorithm with varying context. Two reference models, the RM and LM are defined in Section 5.5.

position, we insert a boundary at that position. Algorithm 6.2 describes this version of the segmentation method using majority voting. For the forward measures, context size defines the length of the sequence l, while for the reverse measures context size defines the sequence r. For all boundary positions, the number of votes Algorithm 6.2 considers is equal to 'the maximum context size' times 'the number of measures'. For example, assuming that we run the algorithm only for H and H_r with the maximum context size of two, and the algorithm is about to decide if there is a boundary after *ki* in *akitty*, it checks each condition

1. $H(i) > H(k)$ *and* $H(i) > H(t)$
2. $H_r(i) > H_r(k)$ *and* $H_r(i) > H_r(t)$
3. $H(ki) > H(ak)$ *and* $H(ki) > H(it)$
4. $H_r(ki) > H_r(ak)$ *and* $H_r(ki) > H_r(it)$

Then, the algorithm increases the vote count by one for each condition met. If the vote count is greater than half of the votes (two in this case) it inserts a boundary.

The results of combining all measures with varying context size using majority voting on the BR corpus are presented in Table 6.5. Each line in the table lists the common segmentation scores we use in this chapter for context size between one and eight. Maximum context size one means that the measures are calculated with single phoneme context. As a result, the scores in the first line of Table 6.5 are obtained by the majority decision of seven voters (TP, MI, SV, H, TP_r, SV_r and H_r, all calculated on single phoneme context), while the scores in line two are obtained by the majority decision of 14 voters, each representing context sizes one or two for all seven measures.

The results certainly improve compared to single-measure segmentation results presented in Table 6.4. Some of the scores also exceed the performance of the LM, the recognition-based reference model. The performance of the majority voting algorithm is good at spotting boundaries and words. The boundary and word precision scores are consistently better than the corresponding recall scores. When increasing maximum context parameter, both precision and recall increase at first. This is expected since we incorporate information from higher level n-gram frequencies that are good predictors of the boundaries. After context length three, the recall starts to go down, while precision still gets better with the increased parameter value. Since the increased number of voters requires a higher consensus, it is natural that the precision is high. However, the higher number of voters also means that the disagreement on real boundaries will also increase. As a result recall drops. With the decreased boundary recall, the word and lexicon precision start going down as well. One of the reasons for this may be because higher level n-grams suffer from data sparseness, the voters that use higher level n-grams start to become less competent. As a result, increasing the number of voters that calculate the results on higher level n-grams violates the requirement of the successful combination that the individual voters need to perform better than random.

Despite being precise at spotting boundaries (and as a result words) the majority voting algorithm is still bad at lexical precision. The low lexical scores mostly stem from two causes. The first reason has to with the fact that this algorithm does not build and use an explicit lexicon. As a result it does not get any reward for reusing the previously discovered lexical items. Second, the algorithm starts with no prior knowledge at all, and it takes time to build useful n-gram statistics. Until a reasonable amount of statistics is collected, many wrong word-types are inserted into the lexicon, and this affects the lexical precision adversely. The use of an explicit lexicon will be investigated in Chapter 8. The effects of the lack of information at the beginning of the segmentation process will be discussed in Chapter 9. Before concluding, the rest of this chapter will present some improvements to Algorithm 6.2.

6.2.3 Weighing the competence of the voters

The majority voting algorithm presented in Section 6.2.2 treats all the voters equally. Even though this may be a virtue in the social and political context, it is a shortcoming for a learner. A better learner is expected to identify the value of the information provided by each source, and increase the weight of the sources that perform well consistently. Weighted majority voting is an extension of the majority voting algorithm which weighs the vote of each source according to their competence (Littlestone and Warmuth, 1994).

For the particular instantiation of the weighted majority voting algorithm used here, we will first assign a weight, w_i, in range $[0, 1]$ to each voter. Second, instead of increasing or decreasing vote count by one, we will increase or decrease the vote count by w_i. To do that we replace line 11 in Algorithm 6.2 with '$\mathit{votecount} \leftarrow \mathit{votecount} + w_i$' and replace line 13 with '$\mathit{votecount} \leftarrow \mathit{votecount} - w_i$'. The rest of the segmentation algorithm is essentially the same. Note that if all weights are set to one, the algorithms are equivalent.

	boundary			word			lexicon			error	
max. context	P	R	F	P	R	F	P	R	F	E_o	E_u
1	72.1	71.7	71.9	57.0	56.8	56.9	17.4	48.3	25.6	10.5	28.3
2	83.7	77.6	80.5	70.3	66.6	68.4	25.1	59.4	35.3	5.7	22.4
3	89.3	78.2	83.4	75.6	68.9	72.1	28.0	62.8	38.8	3.5	21.8
4	92.7	76.0	83.5	77.2	67.4	72.0	28.4	65.1	39.6	2.3	24.0
5	94.1	71.4	81.2	75.8	62.8	68.7	26.3	64.8	37.4	1.7	28.6
6	94.7	66.8	78.3	73.9	58.5	65.3	23.5	63.2	34.3	1.4	33.2
7	95.1	62.1	75.2	71.9	54.3	61.8	21.1	60.6	31.3	1.2	37.9
8	95.1	58.5	72.4	70.2	51.1	59.1	19.6	58.5	29.4	1.1	41.5
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 6.6: Performance error scores for peak-based weighted majority voting algorithm with varying context. Two reference models, the RM and LM are defined in Section 5.5.

So far we have described how to adjust the majority voting algorithm to be able to weigh its sources of information. However, we also need a way of setting the weights, so that they reflect the usefulness of the particular voter's decision. As with many examples in the literature, we will set all the weights to one at the beginning. After each decision, we will update the weights. In supervised models, where exact error is known, one can adjust weights in a way to reduce the error. Here we do not know boundary locations, and we cannot be certain about which decisions are correct. However, we will take the (weighted) majority decision as the correct decision. That is, if the voter agrees with the majority decision, we count this as a correct decision, and if it disagrees we will assume that it made an error. To finalize our adjustments to Algorithm 6.2, we keep count of errors made by each voter i, e_i, which is incremented when the voter does not agree with the majority decision. After the every boundary decision, first the error counts are updated for each voter. Then, the weights w_i, of all voters are updated using,

$$w_i \leftarrow 2\left(0.5 - \frac{e_i}{N}\right)$$

where N is the number of boundary decisions so far, including the current one.

This update rule sets the weight of a voter that is half the time wrong (a voter that votes at random) to zero, eliminating the incompetent voters. If the votes of a voter are in accordance with the rest of the voters almost all the time, the weight stays close to one.

The performance scores of the weighted majority algorithm for the maximum phoneme context parameter between one to eight on the BR corpus are presented in Table 6.6. In general weighted majority voting algorithm performs slightly better than majority voting algorithm. The performance of the algorithm can be improved by further extensions, for example, by using a better method for setting weights, or

	boundary			word			lexicon			error	
max. context	P	R	F	P	R	F	P	R	F	E_o	E_u
1	52.5	89.2	66.1	34.2	51.2	41.0	24.9	30.3	27.3	30.5	10.8
2	63.7	92.5	75.4	49.6	65.4	56.4	34.3	39.7	36.8	19.9	7.5
3	72.4	92.7	81.3	60.5	72.5	66.0	36.8	50.8	42.7	13.3	7.3
4	79.8	90.3	84.7	68.5	74.9	71.5	38.1	60.6	46.8	8.7	9.7
5	84.0	85.6	84.8	71.8	72.8	72.3	34.8	65.9	45.5	6.2	14.4
6	86.2	80.2	83.1	72.5	69.0	70.7	30.2	66.0	41.4	4.8	19.8
7	87.5	75.1	80.9	72.2	64.9	68.4	26.2	63.6	37.1	4.0	24.9
8	88.1	70.8	78.5	71.2	61.3	65.9	23.8	61.7	34.3	3.6	29.2
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 6.7: Performance and error scores for peak-based weighted majority voting algorithm that incorporates the information from local changes at the both sides of the boundary candidate. Two reference models, the RM and LM are defined in Section 5.5. The performance and error scores are defined in Section 5.4.

using modified versions of peak-based boundary detection. However, the purpose of the current work is not to find the best performing segmentation algorithm, but rather proposing an explicit model of segmentation that combines information from multiple sources. The weighted version of the algorithm is more attractive in this regard. First, it makes it easy to include possibly irrelevant sources of information. If they are irrelevant they will be left out by the weight update procedure reducing their weights to zero. Second, it may explain certain shifts during learning. For example, a weak cue that is not very useful before enough input is seen may become stronger in time, as its predictions become more effective with the additional information. In other words, a weak source of information may be bootstrapped by other sources if the information it collects is relevant.

6.2.4 Two sides of a peak

While describing the peak-based segmentation decision, Section 6.2.1 also pointed out a particular weakness of the peak criterion defined here. It is too conservative, and in some cases this is a serious problem. For example, since there cannot be two peaks in a row, it can never find single-phoneme words. This is also evident in the performance scores presented so far, all combinations presented have high precision (and low oversegmentation error), but low recall (and high undersegmentation error).

The solution to this problem has been delayed up to this point since the combination methods described in previous sections provide a natural approach. We can interpret the increase or decrease of uncertainty on either side of a boundary differently. The majority voting algorithm can easily incorporate these additional voters' decisions. The weighted majority voting provides an additional reassurance by eliminating useless

	boundary			word			lexicon			error	
max. context	P	R	F	P	R	F	P	R	F	E_o	E_u
1	51.6	87.0	64.8	32.6	48.4	39.0	23.7	31.6	27.1	30.8	13.0
2	63.1	92.0	74.8	48.9	64.8	55.7	33.3	40.2	36.4	20.4	8.0
3 (*PM*)	69.6	92.5	79.5	56.9	70.2	62.9	36.7	49.8	42.3	15.3	7.5
4	76.6	90.9	83.2	65.0	73.5	69.0	38.3	60.3	46.8	10.5	9.1
5	81.5	87.2	84.3	69.6	73.0	71.3	34.9	64.7	45.3	7.5	12.8
6	84.3	82.7	83.5	71.5	70.6	71.0	32.5	67.5	43.8	5.8	17.3
7	85.3	77.6	81.3	70.6	66.1	68.3	28.3	65.9	39.6	5.0	22.4
8	86.1	73.2	79.1	69.9	62.5	66.0	25.1	63.5	36.0	4.5	26.8
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 6.8: Performance and error scores of the weighted majority algorithm considering the local changes on both sides of the boundary candidate using measures MI, H, and H_r with varying context. Two reference models, the RM and LM are defined in Section 5.5.

votes. Furthermore, since most of the measures discussed here are asymmetric, their indication in one direction is stronger. For example, one expects TP to give better indications while processing the stream left-to-right, so on the left side of a boundary candidate. On the contrary, TP_r should provide a better indication on the right side. Weighted combination will automatically discover the value of these decisions.

As a result, the last improvement to the boundary discovery algorithm discussed here is to incorporate the local changes on the two different sides of a boundary candidate as separate voters in the weighted majority voting algorithm. Table 6.7 presents these results for varying maximum context size. Except for lexical scores, in comparison to previously presented results the benefit of this may not be immediately clear. It seems the approach trades the precision for increased recall. However, as well as in the increased lexical performance in Table 6.7, the benefit of this more eager segmentation approach will be clearer as other, more varied cues are added.

6.2.5 Reducing redundancy

In Section 6.1 we concluded that even though all the measures we discussed so far measure something relevant to word boundaries, they are different. Further analysis showed, however, that not all of them are different from one another. In particular, they seem to form two groups, TP and MI in one, SV and H in the other. The analysis also indicated that if we pick one of the measures from each group, we do not miss much information. Since this option also simplifies the computational system, it is attractive to pick a subset of the measures instead of all of them.

Table 6.8 presents the result obtained only using MI, H, and H_r. Compared to results in Table 6.7, most scores go down slightly. It seems the SV and TP have minor contributions to the performance even in the presence of MI and H. However, for the sake of simplicity, I will take these measures as representative of predictability for the rest of this thesis.

6.3 Summary and discussion

It is time for the rather long discussion of a set of measures of predictability and predictability-based segmentations strategy in this chapter to be wrapped up and concluded. However, a last simplification will be provided, and the model developed so far will be evaluated further before concluding.

The previous sections presented a set of performance measures according to varying context size that the measures are calculated with. Even though the effect of increasing the maximum context size may be insightful, it is not one of main interests in this thesis. It is clear from the results presented so far that increasing maximum context size increases the precision, and reduces the oversegmentation errors, but it also decreases the recall, and increases the undersegmentation errors. If we were to pick a 'maximum context size' based on the results in Table 6.8, the maximum context size in the range three to six look like well-performing options. However, we should also note that using higher context sizes is likely to cause memorization of the complete words or phrases. If we want more generalization, smaller context sizes are more appropriate. As a result, to simplify the discussion for the rest of this thesis, I will pick the maximum context size of three, and use results obtained using this setting (Table 6.8 row three) as the representative results of the predictability-based segmentation strategy. In the rest of this thesis, the model using three predictability measures, MI, H and H_r with context sizes between one and three will be called *predictability-based segmentation model* (PM). The PM will be used for all further simulations combining results from multiple cues and strategies that will be presented in the next two chapters. The decision for number three is still somewhat arbitrary, and fine tuning this parameter may lead to better performance values. However, for the sake of simplicity, I will stick to this setting for the other models presented in the later chapters as well. The choice of the maximum context size parameter is further discussed in Chapter 9.

Now that we picked a representative model, without space concerns for presenting too many graphs, we can present one more aspect of the model's performance: the change of performance with the increasing input. Following the evaluation strategy described in Section 5.4, Figure 6.10 presents change of f-score and error values for the PM for each 500 utterance block of the BR corpus. For the last 290 utterances, the performance scores are significantly better than the scores calculated for the complete corpus (BF=83.8%, WF=69.8%, LF=60.2%, E_o=13.7%, E_u=3.4%. Cf. scores in Table 6.8 row three).

This chapter described a set of well-known measures of predictability or uncertainty. After a careful analysis of these measures and their combination, I have described a completely unsupervised method to combine multiple measures calculated on varying phoneme-context size. Arguably, we could do with a single measure of predictability.

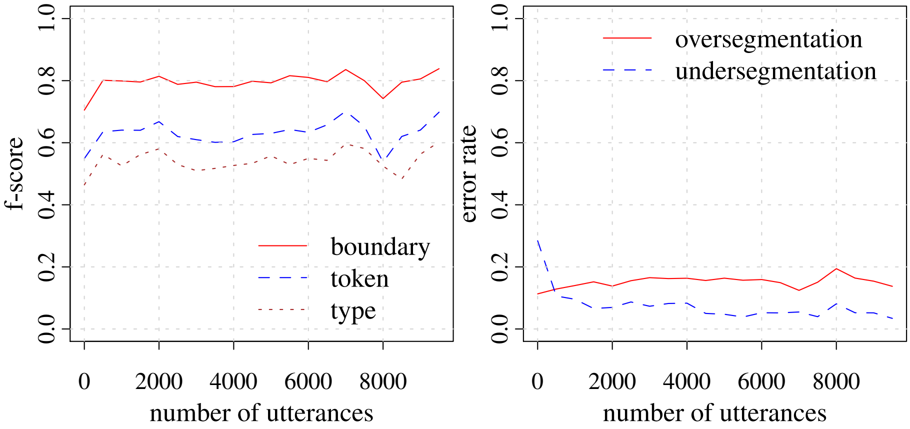


Figure 6.10: (a) Boundary, word token and word type f-scores and (b) oversegmentation and undersegmentation rates of the PM on the BR corpus for successive blocks of 500 utterances each.

The reason for the effort spent for combination of these measures here is twofold. First, it seems none of the measures alone performs as well as the combination of multiple measures. With the interest of getting the most out of predictability cues, it makes sense to combine them. Second, the methods developed here for combination of multiple cues will be used for combining other cues in next two chapters. For the simulations of multiple cue combinations, the algorithms presented in this chapter will set an example.

The strategy defined in this chapter works, and certainly performs well in the segmentation task. This is very clear when we compare the performance of the final model with the baseline model RM. When compared to the performance of the state of the art LM strategy, on the other hand, the performance of the PM is not that impressive. However, it is not too far behind either. The main question of interest in this study is not to achieve best performance, but investigate how multiple cues may be contributing to the performance of a cognitively relevant cue combination model. In this respect, this chapter is only the first step, and coming chapters will continue taking steps in this direction, and increasing the performance of the model along the way.

7 Learning from Utterance Boundaries

> You don't understand anything until you learn it more than one way.
>
> Marvin Minsky

One of the attractive aspects of predictability-based segmentation is that it does not require any lexical knowledge in advance. Most of the other cues discussed in Section 4.2 need at least some lexical knowledge to be useful for discovering lexical unit boundaries. However, certain aspects of lexical unit boundaries, such as the regularities found at the beginning and end of words, can be induced from the boundaries already marked in the input without the need for a lexicon. There are a number of acoustic cues, such as pauses, that are highly correlated with lexical unit boundaries. However, these are generally considered to be unreliable (see the discussion in Section 4.2.4), and they are rarely marked in available corpora. Utterance boundaries, however, are typically marked well during a conversation, and they are marked clearly in all available corpora. Even though the segmentation method that will be presented in that section can be used with boundaries marked by any cues, the results reported here are obtained using only the information collected at the utterance boundaries.

Section 4.2.4 suggested two possible uses for pauses in general which naturally apply to utterance boundaries. First, one can use pauses to restrict the possible segmentations by disregarding possible words that straddle pauses. This use of utterance boundaries for segmentation has been utilized (implicitly) by almost all computational models of segmentation.[1] The predictability-based segmentation model discussed in the previous chapter is no exception. Here, I will investigate the second use: since utterance boundaries are also word boundaries, therefore, by paying attention to the beginnings and ends of the utterances, one can gain some insight into the way words are built—i.e., learn some aspects of the phonotactics of the input language.

[1] A notable exception is presented by Perruchet and Vinter (1998) which simulates the experimental setup of Saffran et al. (1996a) where the stimuli do not have any utterance boundaries.

7.1 Related work

The use of utterance boundaries for lexical segmentation is typical in connectionist models (e.g., Aslin et al., 1996; Christiansen et al., 1998; Stoianov and Nerbonne, 2000). These models try to predict whether there is an utterance boundary after a given input sequence or not. If the model indicates a high likelihood of an utterance boundary where there is none, it is taken as an indication of a word boundary. The non-connectionist models, even the ones that take phonotactics seriously (e.g., Blanchard et al., 2010), do not typically pay attention to utterance boundaries. Two exceptions to this are the models described by Fleck (2008) and Monaghan and Christiansen (2010). Both models make use of pauses and utterance boundaries to learn phonotactics and use them for lexical segmentation.

In segmentation models that use some form of language modeling (such as, Brent, 1999a; Goldwater et al., 2009; Venkataraman, 2001), including the reference model LM, described in Section 5.5, a simple model of phonotactics is used for estimating probabilities of unknown words (Equation 5.4). More elaborate models use a similar phonotactics component but calculate it over higher-level phoneme n-grams (e.g., Blanchard et al., 2010). The probabilities of the phoneme n-grams are estimated from the word tokens (Venkataraman, 2001) or from word types (Blanchard et al., 2010; Brent, 1999a). However, none of these models use utterance boundaries explicitly to infer phonotactics.

As far as I could determine, the only two non-connectionist models that make use of utterance boundaries explicitly are presented by Monaghan and Christiansen (2010) and Fleck (2008). The model presented by Monaghan and Christiansen (2010) learns phonotactics from the lexicon learned in an on-line fashion using a heuristic algorithm. Thus, it is a different source of information than the utterance boundaries, and it is more similar to the word-based segmentation model that will be described in Section 8.2. The model presented by Fleck (2008) is probably the most similar to the segmentation model that will be outlined in this chapter, and it will be described here briefly.

Fleck (2008) presents a batch model that guesses boundaries based on their left and right (phoneme) context. For any boundary position, the model considers its left (l) and right (r) phoneme contexts. The boundary decision is given if $\mathsf{P(b|l,r)} \geq 0.5$, where b denotes boundary. That is, the model decides for a boundary if, given l and r, the probability of observing a boundary is higher than 0.5. Using the Bayesian inversion and assuming independence between l and r,

$$\mathsf{P(b|l,r)} = \frac{\mathsf{P(b)P(l,r|b)}}{\mathsf{P(l,r)}} = \frac{\mathsf{P(b)P(l|b)P(r|b)}}{\mathsf{P(l)P(r)}}. \tag{7.1}$$

As a result, learning in this model can be considered as estimating five probability values $\mathsf{P(b)}$, $\mathsf{P(l|b)}$, $\mathsf{P(r|b)}$, $\mathsf{P(l)}$ and $\mathsf{P(r)}$. Notice that the measure $\mathsf{P(b|l,r)}$ estimated here is the measure presented by Hockema (2006) discussed in Section 6.1. As demonstrated in this section, the measure is rather accurate, and if the estimate is successful, it is expected to lead to high performance.

For each candidate boundary position, Fleck (2008) uses variable length left and right n-grams up to five phonemes so that the selected n-gram is the longest sequence that occurs at least 10 times in the corpus. The segmentation algorithm works in three steps, and requires multiple passes over the corpus. First, the model estimates $\mathsf{P}(\mathtt{l}|\mathtt{b})$ and $\mathsf{P}(\mathtt{r}|\mathtt{b})$ from the utterance boundaries. At the first step, the segmentation decision is given if these probabilities are over a threshold value. At the second phase, a heuristic morphological analyzer fixes certain boundaries decided during the first phase. At the last step, the model estimates the $\mathsf{P}(\mathtt{b})$, and re-estimates $\mathsf{P}(\mathtt{l}|\mathtt{b})$ and $\mathsf{P}(\mathtt{r}|\mathtt{b})$ from the output of the second step, and also estimating $\mathsf{P}(\mathtt{l})$ and $\mathsf{P}(\mathtt{r})$ from the corpus, it uses Equation 7.1 to calculate the estimate of $\mathsf{P}(\mathtt{b}|\mathtt{r},\mathtt{l})$. A segmentation decision is given if this value is greater than 0.5. Fleck (2008) reports relatively good results (BF = 82.9%, WF=70.7%, LF=36.6% on the BR corpus, see Table 7.3 for details).

7.2 Do utterance boundaries provide cues for word boundaries?

From previous studies (e.g., Christiansen et al., 1998; Fleck, 2008), it is clear that utterance boundary information is useful for finding word boundaries. In particular, utterance beginnings and endings can be used for guessing what word beginnings and endings look like. In other words, utterance boundaries help in learning phonotactics. This subsection provides an analysis of the BR corpus to demonstrate the usefulness of the utterance boundaries in finding word boundaries, and introduces a measure that will be used by the unsupervised segmentation model that will be described next.

In Fleck's formulation (Equation 7.1) of the problem, one thing to note is that we can actually estimate $\mathsf{P}(\mathtt{b}|\mathtt{r},\mathtt{l})$ directly from a corpus with pauses. This would probably assume a large number of pauses, but the required number of pauses is not necessarily unrealistic in comparison to the input children receive during language acquisition. Estimating $\mathsf{P}(\mathtt{b}|\mathtt{r},\mathtt{l})$ from the utterance boundaries alone is problematic, since utterance boundaries provide information about $\mathsf{P}(\mathtt{b}|\mathtt{r})$ and $\mathsf{P}(\mathtt{b}|\mathtt{l})$ separately. Nevertheless, estimating the relation of boundaries with left and right context separately is not a bad idea. In essence, what we are interested in is whether $\mathtt{r}$ is a good candidate for a word ending, and whether $\mathtt{l}$ is a good candidate for a word beginning. An accurate estimation of the joint indication of $\mathtt{l}$ and $\mathtt{r}$ together may help, but a word ending is relatively independent of the beginning of the next word, and separate indications of $\mathtt{l}$ and $\mathtt{r}$ as possible word endings and as possible word beginnings, respectively, are useful criteria in identifying boundaries.

To determine whether the utterance beginnings and endings are good estimators of the word beginnings and endings, Figure 7.1 presents results of an analysis that takes a look into the BR corpus. The first two columns in Figure 7.1 present familiar histograms for $\mathsf{P}(\mathtt{ub}|\mathtt{r})$ and $\mathsf{P}(\mathtt{ub}|\mathtt{l})$, respectively, where $\mathtt{ub}$ represents the presence of an utterance boundary. Figure 7.1a presents the distribution of $\mathsf{P}(\mathtt{ub}|\mathtt{r})$ for all

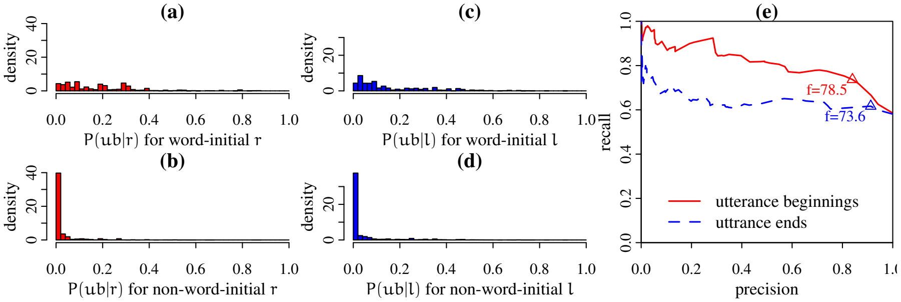


Figure 7.1: For phoneme pairs l and r, distribution of (a) $P(ub|r)$ over word-initial phoneme pairs, (b) $P(ub|r)$ over non-word-initial phoneme pairs, (c) $P(ub|l)$ over word-final phoneme pairs, (d) $P(ub|l)$ over non-word-final phoneme pairs. The precision/recall curves in (e) present the performance of two simple algorithms that decide on a boundary if $P(ub|l)$ or $P(ub|r)$ is higher than a threshold.

word-initial phoneme bigrams that are not utterance-initial, while Figure 7.1b presents the distribution of the same measure plotted for phoneme bigrams that are found in non-word-initial positions. Similarly, Figure 7.1c–d present the distribution of $P(ub|l)$ for the word-final phoneme bigrams that are not utterance final, and non-word-final phoneme bigrams, respectively. The utterance boundaries are excluded while calculating all distributions presented in Figure 7.1.

The histograms in Figure 7.1a–c clearly show that the phoneme pairs that occur at utterance beginnings and utterance ends are more likely to occur at word beginnings and word ends, respectively. Particularly, most of the non-word-initial phoneme bigrams never occur at the beginning of utterances, and most of the non-word-final phoneme bigrams never occur at the end of utterances. As a result, a large portion of the probability mass on the lower histograms is grouped together at the very low end of the scale. In general, the distributions are fairly different for target locations (i.e., word beginnings and ends) and non-target locations.

Figure 7.1e presents boundary precision/recall curves for two simple threshold based segmentation algorithms, one segmenting before a phoneme bigram r, such that $P(ub|r)$ is higher than a threshold, and the other segmenting after a phoneme bigram l such that $P(ub|l)$ is higher than a threshold. When used in this manner, the information from utterance beginnings seems to be slightly more useful compared to utterance endings (this will be discussed further in Section 7.3). In general, the overall performance is rather good, and the best f-scores for both are above 70%. However, as discussed before, an unsupervised learner faces the problem of determining a good threshold, or using another method for an operational definition of how to decide for or against a boundary. The next subsection will address the problem of how to use the information collected at utterance boundaries in an unsupervised learning method. Before that two additional notes are in order.

First, we use conditional probabilities $P(ub|r)$ and $P(ub|l)$ as measures of association between the utterance boundaries, and the phoneme sequences following or preceding them. An arguably better measure would be pointwise mutual information (MI, see Section 6.1.3). Indeed, preliminary experiments using MI (not reported here) resulted in slightly better performance. However, the only major addition MI brings in this particular task is correcting the probability value for frequency of boundaries. Since the frequencies of boundaries are relatively stable compared to other phoneme n-grams, conditional probability performs similarly. The choice of conditional probability here is mainly motivated by the fact that it is a simpler measure. Hence, it makes the exposition here more transparent and easier to follow.

Second, the statistical analysis presented in Figure 7.1 is based on phoneme pairs (or bigrams). Phoneme n-grams of varying length may provide different information. As the length of the n-gram increases, it is more likely to capture whole words, while shorter n-grams are likely to capture regularities shared by many words. The varying size of phoneme n-grams will be one of the points that will be investigated with the unsupervised model that is described next.

7.3 An unsupervised learner using utterance boundaries

The analysis in Section 7.2 showed that $P(ub|r)$ and $P(ub|l)$ are useful measures for deciding for a boundary before r, or after l. In other words, if sequence r occurs consistently at the utterance beginnings, it is likely that it follows word boundaries. Similarly if l occurs consistently at the end of utterances, it is likely that it precedes word boundaries as well. For the rest of this section, we will refer to $P(ub|r)$ as 'UB_b' and $P(ub|l)$ as 'UB_e', signifying that these quantities are utterance boundary cues learned from utterance beginnings and utterance ends respectively. Additionally, where necessary, a superscript indicating the length of relevant phoneme n-gram (l or r) will be used. For example, UB_b^3 is equal to $P(ub|r)$ where r is a phoneme n-gram of length three, i.e., given r, the probability of observing an utterance boundary before r.

Following the conventions suggested in Section 6.2, this section reports results of an unsupervised learner that uses a peak strategy. That is, the learner posits a boundary if the measure (UB_e or UB_b) calculated at the current position is higher than it is in the surrounding candidate boundary positions.

Figure 7.2 plots UB_b^2 and UB_e^2 values for each possible boundary position in the utterance /IzD&t6kIti/ 'is that a kitty'. Like predictability values, the relevant score is calculated both at the beginning and at the end of the utterance. We do not look for boundaries in these positions, however, the values are used for detecting peaks. It should be noted that UB_e is not defined at the beginning of the utterance. Similarly UB_b is not defined at the end of the utterance. As a result, UB_e can measure only right side of the 'peak' after the first phoneme, and UB_b can measure only the left side

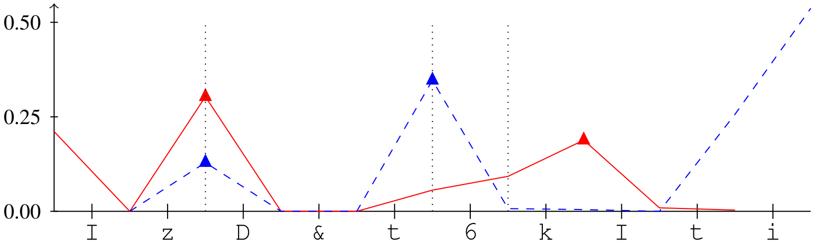


Figure 7.2: UB$_{\mathrm{b}}^{2}$ (solid line) and UB$_{\mathrm{e}}^{2}$ (dashed line) values for example utterance /`Iz D&t 6 kIti`/ ‘ is that a kitty’. The dotted vertical lines mark expected boundary locations. The values are calculated on the BR corpus using bigrams, falling back to unigrams at the edges of the utterance where bigrams are not available.

	boundary			word			lexicon			error	
measure	P	R	F	P	R	F	P	R	F	E_o	E_u
UB$_{\mathrm{b}}^{1}$	61.0	67.9	64.2	46.5	50.2	48.3	15.6	40.8	22.6	16.4	32.1
UB$_{\mathrm{b}}^{2}$	79.5	72.9	76.1	65.1	61.3	63.1	21.0	53.3	30.1	7.1	27.1
UB$_{\mathrm{b}}^{3}$	89.2	63.1	73.9	67.7	53.7	59.9	16.3	49.1	24.5	2.9	36.9
UB$_{\mathrm{e}}^{1}$	58.0	60.5	59.2	44.5	45.9	45.2	16.5	43.4	23.9	16.5	39.5
UB$_{\mathrm{e}}^{2}$	73.9	71.1	72.5	60.1	58.5	59.3	21.6	53.3	30.8	9.5	28.9
UB$_{\mathrm{e}}^{3}$	81.3	61.1	69.7	61.4	50.6	55.5	18.7	55.8	28.0	5.3	38.9
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 7.1: Performance scores for segmentation using utterance beginnings and utterance endings on the BR corpus. ‘UB$_{\mathrm{b}}$’ stands for segmentation using utterance beginnings, and ‘UB$_{\mathrm{e}}$’ stands for segmentation utterance endings. The subscripts indicate the size of the phoneme n-gram which is used to calculate relevant statistics. Two reference models, the RM and LM are defined in Section 5.5.

of the ‘peak’ before the last phoneme. This limits each individual measure’s utility. However, as the performance scores in Table 7.1 indicate, the measures are still quite useful as indications of word boundaries. Furthermore, the final aim of this study is using these measures in combination with each other and in combination with other measures, which will hopefully complement these measures and compensate for their shortcomings. Similar to the predictability scores, there is a tendency to observe peaks in real boundary positions. For example, in Figure 7.2, the UB$_{\mathrm{e}}^{2}$ measure catches two of the boundaries correctly, misses one and gives no false positives. UB$_{\mathrm{b}}^{2}$, on the other hand, finds one of the boundaries correctly, misses two and gives one false positive.

Table 7.1 presents the performance scores of the peak-based segmentation algo-

		1	2	3	4	5	6	7	8
$\mathrm{UB_b}$	novel	8.1	5.7	3.7	3.4	3.3	3.5	3.6	3.7
	$\mathrm{novel}_{>\mathrm{n}}$	100.0	99.0	62.2	24.2	7.7	5.2	4.0	3.8
$\mathrm{UB_e}$	novel	11.7	7.8	3.4	2.6	1.9	1.9	1.9	1.9
	$\mathrm{novel}_{>\mathrm{n}}$	100.0	99.8	59.4	17.9	6.1	3.6	2.4	2.0

Table 7.2: Percentage of novel words discovered by the peak-based algorithm using $\mathrm{UB_b}$ and $\mathrm{UB_e}$ with changing phoneme n-gram lengths. The rows labeled 'novel' present the percentage of correctly discovered words that were not detected at the utterance boundaries before. The rows labeled '$\mathrm{novel}_{>\mathrm{n}}$' consider a word also novel if the n-gram is not novel, but the word is longer than the n-gram length.

rithm, using only information from utterance beginnings and utterance endings. The algorithm is the same as Algorithm 6.1 (on page 105), except instead of predictability scores, the $\mathrm{UB_b}$ and $\mathrm{UB_e}$ scores are used. The segmentation performance is not impressive. However, all performance scores are consistently better than random, and some performance scores are competitive with the language modeling baseline. Intuitively, the different n-gram lengths capture different aspects of utterance boundaries, and combination of n-grams of varying lengths is expected to result in better performance compared to the performance obtained by using each n-gram length alone. The effect of combining varying n-gram length $\mathrm{UB_b}$ and $\mathrm{UB_e}$ scores will be presented shortly. However, there is another interesting question that is easier to investigate in this simple form: is the method learning words, or phonotactics? In other words, can the method discover previously unseen words by using regularities that make up word beginnings and word endings?

Calculating $\mathrm{UB_b}$ and $\mathrm{UB_e}$ scores using a shorter n-gram length is more likely to capture the regularities stemming from phonotactics. On the other hand, higher level n-grams would cause the method to learn frequent lexical items. Of course, both cases are useful for segmentation, but learning phonotactics would allow the method to extract more novel words rather than learning individual words. Table 7.2 presents percentages of correctly identified novel words, for two different definitions of 'novelty'. For the first definition, we count the correctly identified words that were not seen at the utterance boundaries before. For the second definition, we count the correctly identified words that are either not observed at utterance boundaries, or are longer than the phoneme n-gram size in use. The first definition gives an indication of the algorithm's success in discovering completely new words. The second definition is also interesting, because it shows another effect of generalization. Even though most of the words counted according to the second definition are seen before, since they are larger than the n-gram length, there is no way for the model to completely memorize

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	$\mathtt{E_o}$	$\mathtt{E_u}$
UB$_\mathtt{b}$	71.9	77.3	74.5	57.1	60.0	58.5	23.9	50.6	32.4	11.4	22.7
UB$_\mathtt{e}$	73.3	78.6	75.8	57.5	60.4	58.9	22.0	51.1	30.8	10.8	21.4
UM	82.9	84.8	83.8	70.5	71.7	71.1	33.8	66.9	44.9	6.6	15.2
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3
Fleck (2008)	94.6	73.7	82.9	–	–	70.7	–	–	36.6	–	–

Table 7.3: Performance scores for the combination of utterance boundary measures on the BR corpus. The first two rows present performance scores for UB$_\mathtt{b}$ and UB$_\mathtt{e}$ measures, respectively, with n-grams of length one to three combined using the weighted majority voting. The third line, labeled UM, combines both UB$_\mathtt{b}$ and UB$_\mathtt{e}$ using the same method and parameters. Two reference models, the RM and LM are defined in Section 5.5, and the last row reports the performance scores for the related model presented by Fleck (2008).

them. As a result, both definitions indicate whether the algorithm is segmenting words by generalizing how words are built, or by memorizing the frequent words or phrases. The results presented in Table 7.2 demonstrate the expected contrast between short and long phoneme n-grams used in calculation of UB$_\mathtt{b}$ and UB$_\mathtt{e}$. If the algorithm is used with short n-gram lengths, it learns more general aspects of phonotactics. On the other hand, if longer phoneme n-grams are used, the method starts memorizing the frequent words or phrases.

Besides the expected change in generalization by varying phoneme n-gram length, another interesting observation in Table 7.2 is that the UB$_\mathtt{e}$ measure generalizes better, especially for shorter n-gram lengths. This is mostly due to the fact that English is primarily a suffixing language, and the morphemes at the end of the words provide a better opportunity for identifying the novel words which share the previously encountered morphemes. However, this finding is not reflected well by the performance of UB$_\mathtt{e}$-based models presented earlier. For all results presented in Figure 7.1c and Table 7.1, UB$_\mathtt{e}$-based models are outranked by the UB$_\mathtt{b}$-based models. This conflicts with the expectation that if the ends of words allow better opportunities for generalizations, UB$_\mathtt{e}$-based segmentation should be more accurate. The reason for UB$_\mathtt{e}$-based models performing worse than UB$_\mathtt{b}$-based models stems from another fact, at least valid for the BR corpus: utterance ends are more varied compared to utterance beginnings. In the BR corpus, only 579 of 1324 word types appear at the beginning of utterances, while 903 word types appear at the end of utterances.

7.4 Combining measures and cues

The overall aim of the segmentation measures and methods developed here is to be able to combine information from multiple sources. The first question for a segmentation strategy that makes use of utterance boundaries is whether the information from utterance beginnings and ends, as well as the information from varying phoneme n-gram length can be combined for better performance. The second question is whether information from utterance boundaries can be combined with the information obtained using predictability measures presented in Section 6.2. Both questions can be tested using a combination method similar to the one used for predictability measures presented in Section 6.2.2 and 6.2.3.

Table 7.3 presents the performance scores for three models that combine UB_b, UB_e and both, respectively. For all three, the simulations are run using the weighted majority voting algorithm (described in Section 6.2.3) by combining the phoneme n-gram lengths between one and three. For simplicity, UB_b and UB_e without superscripts denote this combination, i.e., the combination of the respective measure with varying the phoneme n-gram length between one and three. Combination of both UB_b and UB_e will be called UM, and it will be used as a representative model for the strategy developed in this chapter.

Table 7.3 also presents the usual reference results, as well as the performance scores for the model presented by Fleck (2008) are also reported in the same table for comparison. Compared to some of the scores presented earlier in Table 7.1, the performance gain by the combination of varying phoneme n-gram length is questionable for individual measures (UB_b and UB_e) for some performance scores. Since the individual measures perform better than random segmentation, the low performance of the combined result is likely to be due to the fact that the voters in these cases are not independent enough. On the other hand, the combination of the UB_b and UB_e yields a clear performance gain. The performance results for UB_b and UB_e combination certainly show an improvement over the individual measures, and the word and lexical scores are better than the results reported by Fleck (2008), only showing slightly worse performance in boundary scores. It should also be noted that, unlike the segmentation model presented in Fleck (2008), these results are obtained using only a single pass through the corpus, with no ad hoc corrections and/or parameters. The only parameter used in this segmentation method is the maximum phoneme n-gram size (set to three in all instances).

To answer the second question, whether the utterance boundary cue is useful in combination with predictability cue or not, Table 7.4 presents results of combining the reference models for each cue, the UM with the PM. The combined model, PUM, combines all measures from both models for phoneme n-gram size one to three using the weighted majority voting algorithm (described in Section 6.2.3). The combination is flat, that is, all measures are considered equal, and no attempt is made to group their results. The usual baseline scores, and the performance scores of the predictability and the utterance boundary-based segmentation algorithms with the same phoneme context range are repeated for comparison. Combining the utterance boundary and the predictability measures results in better performance scores than the combinations utilizing each group of measures alone. Furthermore, the combined model performs

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
PM	69.6	92.5	79.5	56.9	70.2	62.9	36.7	49.8	42.3	15.3	7.5
UM	82.9	84.8	83.8	70.5	71.7	71.1	33.8	66.9	44.9	6.6	15.2
PUM	81.3	87.1	84.1	69.2	72.7	70.9	36.9	65.8	47.3	7.6	12.9
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3
Fleck (2008)	94.6	73.7	82.9	–	–	70.7	–	–	36.6	–	–

Table 7.4: Performance and error scores for combination of the models UM and PM (PUM) on the BR corpus. The first two rows display previously presented results for the models PM and UM. Two reference models, the RM and LM, and the performance scores for the related model by Fleck (2008) are also repeated here for ease of comparison.

better than Fleck (2008) and the reference model LM for most scores. This comparison may not be fair because the PUM uses information from both predictability and utterance boundaries. However, it is also not fair, because the model presented by Fleck (2008) is a batch model which makes use of multiple passes over the data. The relatively fair comparison would be comparing the final state of the UM with the results reported by Fleck (2008). Figure 7.3 present the progression of performance of UB and PUB during the learning as more input is processed.

Figure 7.3 demonstrates that both the UM and the PUM perform better at the later stages of learning, and the combined model's performance is clearly better than th UM's performance. Additionally, at the end of the BR corpus, the UM's performance is substantially better (BF=88.3%, WF=77.7%, LF=72.1%, E_o=7.3%, E_u=6.7%), in comparison to the scores presented by Fleck (2008, BF=82.9%, WF=70.7%, and LF=36.6%). Compared to the LM, on the other hand, the lexical scores are still rather low, to which we will return in Chapter 8.

7.5 Summary

This section described another segmentation strategy that does not require knowledge of lexical items, and as a result is suitable for bootstrapping the lexicon. The strategy presented here makes use of utterance boundaries, i.e., learning common beginnings and endings of utterances. Since utterance boundaries are also lexical unit boundaries, this information is useful for lexical segmentation in two ways. First, learning the beginnings and the ends of words that appear at utterance boundaries is useful for spotting them later when they appear at utterance-internal positions. Second, and more interestingly, words beginnings and ends show some regularities, and learning these regularities from already existing boundaries helps in discovering novel words that conform to these regularities even if they are encountered at utterance-internal

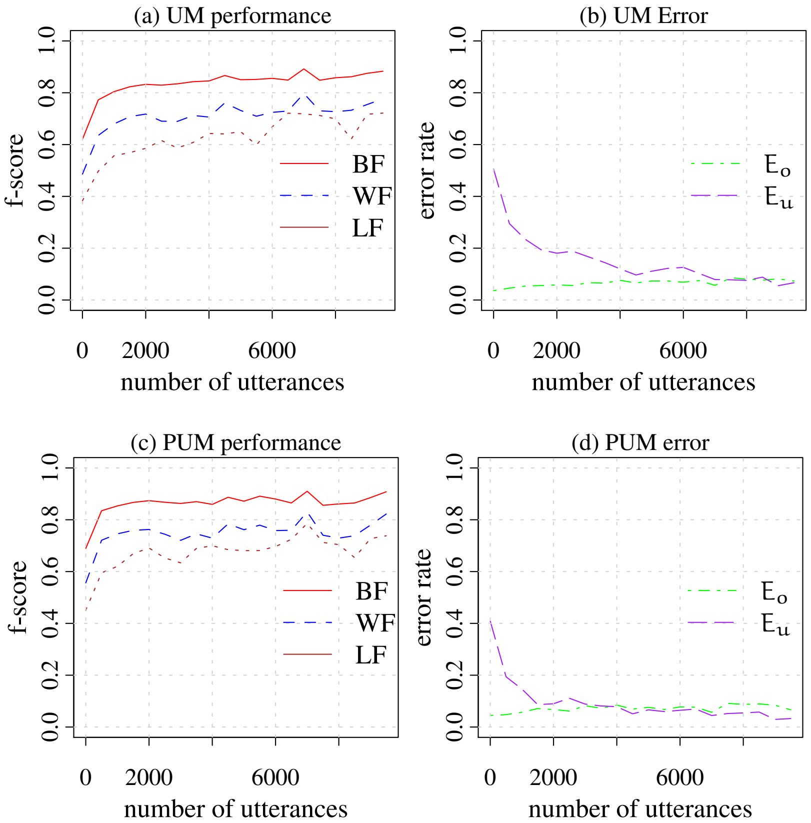


Figure 7.3: Boundary (BF), word token (WF), and word type (LF) f-scores and oversegmentation (E_o) and undersegmentation (E_u) error rates for the phonotactics model based on utterance boundaries, the UM, and its combination with previous the predictability-based model (PUM). The scores are calculated for each 500-utterance block in the BR corpus during the learning process.

positions.

This section first provided some evidence through a statistical analysis of the BR corpus showing that utterance boundaries provide information useful for spotting word boundaries. Second, an unsupervised segmentation algorithm that combines the information gathered at the utterance beginnings and utterance ends has been developed. The performance results obtained using utterance boundaries (Table 7.1 and Table 7.3) are encouraging, and competitive with the other models presented in the literature. More importantly, the segmentation based on utterance boundaries and the predictability-based segmentation strategy described in Section 6.2 seem to complement each other. The combination of the two leads to better performance than each individual cue alone.

The results presented in Table 7.2 showed that the unsupervised algorithm presented in this section not only memorizes words and/or word beginnings and endings, but also gains some knowledge of phonotactics. As a result, it is also useful for discovering words that have not been observed at the utterance boundaries before. This generalization was possible even though the input representation treats phonemes as unrelated symbols that have no relation to each other. It is natural to expect a better generalization if a representation based on phonetic features were to be used (e.g., as in Christiansen et al., 1998). For example, a richer input representation would be sensitive to phonemic tendencies in the data. For example, a vowel–vowel sequence, e.g., /ae/, is less likely in English than a consonant-vowel sequence, /ba/, even if the bigrams in question have never been seen by the learner before. The investigation of effects of a richer input representation is a possible direction for future research.

The method described in this chapter only learns phonotactics from utterance boundaries. In principle, more can be learned from utterance boundaries. A relevant cue that may be bootstrapped from utterance boundaries is lexical stress. At least for the languages with word-initial of word-final lexical stress, it may be possible to learn these patterns from utterance boundaries. However, preliminary experiments (not reported here) with learning stress from utterance boundaries were not fruitful. This will be discussed further in Section 8.3 in detail, where a model of segmentation using lexical stress learned from previously discovered words will be presented.

The segmentation strategies based on predictability and utterance boundaries discussed so far have the advantage that they do not require prior lexical knowledge for discovering words or phrases. This is useful since it provides a way to bootstrap the process of learning lexical items. However, being ignorant about the lexicon is also a disadvantage. This is clearly visible in the difference in their boundary- and lexical-performance scores (e.g., in Table 7.4). Even though the models discussed so far are good at spotting boundaries and words, they do not show the same proportional success rates in lexical scores. The next chapter will be dealing with this issue, introducing an explicit lexicon, and using the information in the lexicon for the advantage of an incremental segmentation algorithm.

8 Learning From Past Decisions

> Those who do not remember their past are condemned to repeat their mistakes.
>
> George Santayana

We learn best from our mistakes, provided that we are aware of the fact that we made a mistake. The formal studies reviewed in Chapter 3 also confirm this common sense statement. Strong learnability results in general settings are possible only with negative evidence. In Chapter 3, I argued that for most real-world learning problems, where the distribution of the input data is constrained, negative evidence is not strictly necessary. That is, if we always observe positive examples of what we try to learn and if we know that they are accurate (to some degree), learning is still possible.

The models of segmentation discussed in this thesis fit in neither of these schemes. The learner does not know whether the segmentation he/she decided on is correct or not. In the real world, mistakes in segmentation would eventually be translated to communication failures or processing inefficiencies which may provide feedback to the learner. However, the models discussed in this thesis do not have access to this type of feedback. Instead, the models presented so far act upon general guiding principles for detecting word boundaries. In Chapter 6 the principle was 'predictability within the lexical units is high, predictability between the lexical units is low'. The models in Chapter 7 utilized the fact that the utterance boundaries are also word boundaries, and words share common beginnings and endings. These models learn statistical relationships between phoneme sequences, and they learn to weigh the individual boundary indicators. However, these models do not learn from negative or positive boundary labels in the input.

This chapter presents models that learn from positive examples. Since the input does not contain any labels (the boundaries), we will take earlier boundary decisions given as positive examples. The learners in this chapter will use the boundaries decided in the past as a source of information, in order to contribute to the decisions in the future. At first sight, this may look circular. However, this works because the data is structured, and we have relatively accurate indications (predictability and utterance boundaries). The type of learning investigated here is closely related to many

IzD*EniTINElsD&tyuwanttubrINApIntuD6h9c*

is	there	anything			else	that	you	want	to	bring	up	into		the	highchair	
is	there	anything			else	that	you	want	to	bring	up	into		the	high	chair
is	there	anything			else	that	you	want	to	bring	up	in	to	the	highchair	
is	there	anything			else	that	you	want	to	bring	up	in	to	the	high	chair
is	there	any		thing	else	that	you	want	to	bring	up	into		the	highchair	
is	there	any		thing	else	that	you	want	to	bring	up	into		the	high	chair
is	there	any		thing	else	that	you	want	to	bring	up	in	to	the	highchair	
is	there	any		thing	else	that	you	want	to	bring	up	in	to	the	high	chair
is	there	N	E	thing	else	that	you	want	to	bring	up	into		the	highchair	
is	there	N	E	thing	else	that	you	want	to	bring	up	into		the	high	chair
is	there	N	E	thing	else	that	you	want	to	bring	up	in	to	the	highchair	
is	there	N	E	thing	else	that	you	want	to	bring	up	in	to	the	high	chair

Figure 8.1: One of the most ambiguous utterances according to 'gold standard segmentation' of the BR corpus. 'N' and 'E' are letters in the alphabet.

bootstrapping proposals in the language acquisition literature.

This chapter presents two strategies, or cues, that make use of previously discovered word boundaries. The first model makes use of frequencies and forms of previously discovered words. Section 8.1 informally discusses the use of already known words for segmentation in detail. Section 8.2 describes and presents results of simulations carried out by a segmentation model that uses information from previously discovered words. The second model, described in Section 8.3, utilizes lexical stress for segmentation. Section 8.4 provides a summary of findings presented in this chapter.

8.1 Using previously discovered words

In the ideal case where the listener knows all the words in the input, the preferred segmentation of an utterance is the one that spans the complete utterance by a non-overlapping sequence of words in the lexicon. For example, assuming that the lexicon contains the words *is*, *that*, *a*, and *kitty*; and the utterance *isthatakitty* is heard, then *is that a kitty* is such a segmentation. For a large number of utterances, the segmentation using the words in the lexicon is not as clear-cut as in this example. The word forms in the lexicon often suggest multiple segmentations, which makes the segmentation problem non-trivial even in this ideal scenario.[1] Figure 8.1 presents all 12 segmentations suggested for one of the most ambiguous utterances in the BR corpus, *is there anything else that you want to bring up into the high chair*.

Most of the ambiguous segmentations stem from embedded words. For example the word /&t/ 'at' is part of words /b&t/ 'bat' or /h&t/ 'hat'. Furthermore, some words can be segmented completely using other words. For example, the word /bIgIn/ 'begin' can be segmented as /bIg In/ 'big in', and /6go/ 'ago' can be segmented as /6 go/

[1] This is true despite the fact that use of a phonemically transcribed corpus reduces the difficulties that stem from the variability in the speech signal. See Section 4.1 for a discussion of the broader problem, and Section 5.1 for a discussion of the input representation.

'a go'. As well as the embedded words, there are other cases where utterances can be segmented in multiple ways. Fore example, /`Itsnoz`/ can be segmented as /`Its noz`/ 'its nose' or /`It snoz`/ 'it snows'. Similarly, the utterance segment /`&nd9v`/ which can be segmented as /`&nd 9v`/ 'and I've' or /`&n d9v`/ 'Anne dive'. Of 1320 word types in the BR corpus, 32% of the words are parts of other words, and 15% of the words can completely be segmented by two or more other words.

Even though the lexicon is not enough to solve the segmentation problem completely, its apparent utility is clear. Constraining the boundaries to the locations that allow complete non-overlapping sequences formed by the words in the lexicon allows relatively good segmentation. If we take the word types in the gold standard segmentation of the BR corpus, this strategy assigns 84% of the utterances in the corpus a single correct segmentation. If we pick a segmentation at random for the ambiguous 16%, we can achieve f-scores of 99.0%, 98.4% and 98.8% for boundaries, words and lexical units respectively. These scores are obtained using only the 1324 word types that are used in the gold standard segmentation of the BR corpus. With increasing lexicon size and the longer utterance lengths, the ambiguous segmentations are expected to increase, and segmentation performance is likely to go down. Nevertheless, it is still possible to achieve very good segmentation performance if a comprehensive lexicon is available.

While words in the lexicon are a great help for segmentation if one has a complete and accurate lexicon, the story is different for a learner. The learner neither knows all the words in the input nor can he/she be as confident about his/her knowledge of words as a competent language user. However, this does not mean that a partial lexicon is useless for segmentation. Even a small repository of words can be useful for learning new words. This will be demonstrated through a simple strategy that uses previously discovered lexical units to extract novel lexical units from the input utterance. This strategy segments the utterance at locations that start and end at known lexical units. The sequences of phonemes between the boundaries but not in the lexicon are identified as new words and added to the lexicon. An informal demonstration of this strategy is given in Figure 8.2. The example in Figure 8.2 starts with an empty lexicon. Since there is no matching part in the current lexicon, the first utterance cannot be segmented, and it is analyzed as a single word and inserted into the lexicon. This is consistent with the findings that adults and children tend to take whole utterances as lexical units when they cannot segment them (Bannard and Matthews, 2008; Dahan and Brent, 1999). In the second step, the segment *akitty* matches an already known lexical unit. The input is segmented as *thats akitty*, and the newly discovered segment, *thats*, is added to the lexicon. Utterance in step three, *kitty*, cannot be segmented, and inserted into the lexicon as it is. In step four, we use lexical item *kitty* to segment the utterance as *that-kitty*. The new word *that* is inserted into the lexicon, and we increase the frequency of the known word *kitty*. In the last step, using known words *is*, *that*, *a* and *kitty*, we segment the input utterance as *is that a kitty*.

This segmentation strategy is admittedly naive, and it fails on real world input. However, this strategy and its weaknesses provide insights into many successful

step	input	output	lexicon
0			{}
1	*akitty*	*akitty*	{*akitty*(1)}
2	*thatsakitty*	*thats akitty*	{*akitty*(2), *thats*(1)}
3	*kitty*	*kitty*	{*akitty*(2), *thats*(1), *kitty*(1)}
4	*thatkitty*	*that kitty*	{*akitty*(2), *thats*(1), *kitty*(2), *that*(1)}
5	*isthatakitty*	*is that a kitty*	{*akitty*(2), *thats*(1), *kitty*(3), *that*(2), *a*(1), *is*(1)}

Figure 8.2: An informal example of segmentation using already discovered lexical units. The numbers in the parentheses are the number of times the lexical unit is used.

computational models of segmentation, and they deserve some more elaboration here. The first problem we encounter is related to the reliability of the lexical units in the lexicon. For example, in step five of the example presented in Figure 8.2, the decision of segmenting the utterance as *is that a kitty* is arbitrary. The lexicon equally supports the segmentation *is that akitty*. We can remedy this problem by preferring words formed by shorter sequences of phonemes (e.g., *a and kitty*) instead of longer sequences (e.g., *akitty*). However, this leads us to the second problem. In this example, after discovering the word *a*, the naive segmentation strategy has no reason for not segmenting *that* as *th a t*.

On one hand, we want to favor eager segmentation, since otherwise the lexicon would contain whole utterances or phrases like *akitty* in this example. On the other hand, we do not want to oversegment, as in segmenting *that* as *th a t*. These two apparently conflicting problems have a common solution: we need to define what makes a sequence of phonemes a good lexical unit.

8.1.1 What makes a good word?

There are two properties of sequences of phonemes that make them good candidate words. First, the *usage* of the sequence in the language, and in communicative context is an important property of words. There are many clues that usage of a sequence can indicate its 'wordhood'. For example, hearing a certain sequence consistently in the presence of an object is a clue that this sequence may be the name of the object. In this section, we will consider only one usage-related property of a word that is directly observable from a transcribed corpus: its frequency. Returning to the example above, the decision to segment the utterance *akitty* as *a kitty* can be based on the observation that in comparison with *akitty*, both *a* and *kitty* are more frequent in the input.

The second property of a sequence that is helpful for identifying a sequence of phonemes as a word is their *form*. In natural languages, words are typically formed according to certain regularities. Observing these regularities, *phonotactics*, on already known words help identifying new words. For example, one can reject the segmentation *th a t* in the above example by observing that words typically contain a vowel, and the sequences *th* and *t* are not good candidates because they do not contain one. As already introduced in Chapter 7, the models in this thesis use a rather restricted model of phonotactics that is based on regular phoneme sequences at the beginnings and ends of the words. The phonotactics component in Chapter 7 used utterance boundaries to extract information about word boundaries. Having an inventory of lexical items makes these generalizations more direct.

8.1.2 Related work

Most of the state of the art computational models of segmentation use (relative) frequency of a sequence of phonemes as the main indication that the sequence is a good word candidate. The forms of the words are typically used only if the sequence in question is not already in the lexicon. For example, the segmentation models that are based on the language modeling strategy (e.g., Brent, 1999a; Goldwater et al., 2009; Venkataraman, 2001) use a word's relative frequency as its probability if the word is already known. Hence, they favor frequently occurring sequences of phonemes as words. In these models, the form of the candidate word is evaluated only if the candidate word is not already in the lexicon. In most of these models, the word-form component serves for preventing oversegmentation by assigning lower probabilities to combinations of short words in comparison to using long sequences of phonemes as a single word (see Section 5.5 for discussion of this type of modeling). As a result, even though some improvements are possible by modeling phonotactics more carefully (e.g., Blanchard et al., 2010), it is difficult to get much improvement in this setting. For example, as Goldwater et al. (2009) demonstrate, even an idealized phonotactic component that knows the words in the corpus does not improve the performance compellingly compared to the assumption that all phonemes are uniformly distributed.

A segmentation model that primarily uses the frequency of previously discovered words, but does not strictly fit into the language modeling framework, was presented by Monaghan and Christiansen (2010). This model has many similarities to the model I will describe in Section 8.2. The segmentation algorithm in this model follows a similar strategy to the segmentation strategy informally described above (Figure 8.2). It starts searching sequences of phonemes that are already in the lexicon from the beginning of the utterance. The words in the lexicon are matched against the initial substring starting from the current location based on their frequency. When there are multiple alternative segmentations, this gives high frequency sequences an advantage. When a match is found, the part of the utterance that did not match any lexical item is considered a new word, only if the first and last two phonemes of the sequence have been observed at the beginning or at the end of previously accepted words, respectively.

There are studies that model either the usage or the form of the words more carefully. For example Goldwater et al. (2009) present a model that takes frequencies of higher level word n-grams into account. Attempts at modeling phonotactics more carefully include using higher level phoneme n-grams (Blanchard et al., 2010), or taking into account common beginnings and ends of the words explicitly (Monaghan and Christiansen, 2010). However, in almost all of the current computational models of segmentation frequency is the main measure of a word's usage, and phonotactics is mainly used as a way to prevent oversegmentation.

8.2 Using known words for segmentation: a computational model

It is clear that the words we already know do help in discovering new words. However, the problem is not trivial, especially, in the absence of a complete and reliable lexicon. The discussion so far in this chapter identified two properties of words that are useful for deciding whether a string of phonemes is a good word or not. First, a good word is used over and over again in different contexts. Second, word formation follows certain regularities, and does not result in arbitrary sequences. Both properties are exploited by the computational models in the literature to varying degree, and have proven to be useful.

The segmentation strategies we discussed in Chapter 6 and Chapter 7 already make use of these properties to some extent. The predictability strategy implicitly makes use of the fact that words consist of frequent phoneme sequences by requiring word-internal phoneme sequences to be highly associated with each other. The requirement that phoneme sequences across boundaries are unpredictable, is equivalent to the idea that a word needs to occur in many contexts. However, in both cases, we do not make use of the boundaries, and hence the words, discovered so far.

Except the model by Monaghan and Christiansen (2010), the related models in the literature reviewed above follow the language modeling strategy (discussed in Section 5.5 in detail). In these models, the probability of a segmentation is the product of the probabilities estimated from relative frequencies of the words used by that segmentation. The problem then becomes searching for the best segmentation in an efficient way. The segmentation models presented in this thesis follow an eager 'boundary guessing' strategy, evaluating all boundary candidates locally without reference to other boundaries. The model I will describe in this section will also follow the same strategy. In brief, the model assigns higher boundary scores to the positions that are preceded and followed by previously detected words.

8.2.1 The computational model

The computational model, which I will call *word-based model* (WM for short), uses a segmentation algorithm similar to the informally described example in Figure 8.2. It tries to segment given utterances using already known words, and when segmentation is not possible, it inserts the complete utterance into the lexicon.

To decide whether an input utterance should be segmented at a certain position, the model tries to maximize the use of strings that are word-like on both sides. As discussed above, the properties we seek in word-like units are that they are frequent, and they share some features with already known words. In our usual majority voting

		i		s		t		h		a		t		a		k		i		t		t		y	
WF_e			0		0		0		0		0		1		0		0		0		0		0		2
WF_b	0		0		1		0		0		0		2		2		0		0		0		0		
WP_e			0		1		0		0		0		1		0		0		0		1		0		2
WP_b	0		0		2		0		1		0		1		1		0		2		2		1		

Figure 8.3: Lexicon measures for example utterance *is that a kitty*. The measures are calculated assuming that the lexicon contains the words {*akitty*(2), *thats*(1), *kitty*(2), *that*(1)} prior to the input utterance. The numbers in the parentheses are the frequencies of the words.

framework, these form two measures. The first measure is *word frequency* (WF), simply the frequencies of the already known words beginning or ending at the position in question. I will call the second measure *word phonotactics* (WP). The WP is based on the number of times the phoneme sequences surrounding the boundary found at the beginnings or ends of the previously discovered words. It is essentially the same measure as discussed in Chapter 7, except it is calculated using already known word types instead of all utterance boundaries.

Similar to other asymmetric measures discussed previously, we have two flavors for each measure. One indicating the existence of words to the right of the boundary candidate (words beginning at the boundary), and the other indicating the existence of words the left of the boundary candidate (word ending at the boundary). When we need to distinguish them, the measures indicating words ending at the boundary will be suffixed with an 'e' (e.g., WF_e), and the measures indicating words beginning at the boundary will be suffixed with a 'b' (e.g., WF_b).

Figure 8.3 presents the measures calculated for the fifth step of the example in Figure 8.2. For the sake of demonstration, this example uses the orthographic form of the words. The experiments reported below are, as before, run on a phonemically transcribed corpus of child directed speech. For a demonstration of the calculation of the measures, consider the position after the segment *isthata*. There is no matching word that ends at that position (*a* is not a known word), hence, WF_e is zero. WF_b, on the other hand, is two since the frequency of the only word that begins at that position, *kitty*, is two. The phonotactics scores are zero and one respectively, since there is no word in the lexicon that ends with an '*a*', and there is only one word, *kitty*, in the lexicon that starts with a '*k*'. As this simple example demonstrates, there is a tendency for higher values where we would expect a boundary. Before presenting results on real child-directed speech, a more precise description of these measures is in order.

The WF measures are the frequencies of the words ending and beginning in the position to be evaluated. When there are multiple words that begin or end at the position, we take the sum of the frequencies of each word. There are other possible ways to combine frequencies of multiple words. For example, as in the segmentation model described in Monaghan and Christiansen (2010), one can use the frequency of the most frequent word. Another possibility is to use the average frequency of the words. These two alternatives perform similarly, but slightly worse than the summation method. The reason sum works better in this setting has to do with the undersegmentation caused by inserting whole utterances or phrases into the lexicon. For example, with a lexicon that contains lexical units *a* and *akitty*, summation method reflects the role of *a* in both lexical units, while other methods discounts one completely or averages out the effect of both. I will only report results obtained by using summation method.

The WP measure does not require much elaboration either, since it has already been discussed in Chapter 7 at length. The WP is either $P(l|wb)$ or $P(r|wb)$ for phoneme sequences l and r, where wb denotes 'word boundary'. The major difference here is that these conditional probabilities are calculated using the boundaries of word *types*, while in Chapter 7 it was calculated on partial word *tokens* that are observed at the utterance boundaries. As before, we can vary the length of the phoneme n-gram that we use for calculating this measure. For the results reported below, we use the combination of n-gram sizes one to three.

8.2.2 Performance

Given the set of measures described above; the peak-based segmentation strategy (described in Section 6.2.1) for detecting boundaries; and the weighted majority voting algorithm (described in Section 6.2.3) to combine decisions, we are ready for presenting the performance scores. Figure 8.1 presents the usual performance scores for individual measure and the performance scores obtained by combining them using the peak criterion and the weighted majority voting algorithm on the BR corpus. The row before the baseline performance scores, marked as WM (for '*word model*') is the combination of all, and will be used as the representative example of the strategy presented in this section.

All models perform better than random, and combinations consistently give better results. The reason for presenting the models that make use of only one side of the boundary (the models with subscripts) in Table 8.1 is to give a sense of how a strictly left-to-right model performs. Some models in the literature, especially the connectionist models, use only the past information. At first sight, this may seem to be a better choice for modeling human language segmentation. However, for the rest of this chapter, the combined model will be used on the grounds that people make use of information from both sides while processing the utterances (see Section 6.1.7, for a discussion).

I started this chapter by stating that the strategies presented in this chapter learn from previously discovered words. However, the scores presented in Table 8.1 are only based on the measures introduced here. It may not be obvious, at first sight, where the words needed for using these measures come from. The answer is implicit in the general strategy used throughout this thesis: when the learner cannot segment an utterance, it takes it as a word. Hence, the models listed in Table 8.1 are bootstrapped from utterance boundaries. However, having previously introduced strategies, we are not limited only to utterance boundaries. If we combine the model with the models introduced previously, we expect it to provide an improvement. And the performance

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
WF$_e$	48.6	60.6	53.9	32.2	37.8	34.7	14.4	38.8	21.0	24.2	39.4
WF$_b$	78.4	67.7	72.7	61.0	55.1	57.9	17.4	48.0	25.5	7.0	32.3
WF	77.5	71.3	74.3	60.6	57.2	58.9	18.3	47.7	26.4	7.8	28.7
WP$_e$	65.3	67.6	66.4	49.5	50.8	50.1	18.7	47.7	26.8	13.6	32.4
WP$_b$	76.2	70.3	73.1	60.3	57.0	58.6	18.9	51.2	27.6	8.3	29.7
WP	75.2	80.3	77.7	61.4	64.3	62.8	25.4	55.6	34.9	10.0	19.7
WM$_e$	62.7	73.8	67.8	45.2	50.8	47.8	19.5	46.4	27.5	16.6	26.2
WM$_b$	76.4	73.5	74.9	61.0	59.3	60.1	19.0	47.5	27.1	8.6	26.5
WM	77.4	86.0	81.5	65.1	70.2	67.6	30.6	57.6	40.0	9.5	14.0
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 8.1: Performance scores for models that use already known words. WF stands for word frequency, WP stands for phonotactics. WM is the combination of both. The models with subscript e only make use of words that end at the boundary, and the models with subscript b make use of the words that begin at the boundary. Two reference models 'RM' and 'LM' are the same as in Table 6.4.

scores presented in Table 8.2 confirm this expectation. The upper block in Table 8.2 repeats the results presented in Chapter 6 and Chapter 7. The middle row is the combination of all strategies presented so far, followed by repeated baseline results for comparison. The combination provides a consistent improvement on all performance measures. The combined (PUWM) model performs better than the reference LM model on all scores, except boundary precision and lexical precision, and makes slightly more oversegmentation mistakes.

As before, the results in Table 8.1 and Table 8.2 are calculated for the complete corpus. To demonstrate the development of the learner with increasing input, as well as an indication of the final state of the learner, Figure 8.4a presents the performance scores for word-based model (WM), and Figure 8.4b presents the performance of the combined model (PUWM) for each successive 500 utterances. Towards the end of the corpus, the f-score measures PF, WF and LF for word-based model are over 80%, 70%, 60%, respectively. The oversegmentation error is close to 10%, and undersegmentation error is close to 8%. The combined model, on the other hand, achieves over 90.9%, 81.9% and 75.6% for BF, WF, and LF, respectively, and oversegmentation error drops to 7.2%, and undersegmentation error drops to 2.0%. Like the general scores, these scores are an improvement over combined model (PUM) presented in Chapter 7.

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
WM	77.4	86.0	81.5	65.1	70.2	67.6	30.6	57.6	40.0	9.5	14.0
PM	87.9	77.4	82.3	74.2	67.9	70.9	28.0	65.6	39.3	4.0	22.6
UM	82.9	84.8	83.8	70.6	71.7	71.1	33.7	67.0	44.9	6.6	15.2
PUM	82.6	90.7	86.5	72.4	77.4	74.8	42.8	65.3	51.7	7.2	9.3
PUWM	83.7	91.2	87.3	74.1	78.8	76.4	43.9	67.7	53.2	6.7	8.8
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 8.2: Performance scores for combination of WM with the models discussed previously, predictability (PM) in Chapter 6 and utterance boundaries (UM) in Chapter 7. PUM denotes the combination of PM and UM, and PUWM is the combination of all cues discussed so far.

8.3 Making use of lexical stress

As discussed in Section 4.2.2, lexical stress is one of the cues for segmentation that is well supported by psycholinguistic research (e.g., Cutler and Butterfield, 1992; Jusczyk, 1999; Jusczyk et al., 1999b). Lexical stress is used in many languages for marking the prominent syllable in a word. As a result, in combination with the regularities regarding the location of the stress, it is useful for deciding where the word boundaries are. However, as is also noted by Christiansen et al. (1998), what we perceive as lexical stress is a combination of multiple acoustic and physical cues including amplitude, duration, segmental quality and pitch contours. Furthermore, the features that correlate with lexical stress have different functions in different languages. For example in tone languages (such as Chinese), the difference in pitch accent indicates meaning differences rather than marking the prominent syllable in the word. Hence, the notion of lexical stress itself has to be learned. Nevertheless, once the learner is aware or the notion of stress, it can be useful for segmenting words.

Despite the prominence of stress as a cue for segmentation, there are relatively few computational studies that incorporate this cue. There are a number of reasons for neglecting stress in computational models of segmentation. First, there is no neutral way of including the stress cue in the most popular and successful segmentation strategy, the language modeling strategy, in the computational modeling literature. Second, most computational studies use phonemes as the basic unit in the input. Since stress is strongly related to the unit of syllable, it is generally difficult, or unnatural to mark phonemes as having levels of stress or not. Last, and most importantly, currently we do not have child directed speech corpora that mark stress realistically and reliably. As a result, computational methods that study use of stress for segmentation tend to make relatively unnatural assumptions regarding stress assignment.

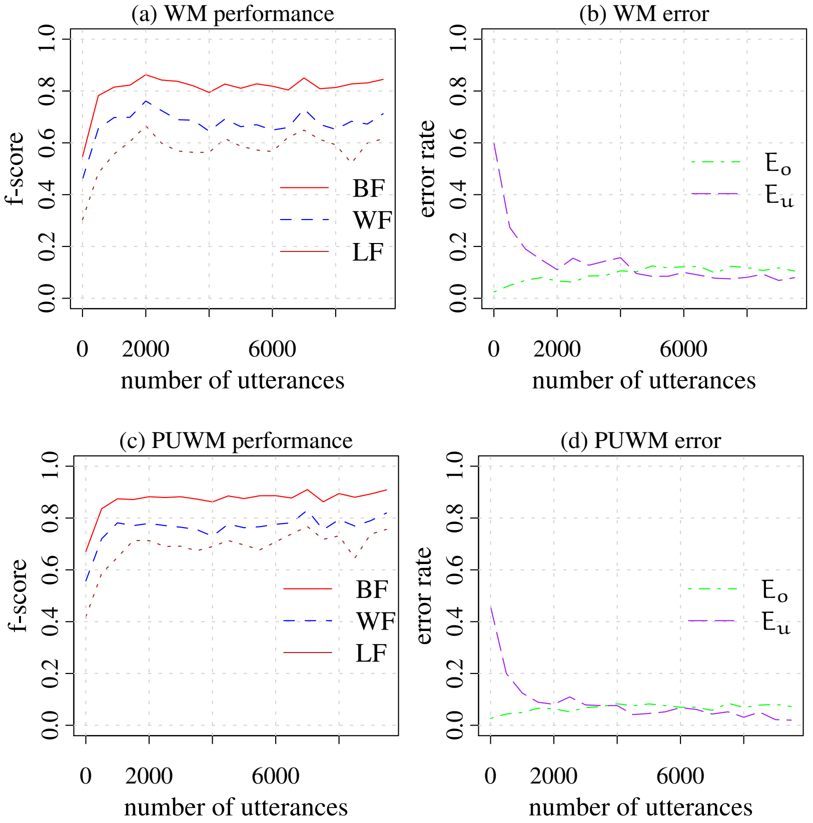


Figure 8.4: Boundary (BF), word token (WF), and word type (LF) f-scores and oversegmentation (E_o) and undersegmentation (E_u) error rates for the word-based segmentation model, the WM, and its combination with previous strategies, the PUWM. The scores are calculated for each 500-utterance block in the BR corpus during the learning process.

8.3.1 Related work

A well-known computation model of segmentation that does incorporate stress as a cue was presented by Christiansen, Allen and Seidenberg (1998). This model is a multiple-cue integration model using a connectionist architecture, more specifically, a simple recurrent network (SRN, see Section 2.2.2 for a description). The model is based on earlier model presented by Allen and Christiansen (1996), and used with modifications in later models (e.g., Christiansen, Conway and Curtin, 2005; Rytting, Brew and Fosler-Lussier, 2010). In the Christiansen et al. (1998) study, the input to the SRN was a set of 11 phonological features for each phoneme; an indication of whether the phoneme precedes an utterance boundary or not; and two additional features representing whether the phoneme belongs to a syllable with primary or secondary stress. Christiansen et al. (1998) tested the model on the Korman corpus (Korman, 1984), by assigning canonical stress patterns found in the MRC psycholinguistic database. Since the MRC database does not mark stress on mono-syllabic words, they marked all monosyllabic words as having primary stress. The overall performance of the model is rather low, 71% BF and 43% WF after an initial training period (the detailed scores are presented in Table 5.4 in comparison with the other segmentation studies in the literature). However, this study showed that adding stress as a cue improves the performance of the SRN. Crucially, similar to the use of stress in this section, Christiansen et al. (1998) stress that a learner may learn how to use lexical stress for segmentation from the boundaries found by the other cues.

Swingley (2005) presents a word discovery procedure based on mutual information and frequency. In this work, lexical stress is not used as a criterion for word discovery. However, a careful analysis of stress patterns of the bisyllabic words found by the discovery procedure was presented. The results, in summary, show that the preference of the trochaic (strong–weak) stress pattern for bisyllabic words may emerge from a word-clustering method based on mutual information and frequency. An interesting finding is that, to learn the trochaic bias, one needs to pay attention to the stress pattern in the lexicon rather than stress pattern found in the corpus. For English Swingley (2005) found that only 14.2% of the bisyllabic word tokens (counted in the corpus) was trochaic, while 19.7% of them were iambic (weak–strong). Counting the pattern in word types (in the lexicon) reverses this balance: 21.2% trochaic, and 6.4% iambic. In both cases, the dominant stress pattern was strong–strong. Swingley also used the Korman corpus, but, with a more careful stress assignment. Rather than assuming that all monosyllabic words have primary stress, they are assigned weak or strong stress depending on their primary use. A word is still assigned a single stress pattern for all of its occurrences in the corpus. However, this method of marking stress certainly provides an improvement over the stress-marking by Christiansen et al. (1998).

In another segmentation model that uses stress, Yang (2004)[2] reports rather good performance using a rule-based system. The model essentially segments utterances before strong syllables. If there are one or more weak syllables between two strong syllables, it either inserts a boundary randomly (random model), or ignores this part of the utterance (agnostic model). The words ignored in multi-word utterances are learned later if they are observed as a single utterance. Gambell and Yang (2006) report 95.9% word precision, and 93.4% word-recall for their random model; and 85.9% precision and 89.9% recall for their agnostic model on child-directed utterances from Brown (1973) corpus from CHILDES.[3] The stress information is obtained from the CMU pronunciation dictionary. Even though the results look impressive, there are a

[2] An extended version is presented by Gambell and Yang (2006, unpublished manuscript).

[3] Not to be confused with the well-known *Brown corpus* of written English (Francis and Kucera, 1979).

Word Form	Stress	Tokens	Types
Monosyllabic	2	28566	812
Bisyllabic	20	3437	417
	22	516	8
	02	347	52
	00	38	6
Trisyllabic	200	297	53
	020	149	20
	202	12	4
	102	4	2
Quadrisyllabic	2000	6	3
	2010	3	1
	0200	2	2
Sum	12 forms	33377	1380

Table 8.3: The stress distribution on BR corpus. As in the MRC database, the stress patterns are coded so that '2' indicate primary stress, '1' indicates secondary stress, and '0' indicates no stress.

number of assumptions this strategy makes that are difficult to justify. First, it assumes that word boundaries always correspond to the syllable boundaries (see Section 5.1 for a discussion of problems with this assumption). Second, since stress is neither universally available in all languages, nor is its location the same for all languages with lexical stress, it is not clear how children might come to decide that the onset of a strong syllable is the position to segment words.

8.3.2 An analysis of stress patterns in the BR corpus

Throughout this thesis the segmentation simulations are tested on the BR corpus which has become the de facto standard for testing computational models of segmentation in the literature. This allows direct comparison of the models with each other and with the other models in the literature. Unfortunately, none of the previous studies that use stress for segmentation were tested on BR corpus.

For the results reported in this section, the BR corpus is stress-marked semi-automatically following the procedure of Christiansen et al. (1998). The stress assignment is done according to stress patterns of the MRC psycholinguistic database. All single-syllable words are coded as having primary stress, and the words that were not found or did not have stress assignment in the MRC database were annotated manually. Further details on the annotation process can be found in Appendix A.

Pattern	Boundary	Word internal	
		Syllable	Phoneme
00	50	349	5830
01	0	3	3
02	2080	518	518
10	0	7	7
11	0	0	7
12	0	0	0
20	382	3911	3911
21	4	0	0
22	21071	516	52156
sum	23587	5295	62432

Table 8.4: Stress patterns at boundaries and word-internal positions for both phoneme- and syllable-transitions.

Table 8.3 presents the stress patterns of words in the resulting corpus. Although the corpora are different, these figures are similar to the ones reported by Christiansen et al. (1998) for the Korman corpus. First thing to note is that majority of the words are monosyllabic, 85.6% of the word-types and 58.8% of the word tokens are monosyllabic. This means one can get 85.6% of the words correctly by inserting boundaries before and after every syllable. Furthermore, according to this stress assignment, 98.4% of the words start with a strong syllable.[4]

Another way to look at the stress data is to analyze the distribution of adjacent stress patterns over boundaries and word-internal locations. This would give an indication of stress transitions that signal possible word boundaries. Table 8.4 presents this distribution for all possible transitions of stress levels specified in the MRC database. For example, the last row of this table indicate that the stress pattern 22 (primary–primary) straddles word boundaries 21,071 times. Within the words, this pattern occurs at the syllable boundaries only 516 times, and it occurs 52,156 times during phoneme transitions. Since the stress levels change only on syllable boundaries, that when there is a transition between stress levels the values for syllables and phonemes are the same.

As expected, the majority of the transitions (or lack thereof) are from primary stress to primary stress. For boundary detection most promising pattern is weak–strong transition, which is expected because of the trochaic bias in English. A trivial segmentation algorithm that inserts a boundary between all phonemes pairs with this pattern would guess 2080 boundaries correctly, which would give a 80.0% precision, but it would cover only 8.8% of the boundaries. On the other hand, it is clear that

[4] This further explains the success of the rule-based model of Gambell and Yang (2006).

inserting a boundary is not a good idea when stress transition is strong–weak. Inserting a boundary between two phonemes would discover 382 boundaries correctly with a precision of 8.9%, and a recall of 1.6%. Since these transitions only occur at syllable boundaries, for these two cases the difference between using the syllable or the phoneme as the basic unit, does not cause much difference. To achieve a good segmentation performance from stress patterns, the key factor is the strong–strong transitions that are most frequent. Unfortunately, this is only possible with the assumption that the word boundaries can only be at the syllable boundaries. If we insert a boundary between every strong–strong transition, we get 97.6% precision, and 89.3% recall with syllables. However, the same decision results in a mediocre precision of 28.7% if we do not assume syllable boundaries are known (note that since number of boundaries do not change with the change of units, recall is the same for both syllables and phonemes).

The analysis presented above does indicate that lexical stress can indeed be useful in certain ways. However, there are a number of problems with its use for the type of models presented here. The first problem is the assumption that word boundaries coincide with the syllable boundaries. Note that this is not the same as assuming the knowledge of syllable, which is well known to be a salient unit of language processing. The problem stems from the fact that, in fluent speech, frequently the syllables can often straddle word boundaries (see Section 5.1 for a discussion). Second, the use of stress to mark prominent syllable in a word is not universal. And last, a more practical problem is that currently corpora with realistic stress marking are not available.

8.3.3 Segmentation using lexical stress

Even though the analysis in Section 8.3.2 is somewhat discouraging for the segmentation strategy followed throughout this thesis, this section presents results from a stress-based segmentation model following the same strategy used throughout this thesis.

As briefly discussed in Chapter 7, a possible strategy for using stress for segmentation is to use the stress patterns at the utterance boundaries. This method does not provide much help for the learner, because, as presented in Table 8.3 most of the syllables in the corpus have primary stress. As a result, the beginnings and ends of the utterances are also dominated with primary stress. In the BR corpus 98.9% of the utterances begin with, and 82.5% of the utterances end with strong syllables. Even though this indicates that it is more likely for words to end in weak syllables, compared to word beginnings, it is a big leap of faith for a learner to choose segmenting after weak syllables given that they occur 17.5% of the time at the end of utterances. As a result, if we train any learner on utterance boundaries, the likely outcome is not segmenting at all. However, as Swingley (2005) also pointed out, if we take the word-types instead of the word tokens, we have a better chance of generalization. If we count the words that start with strong syllables in the gold standard lexicon, it is similar to utterance beginnings: 94.1%. However, if we count the stress at the end of the words, strong

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
SM	78.2	8.2	14.8	26.5	9.7	14.2	8.2	38.7	13.5	0.9	92.8
PUWM	83.7	91.2	87.3	74.1	78.8	76.4	43.9	67.7	53.3	6.7	8.8
PUWSM	92.8	75.7	83.4	78.3	68.1	72.9	26.8	62.7	37.5	2.2	24.3
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 8.5: The performance measures for stress-based segmentation model (SM). The third row (PUWSM) is the combination of stress with previously presented models. As well as the usual reference models, the performance scores for combination of the previous models without stress (PUWM) are provided for comparison.

syllables are the last syllable of only 36.5% of the word types.

The model presented in this section, the stress-based model, uses a similar strategy like the phonotactics models described before, except, instead of collecting statistics about phoneme n-grams, the stress-based model collects statistics about stress assignment on each phoneme. As before, the measures used are the probability of observing a certain stress pattern l at the end of words, $P(l|wb)$, and the probability of observing a certain stress pattern r at the beginning of the words, $P(r|wb)$. The measures can be calculated using stress patterns spanning variable length l and r. For the results reported below, both l and r are varied between one to three. Using the peak based boundary decision (Section 6.2.1) and weighted majority voting algorithm (Section 6.2.3), Table 8.5 presents performance scores for stress-based algorithm and its combination with previously presented models. The line labeled ‘SM’ (short for ‘stress-based model’) presents the scores for calculating stress scores from already discovered word types. Like the results presented previously in Table 8.1, the discovery of words is bootstrapped by utterance boundaries.

The first row, labeled SM, in Table 8.5 presents the results of using the stress-based model alone, bootstrapping from the utterance boundaries. As expected, the performance is not impressive. However, even though it heavily undersegments, the stress-based model makes very few oversegmentation errors. This is also reflected in the performance scores with high precision and low recall. Note that the low error rate is not only due to conservative boundary decisions. The stress-based model indeed makes very few positive decisions for boundaries, however, a random model constrained to insert same number of boundaries (leading to 92.8% undersegmentation) is expected to make 7.2% oversegmentation errors, where the oversegmentation errors the stress-based model makes are less than one percent. The findings are also in line with the analysis provided in Section 8.3.2. As expected, the model learns to segment at weak–strong transition, which is expected to be precise. However, since majority of the stress transitions are strong–strong, this covers rather a small portion of the boundaries.

As shown in the second row of Table 8.5, the combination of the stress-based model with others (performance scores presented in the row labeled ‘PUWSM’), increases the performance over stress-based model. However, the contribution of the stress-based model to the performance of the previously discussed models (repeated in the row labeled ‘PUWM’) is not worthwhile.

Combining the stress-based model with others adversely affects the performance scores calculated for the complete corpus, except BP and E_o. However, if we plot the performance change as more input is provided, the benefit of the stress cue becomes visible. Figure 8.5 presents the changes of BF, WF, LF for stress-based model alone (SM), and the combination of the stress-based model with others (PUWSM). For the last 290 utterances of the BR corpus, BF, WF, and LF reaches to 91.4%, 83.9% and 78.8%, respectively. The oversegmentation error rate drops to 4.5% and oversegmentation error drops to 6.7%. These results present an improvement over most of the final performance values of the combined model without stress (PUWM) presented in Section 8.2. Even though stress-based model does not perform well alone, and even if it seems to have an adverse effect on the performance scores calculated for the complete corpus, its contribution to the final performance of the combined model is clearly positive. The stress cue seem increase the errors at the beginning of the learning process. However, as more data becomes available, the contribution of the stress to causes an increase in precision, and this contribution is visible in the final performance of the PUWSM.

The reason that the lexical stress has not been very successful alone in the simulations reported here may be due to a number of different problems. First, the stress marking used here is rather crude and unrealistic. A better, more realistic, stress marking may help improve the situation. Second, the weighted majority voting algorithm used for combining results is not flexible enough for deciding when and where a particular cue is useful. A more elaborate combination model may exploit usefulness of the stress cue in better ways. Third, despite its problems, a realistic inclusion of syllable structure of the utterances may provide a major improvement.

8.4 Summary

This chapter investigated two ways of extending the segmentation models developed so far in Chapter 6 and Chapter 7 using information from already discovered words.

The first strategy makes use of already known words and is based on the words themselves. Even a partial lexicon with some incorrect words is useful for segmentation, and this has been demonstrated by the performance of the model presented in Section 8.2. Furthermore, the combination of this model with the previous models improved overall performance of the combined model resulting in the best performance values presented in this thesis. The model is still relatively simple, and can be improved in many ways. For example, the phonotactic models used in this thesis pay attention to only the beginning and ends of the words. However, the structure of words can be

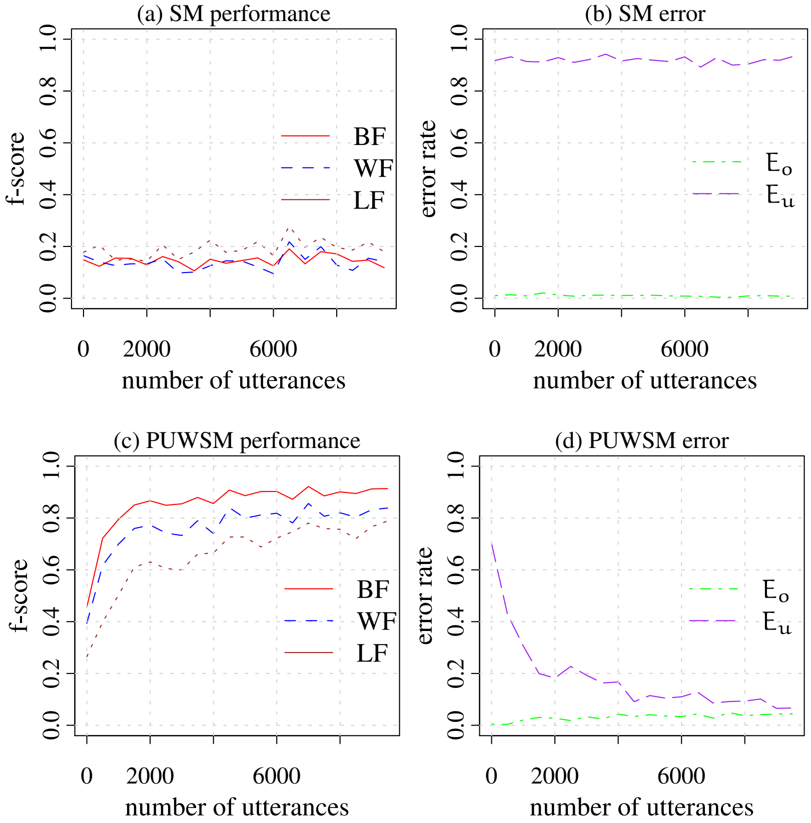


Figure 8.5: Boundary (BF), word token (WF), and word type (LF) f-scores and oversegmentation (E_o) and undersegmentation (E_u) error rates for the word-based segmentation model, the WM, and its combination with previous strategies, the PUWSM. The scores are calculated for each 500-utterance block in the BR corpus during the learning process.

modeled more realistically. For example, the learner can learn certain constraints or tendencies (such as ‘a word should have a vowel’), or the length of a typical word in phonemes or syllables. Similarly, using only the frequencies of words is not the best way to characterize a word’s usage. A possible improvement that does not require additional resources would be to take the number of contexts where the word is observed. This additional modeling component may match the intuition that words are not only frequent, but also used in varying contexts.

The second strategy, the use of lexical stress, does not look as promising as the lexical information. Even though the model presented in Section 8.3 achieved good precision scores on the BR corpus, the recall was very low. At first sight, combining the stress-based model with the others affected the performance of the combined model adversely. However, it seems this is because of the initial mistakes it causes during learning. Once the combined model learns how to use the stress cue, i.e., in the advanced phases of the learning, the results improve over the model without the stress cue. The negative results obtained with the stress-based model have to do with a number of factors that point to future directions for modeling the effect of stress. Most importantly, the stress annotations in currently available corpora of child directed speech are not realistic enough for drawing any strong conclusions. Furthermore, for better use of the stress cue, the knowledge of the syllable (either built into the system, or learned from the input) seems to be crucial for the success of a stress-based segmentation model.

Even though the models presented in this chapter can further be improved in a number of ways, the results presented here already indicate that the combination of information at different levels lead to better performance. The next chapter will provide an overall view of the results presented in this thesis.

9 An Overview of the Segmentation Strategies

> I don't pretend we have all the answers. But the questions are certainly worth thinking about.
>
> Arthur C. Clarke

The last three chapters presented four different segmentation methods and their combinations in the order in which they were presented. This chapter brings all the results discussed in these three chapters together under a unifying perspective. In the sections that follow, I will summarize the results presented in the preceding chapters, compare them to each other, and discuss some common issues whose discussion was deferred during discussion of individual strategies. Throughout this discussion I will point to possible improvements, and future directions for the research.

9.1 The measures and the models

The first unsupervised model of discovering lexical units in continuous speech stream is presented in Chapter 6. First part of this chapter, Section 6.1 defined four measures of predictability (or uncertainty), namely, *transitional probability* (TP), *point-wise mutual information* (MI), *successor variety* (SV) and *boundary entropy* (H). The analysis presented in this section indicated that these measures correlate with word boundaries. Furthermore, variations of these models which condition the calculations right-to-left instead of left-to-right, and/or use multiple, varying sizes of phoneme n-grams are discussed. The analysis indicated that despite the fact that they are not completely independent, all measures provide some additional information regarding boundaries. In addition, it was found that the measures form two groups: TP and MI in one, SV and H in the other. The measures within each group are more closely related to each other than the measures in the other group.

While presenting the use of these measures in segmentation, Section 6.2 defined an unsupervised strategy for deciding whether there is a boundary at a position in the input utterance, given an indication of a word boundary. The strategy suggests boundaries at the positions where there is an increase in the unpredictability followed by a decrease. In other words, boundaries are suggested at the *peaks* of unpredictability. Next, a simple algorithm, *weighted majority voting* for combining multiple indications was described, which also allowed us to improve the peak-based boundary criterion.

The final segmentation model described in Chapter 6 as a representative example of a predictability-based segmentation model (PM) was a model using measures MI and H calculated on phoneme n-grams between one and three.

In Chapter 7 on using utterance boundaries, I presented a model that learned partial phonotactics from utterance beginnings and ends. The segmentation strategy based on utterance boundaries used the probability of observing a certain phoneme sequence at the utterance beginnings and utterance ends as an indication of word beginnings or word ends respectively. Again, a representative model for this strategy, the UM, was defined. The UM combines the boundary measures using varying sizes of phoneme sequences between one and three.

Chapter 8 introduced two strategies based on learning from previously discovered words. First, a strategy that makes use of the words in the lexicon was defined. This strategy favors use of frequent and well-formed words on both sides of the candidate boundary. The models following this strategy make use of two separate measures, first, the frequency of words, and second, the phonotactics of the words. The first measure is simply the frequencies of the words discovered so far. The second measure is exactly the same phonotactics measure defined in Chapter 7, but it uses the beginning and ends of already discovered word types rather than utterance boundaries. The combined model, the WM, uses n-gram sizes one to three for the phonotactics component, and the sum of the frequencies of already known words on both sides of the boundary position for the frequency component.

The second strategy investigated in Chapter 8 was based on lexical stress. This strategy learns the stress pattern from already known words, and uses this knowledge in later boundary decisions. The model is similar to the phonotactics models described earlier, except that instead of phonemes, the model uses the stress levels of the syllable that a particular phoneme is part of. I use the name SM for this model throughout this thesis.

9.2 Performance

Along with their descriptions, a set of performance indications are reported for all models described in the preceding chapters. These results, together with two reference models introduced in Section 5.4, are repeated in Table 9.1. As in the previous presentations, the results presented in these tables are the performance scores for the complete corpus. The performance of the learners towards the end of the learning phase will be discussed below.

Figure 9.1 presents the f-scores for boundaries, word tokens, and word types (BF, WF and LF) and oversegmentation and undersegmentation errors (E_o and E_u) for each individual model presented in previous chapters, including two reference models the reference model based on language modeling strategy (the LM) and the pseudo-random

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
PM	69.6	92.5	79.5	56.9	70.2	62.9	36.7	49.8	42.3	15.3	7.5
UM	82.9	84.8	83.8	70.5	71.7	71.1	33.8	66.9	44.9	6.6	15.2
WM	77.5	71.3	74.3	60.6	57.2	58.9	18.3	47.7	26.4	7.8	28.7
SM	78.2	8.2	14.8	26.5	9.7	14.2	8.2	38.7	13.5	0.9	92.8
PUM	82.6	90.7	86.5	72.4	77.4	74.8	42.8	65.3	51.7	7.2	9.3
PUWM	83.7	91.2	87.3	74.1	78.8	76.4	43.9	67.7	53.3	6.7	8.8
PUWSM	92.8	75.7	83.4	78.3	68.1	72.9	26.8	62.7	37.5	2.2	24.3
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 9.1: Performance scores for all the models discussed in previous chapters. The values are repeated here in a single table for ease of comparison.

segmentation baseline (th RM). The reference models were described in Section 5.5. The results presented in Figure 9.1a–b were calculated over the complete corpus during incremental learning. The results in Figure 9.1c–d reflect the performance of the learner at the end of the learning phase. The first set of results are affected by the initial mistakes the learner makes during the learning process, while the second set of results reflect the performance of the learner in a more advanced phase in learning (see Section 5.4 for a detailed discussion).

Except the stress-based model (SM), the rest of the models perform higher than the random baseline (LM), and comparably to the state of the art reference model LM. The model that learns phonotactics from utterance boundaries (UM) seems to perform slightly better than others. However, in general, the performances of individual models, except (SM), seem to be comparable to each other. Except the SM, all models perform only slightly worse than the state of the art reference model LM.

The stress-based model performs even worse than random segmentation for some scores. The reason, and the nature of this failure is clearer if we take a look at the error scores. The SM performs badly because it is conservative: it makes very few oversegmentation mistakes at the cost missing over 90% of the boundaries. The way it is used here, the lexical stress seems to be a precise cue. However, its coverage is very low.

Figure 9.2 presents the progression of the f-scores and error measures for the combination path taken in the preceding chapters. Figure 9.2a presents the performance of the models on the complete BR corpus, Figure 9.2b presents the performance values for the last 290 utterances. The benefit of combining predictability (PM) and the utterance boundaries (UM) is clear. All performance scores increase. The reason is clear if we take a look at the asymmetry of the error scores of PM and UM in Figure 9.1. The PM makes more oversegmentation errors than the UM, but the

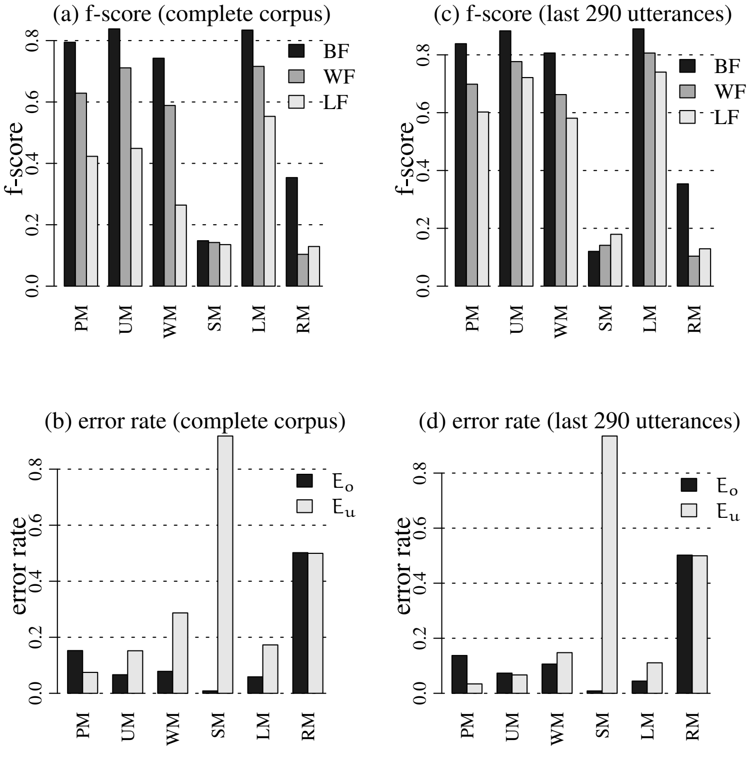


Figure 9.1: Boundary (BF), word-token (WF) and word-type (LF) f-scores and oversegmentation (E_o) and undersegmentation (E_u) rates of all individual models, including the reference models, calculated during learning (a,c) for the complete BR corpus, (b,d) only for the last 290 utterances of BR corpus.

difference is reversed for undersegmentation errors. The combined model reduces the oversegmentation errors of the PM, and achieves a better overall score. Figure 9.2 also indicates a slight increase in undersegmentation errors for the combined model PUM, compared to the PM, but this does not seem to affect the overall performance adversely.

The PUM, combination of the predictability (PM) and the utterance boundary (UM) strategies shows a clear improvement over the individual components. However, the effect of addition of the word-based segmentation model, as can be seen the difference between the performances the PUM and the PUWM (PM–UM–WM combination), is not as clear. However, as can be seen more clearly in Table 9.1, for overall performance, there is a slight but consistent improvement on all measures. As expected, the usefulness

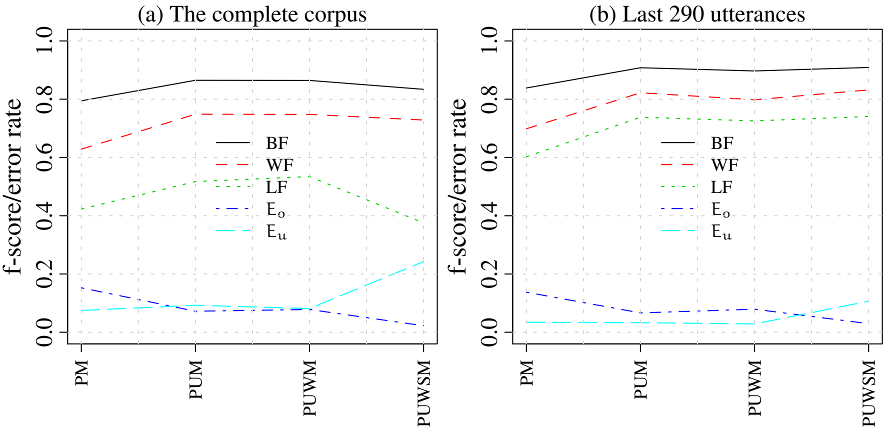


Figure 9.2: Progression of the F-score and error values for the combination of models presented in the last three chapters.

of the WM is particularly visible in increasing the lexical scores. However, the effect of the WM on the final state of the learner is not that clear, and in some cases causes a poorer performance. It seems that the WM speeds up the learning at the beginning, so the overall performance of the combined model is better with the WM included. However, it does not seem to have a clear effect on the final state of the learner. Admittedly, the model of using lexical information for learning segmentation is one of the less deeply investigated areas in this study, and some of the ideas for improvement in the future research were listed at the end of Chapter 8.

Figure 9.2 reveals an interesting effect of combining the SM with others (the model labeled PUWSM). The SM is very conservative, and it affects the overall performance by causing many undersegmentation errors. As a result, all of the performance values calculated on the overall corpus drops. However, its effect at the end of the learning phase is positive. This can be explained by the fact that the model itself learns the stress patterns that are useful, and the weighted majority voting algorithm learns how to weigh decisions of the stress-based model, improving the performance slightly for all performance scores.

The differences of performance scores in Figure 9.1 and Figure 9.2 between the calculations based on the complete learning process and those based on the final state of the learning demonstrates that for the incremental learners, the way performance is evaluated changes the apparent success of the learners on the segmentation task. This seems to be particularly important for lexical scores (LP, LR and LF), since initial phases of learning introduce many incorrect word types, and causes overall performance to degrade.

The reason the inclusion of last two models, the WM and the SM, did not increase the performance of the combined model substantially may be due to a number of reasons. First, it may be that these two cues are utilized rather poorly by the models described in Chapter 8. As discussed in Section 8.4, these models can indeed be extended in a number of ways.

Second, on a related note, it may also be that the information provided by these cues is insignificant. This is particularly true for the SM due to poorly marked stress in the corpus. A corpus with more realistic stress annotation may allow better use of this cue. These two possible reasons for the lack of compelling improvement seem to be supported by the combination of these models in other ways. Compared to combining the PM and the UM, particularly the SM (but also the WM to some extent) does not increase the performance when combined with other models (a complete tabulation of the performance values for all possible combinations is provided in Appendix B).

A third possibility is that combination of predictability and utterance boundaries already uses most of the information in the data usable for segmentation. In other words, we are close to the upper bound of what we can achieve with the information at hand. This may seem likely, because (1) the final combined model performs similar to other state of the art segmentation models in the literature, (2) except stress, the reason all these segmentation strategies work stem from the fact that the input stream is formed by concatenation of a limited set of lexical units. However, numerous possibilities for improving individual models and their combination listed in this chapter suggest that even without additional sources of information we may achieve better performance. Nevertheless, an interesting task for the future research is to establish the upper bound that can be achieved using only the information available in the transcribed corpus.

Fourth, the reason for not observing substantial improvements by adding more cues may also be due to the way they are combined. The majority voting algorithm used for combining these results is not the best known algorithm for this purpose. This possible problem will be discussed at length in Section 9.6

9.3 Making use of prior information

Section 9.2 demonstrated an expected effect. Learners make mistakes at the beginning. The performance measures calculated during the complete learning process may set apart fast learners from slow learners. However, as long as learning is achieved in a reasonable time, what counts more is their final performance. The models presented earlier in this thesis learn two different aspects of the input languages. The first is the statistical relationships between the neighboring phoneme sequences, and the second is the optimal weights for different indications to word boundaries. This allows us to investigate another plausible scenario relevant to acquisition of lexical units.

Children start to show some sensitivity to a limited set of words, such as their names, around 6-months of age (Bortfeld et al., 2005). However, they start to utter their first words later, around their first birthday. Typically, comprehension precedes production:

with a rather large variability, the number of words a one-year-old understands seem to be around 50 (Fenson et al., 1994). The lexical knowledge increases slowly until about 18 moths of age, at which time a rapid increase in lexical knowledge, known as the *vocabulary spurt*, starts (Reznick and Goldfield, 1992, see also the discussion in Section 2.1.1). In the light of the research so far, it seems fair to assume that children start building a lexicon around one year of age. However, infants attend to speech and they start collecting statistics about sound sequences long before they start building a lexicon. As a result, it is plausible that lexical segmentation starts with some prior knowledge of phoneme sequences. To test the usefulness of this type of prior knowledge, the learning algorithm described in Section 6.2.3 was modified to collect phoneme n-gram statistics from a different data set as the first step. After collecting the phoneme sequences, the learning proceeds as before.

The corpus used for initial statistics is gathered from the CHILDES database. For all American English transcripts in the CHILDES, the recording sessions where target children were less than a year of age were selected, and all child-directed speech in these sessions are processed and converted to phonemic transcriptions following Brent (1996). The resulting corpus contained 53,770 child directed utterances for 24 different children recorded in 171 sessions. The ages of children were between 0;6 and 0;11.29 (mean=9;11, sd=48 days).[1] Since the resulting corpus is larger than the BR corpus, and it includes many words not marked for stress in MRC database, annotating this corpus with stress information was not practical.[2]

Table 9.2 presents the performance of the models discussed so far using the modified algorithm. First, the phoneme n-gram statistics from the corpus described above were collected. During this first step no attempt was made to segment the corpus, or to learn the weights of the boundary indications. In the second step, learning proceeded as before. The phoneme statistics are updated over already existing statistics, and the BR corpus is segmented using the methods indicated in Table 9.2, where previously reported results without prior information are also provided for ease of comparison.

Compared to the results without the prior statistics, all models, except the word-based model alone, show an increase in the performance scores. The WM, as expected, shows no difference since the lexicon is not populated with the prior data collection. Figure 9.3 presents these differences graphically only for combined model PUWM. The differences are more visible for the performance scores calculated for the complete corpus, especially increasing LF, and decreasing $\mathsf{E_u}$. This is expected, since having prior data reduces the initial mistakes that the learner otherwise makes. As can be seen in Figure 9.3b, use of prior data also increases the performance scores in the final state of the learner, albeit slightly.

[1] As previously explained in Section 2.1.1, the age notation follows the standard notation in the language acquisition literature. 0;11.29 means 0 years, 11 months and 29 days.

[2] The decision is also related to the poor quality of the stress information available. If a more realistic source of stress information available, the effort may be well justified as a future step.

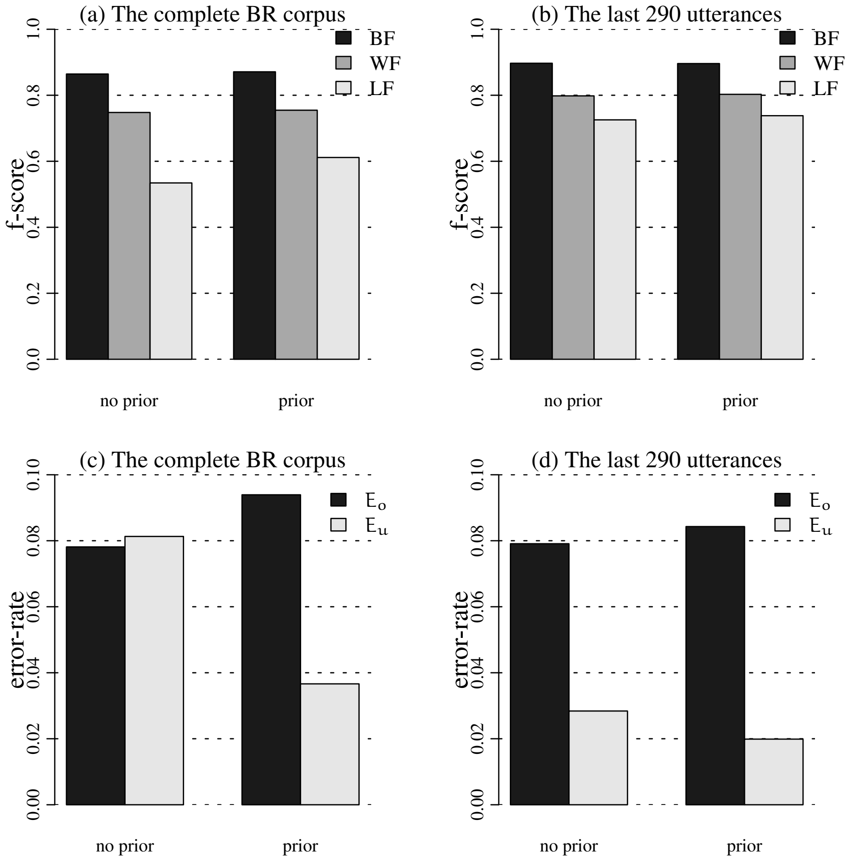


Figure 9.3: Boundary (BF), word-token (WF) and word-type (LF) f-scores and oversegmentation (E_o) and undersegmentation (E_u) rates of the combined model PUWM, with and the without prior phoneme statistics. The scores are calculated (a,c) for the complete BR corpus, and (b,d) only for the last 290 utterances of BR corpus. Note the scale difference between the y-axes error rates and the f-scores.

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
PM	69.6	92.5	79.5	56.9	70.2	62.9	36.7	49.8	42.3	15.3	7.5
	66.9	96.0	78.9	53.7	70.2	60.9	42.7	47.1	44.8	18.0	4.0
UM	82.9	84.8	83.8	70.5	71.7	71.1	33.8	66.9	44.9	6.6	15.2
	78.6	94.0	85.6	68.5	78.0	72.9	54.3	64.4	58.9	9.7	6.0
WM	77.5	71.3	74.3	60.6	57.2	58.9	18.3	47.7	26.4	7.8	28.7
	77.5	71.3	74.3	60.6	57.2	58.9	18.3	47.7	26.4	7.8	28.7
PUM	82.6	90.7	86.5	72.4	77.4	74.8	42.8	65.3	51.7	7.2	9.3
	79.7	95.7	86.9	70.3	80.3	74.9	57.3	62.6	59.8	9.2	4.3
PUWM	83.7	91.2	87.3	74.1	78.8	76.4	43.9	67.7	53.3	6.7	8.8
	79.5	96.3	87.1	70.6	81.1	75.5	59.5	62.9	61.1	9.4	3.7
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3

Table 9.2: Performance scores without and with prior information for the complete BR corpus. For each model listed, the first row presents the previously presented scores without the prior statistics, and the second row presents the performance scores for the complete learning process on the BR corpus after the phoneme statistics were updated using a larger corpus of child directed speech described in this section.

9.4 Variation in the input

A natural concern regarding the validity of the results presented in this thesis is whether these results are representative for child directed speech, or are they due to some peculiarities that may be found only or almost only in the BR corpus. Since the BR corpus is used in segmentation research rather often, the performance can even be a by-product of the research conducted using this corpus for over a decade. The larger child-directed speech corpus used as prior data in Section 9.3 already gives some hints that the model is not learning peculiarities of this data. The models are not specialized for the BR corpus. In this section I will present two more results to reassure about this. First, I will present the performance results obtained using the larger corpus presented above. Second, I will present the performance results obtained by randomizing the order of utterances in the BR corpus.

Table 9.3 presents the results obtained using the larger corpus described in Section 9.3 in comparison to the previously reported results for the BR corpus. Except the WM, all models perform better on the larger corpus. However, even though it performs poorly on its own, the WM's contribution to the combined model is positive. Particularly, it decreases the undersegmentation errors and increases the LF.

The order of natural language utterances is not arbitrary. There are certain regu-

	boundary			word			lexicon			error	
model	P	R	F	P	R	F	P	R	F	E_o	E_u
PM	69.6	92.5	79.5	56.9	70.2	62.9	36.7	49.8	42.3	15.3	7.5
	73.1	98.2	83.8	62.8	77.7	69.5	35.4	56.2	43.4	13.8	1.8
UM	82.9	84.8	83.8	70.5	71.7	71.1	33.8	66.9	44.9	6.6	15.2
	83.2	94.1	88.3	74.6	81.4	77.9	31.6	76.1	44.7	7.3	5.9
WM	77.5	71.3	74.3	60.6	57.2	58.9	18.3	47.7	26.4	7.8	28.7
	51.5	86.8	64.7	33.8	49.9	40.3	15.7	32.7	21.2	31.3	13.2
PUM	82.6	90.7	86.5	72.4	77.4	74.8	42.8	65.3	51.7	7.2	9.3
	85.2	97.2	90.8	78.6	86.2	82.3	44.2	71.2	54.6	6.4	2.8
PUWM	83.7	91.2	87.3	74.1	78.8	76.4	43.9	67.7	53.3	6.7	8.8
	84.2	98.0	90.6	77.7	86.6	81.9	46.5	70.5	56.0	7.1	2.0
RM	27.4	27.0	27.2	12.6	12.5	12.5	6.0	43.6	10.5	27.1	73.0
LM	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3
	88.9	93.1	90.9	81.8	84.5	83.1	50.2	66.0	57.1	4.5	6.9

Table 9.3: The comparative performance results obtained on the BR corpus and a larger child-directed speech corpus (explained in Section 9.3). For each model, the first row presents the previous results obtained on the BR corpus (previously summarized in Table 9.1), and the second row presents the results obtained on the larger corpus. All values are calculated for the complete corpus. As before, the performance scores of reference models are provided for comparison.

larities that can be found in utterances of a natural conversation. For example, it is well known that a word used in an utterance is more likely to be repeated during the same conversation compared to its base frequency of occurrence in the language. The experiments presented next disregard this fact by re-ordering the utterances in the BR corpus randomly. The variance of the performance scores for different orderings of the input gives an indication of the robustness of the model with respect to ordering of the utterances. In addition, the difference between the mean of the multiple randomized runs and the results obtained using the natural ordering of the utterances may indicate the usefulness of the natural ordering.

For this purpose, box-plots of f-scores and error values for 50 runs of the PUWM with randomized input are presented in Figure 9.4. The dotted horizontal lines in the graphs represent the scores obtained on the natural ordering of the utterances. The ordering of the utterances does not seem to cause a large variation. Due to the shorter sample size, the scores calculated on final 290 utterances are more varied. However, in both cases, the standard deviations of the distributions of the scores are rather low. Although it is difficult to get a conclusive interpretation from these values, the natural ordering of the utterances seem to be useful. The scores obtained on natural ordering seem to be higher (and for the error scores lower) than the mean of the random runs,

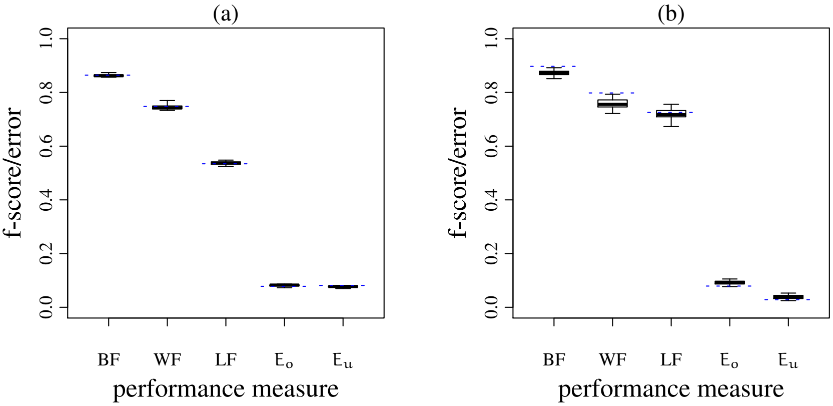


Figure 9.4: Boxplots of the performance scores obtained over 50 learning trials over the BR corpus with randomized orderings of utterances. (a) presents the scores calculated for complete corpus, while (b) presents the scores calculated for last 290 utterances. Whiskers cover the complete range of scores (including outliers).

and this is more visible for the performance scores calculated for last stage of the learning.

9.5 Qualitative evaluation

In the evaluation of the models so far, quantitative measures of success and failure have been presented. This is a reasonable approach, especially when dealing with a large amount of data. However, inspecting the output of a model qualitatively also helps identifying the cases where the model is successful, and where it fails. In this section I will present some examples of utterances segmented and words identified by the combined model (PUWSM) and the reference model (LM) starting with a familiar example.

Figure 4.1 on page 45 presented a sequence of utterances from the BR corpus in the form of a puzzle. The same sequence of utterances will serve here for evaluating the performance of the learners qualitatively. In this section, the results will be displayed using the phonemic transcriptions. To aid the reader in interpreting the results Figure 9.5 presents the phonemic and orthographic forms of the sequence of utterances that were used in the puzzle.

This particular sequence is formed by utterances 422$^{\text{nd}}$ through 429$^{\text{th}}$ in the BR corpus. Besides the repeated use of the word *kitty*, the utterances in this sequence have another interesting property that they occur at an early position in the BR corpus, where all models presented here are still actively learning. As can be observed from the

Orthographic	Phonemic
a kitty	6 kIti
do you think it's a kitty	du yu TINk Its 6 kIti
you can think it's a kitty	yu k&n TINk Its 6 kIti
is that a kitty	Iz D&t 6 kIti
here kitty kitty sleeping kitty	h(kIti kIti slipIN kIti
look you can stick your finger in there	lUk yu k&n stIk y) fINgR In D*
stick your finger in that hole right	stIk y) fINgR In D&t hol r9t
good girl	gUd g3l
now let's see what's the baby saying	nQ lEts si WAts D6 bebi seIN

Figure 9.5: The gold-standard solution of the puzzle presented in Figure 4.1. The orthographic version and the phonemic transcriptions are displayed together to aid easy interpretations of the results presented in this section.

graphs presenting the progress of the performance with increasing input (for example, Figure 8.5), almost all models reach their stable state around the 2000th utterance. After this point, some slight increase and some fluctuation is observed, but, the steepest increase of performance (end decrease of error) is observed until around the 2000th utterance. Figure 9.6a presents the segmentation solutions offered by the combined model, the PUWSM, and the reference model LM. The models make different mistakes, but both of them undersegment. This is expected, since they still are at the beginning of the learning process. This is also in agreement with the progress of the E_u displayed in Figure 8.5 and Figure 5.2 for the PUWSM and the LM, respectively. It is also visible in this example that the LM learns faster. It finds more boundaries compared to the PUWSM.

Figure 9.6b presents the outputs of the models for the same sequence of utterances. However, for this demonstration, the utterances are moved to the end of the BR corpus. This reflects the performance of the learners towards the end of the corpus. As expected, both perform better than in the earlier phase. It is difficult to see the performance differences very clearly. However, even though the LM also makes some oversegmentation mistakes (for example segmenting /fINgR/ 'finger' as /fIN gR/), the PUWSM makes more oversegmentation errors. This can be seen in the first utterance, and the utterances where the PUWSM identifies /k/ as a word. These oversegmentation errors indicate that the PUWSM can benefit from a better phonotactics component (e.g., one that can learn that /k/ is an unlikely word).

This example also demonstrates that both models consistently segment the morpheme /IN/ 'ing'. Even though the comparison with the gold standard will mark this as an error, it is not clear what this would mean for child language acquisition. The fact that the morphemes are learned and used productively at some point in the acquisition process suggests that identifying the morphemes is not necessarily bad for a learner.

LM	PUWSM
6kIti	6kIti
du yu TINk It s6kIti	du yu TINk Its6kIti
yu k&n TINk It s6kIti	yu k&nTINk Its6kIti
Iz D&t 6kIti	Iz D&t 6kIti
h(kItikIti slipIN kIti	h(kIti kIti slipIN kIti
lUk yu k&n stIky)fINgRIn D*	lUk yuk&nstIky)fINgR In D*
stIky)fINgRIn D&t holr9t	stIky)fINgR In D&tholr9t
gUd g3l	gUdg3l
nQ lEtssi WAt sD6 bebi se IN	nQ lEtssi WAtsD6bebi se IN

(a)

LM	PUWSM
6kIti	6k Iti
du yu TINk Its 6kIti	du yu TIN k Its6 kIti
yu k&n TINk Its 6kIti	yu k&n TIN k Its6 kIti
Iz D&t 6kIti	Iz D&t 6kIti
h(kIti kIti slipIN kIti	h(kIti kIti slip IN k Iti
lUk yu k&n stIk y) fIN gR In D*	lUk yu k&n stIk y) fINgR In D*
stIk y) fIN gR In D&t hol r9t	stIk y) fINgR In D&t hol r9t
gUd g3l	gUd g3l
nQ lEtssi WAt s D6 bebi se IN	nQ lEts si WAts D6 bebi se IN

(b)

Figure 9.6: Solutions offered to the puzzle presented in Figure 4.1 by the LM and the PUWSM. (a) presents the outputs of the models obtained where the utterances are in their normal position. (b) presents the outputs obtained by moving the utterances that form the puzzle to the end of the BR corpus.

Another way to inspect the output of the models is to check the most frequent words that they identify. Table 9.4 presents the most frequent 10 words identified by the PUWSM and the LM as well as the most frequent 10 words in the gold-standard segmentation of the BR corpus. In general, the output of both models match well with the gold-standard segmentation. Both miss some occurrences of highly-frequent words, but, the number of times they find these words is also close to the gold standard. All high-frequency words the PUWSM finds in this list are real words. Furthermore, except that it misses the word *a* many times and except that the word *do* gets a higher rank, the words in the list matches the gold-standard perfectly. The LM performs similarly, but here the problem noted above surfaces again. The LM segments the morpheme /z/ ‘-s’.

The high-frequency words that the models find indicate that both models perform

PUWSM		LM		Gold standard	
freq.	word	freq.	word	freq.	word
1488	/yu/ ‘you’	1459	/yu/ ‘you’	1704	/yu/ ‘you’
839	/D6/ ‘the’	931	/D&t/ ‘that’	1291	/D6/ ‘the’
802	/D&t/ ‘that’	891	/WAt/ ‘what’	895	/6/ ‘a’
778	/WAt/ ‘what’	855	/D6/ ‘the’	798	/D&t/ ‘that’
610	/Iz/ ‘is’	749	/z/ ‘**-s**’	783	/WAt/ ‘what’
572	/It/ ‘it’	647	/Iz/ ‘is’	653	/Iz/ ‘is’
525	/DIs/ ‘this’	622	/It/ ‘it’	632	/It/ ‘it’
504	/du/ ‘do’	521	/du/ ‘do’	588	/DIs/ ‘this’
465	/WAts/ ‘what’s’	502	/tu/ ‘to’	569	/WAts/ ‘what’s’

Table 9.4: Most frequent words discovered by the PUWSM and the LM in comparison to the most frequent words in the gold standard. The sequences that do not exist in the gold-standard lexicon are marked with boldface.

well in finding the frequent words. However, it gives limited indication as to what sort of mistakes the models make. To demonstrate the mistakes, Table 9.5 presents the most frequent sequences the PUSWM and the LM suggest as words, but which are not found in the gold-standard lexicon. Since this list is constructed by checking the words that are found regardless of their context, some errors (such as the ones caused by mistakenly identifying a rare word) is not visible in this list. Nevertheless, errors listed in Table 9.5 are useful indications of the mistakes the models make. In this table, there is a mix of undersegmentation and oversegmentation errors for both models. The first thing to note is that the errors listed for LM are more frequent. This is likely because of the fact that the LM learns faster, and reaches to a stable state quickly, and once it starts making a particular mistake, it has more chances to make more of them.

Another interesting difference between the PUWSM and the LM that can be deduced from these examples is that the oversegmentation mistakes the LM makes are typically morphemes. On the other hand, the PUWSM makes rather bold oversegmentation mistakes, such as segmenting *peekaboo* as *pee kaboo*. Despite the fact that /pi/ ‘the letter P’[3] is a word in the gold-standard lexicon, it rarely occurs outside the word *peekaboo*, and /kabu/ is not a word. Similarly, the model oversegments the words *Cindy*, *daddy* and *mommy* in similar settings. This is because of a problem we noted in Section 8.4. The model requires words to be highly frequent, but does not require words to occur in varying contexts. This particular error type indicates that modeling properties of words more carefully is another future direction to pursue for improving the model’s performance.

[3] Surprisingly, the word *pee* which would also match this sound sequence does not occur in the BR corpus.

PUWSM				LM			
rank	freq.	sequence	example	rank	freq.	sequence	example
15	340	/IN/	*-ing*	5	749	/z/	*-s*
79	67	/d&/	***da**ddy*	10	477	/IN/	*-ing*
82	63	/#yu/	*are you*	12	462	/s/	*-s*
84	63	/lUk&t/	*look at*	29	263	/WAtsD&t/	*what's that*
93	56	/nADR/	*a**nother***	33	224	/k&nyu/	*can you*
109	45	/Its6/	*it's a*	48	145	/s6/	*it's a*
113	44	/ma/	***mo**mmy*	50	139	/WAtsDIs/	*what's this*
117	44	/D6d%/	*the door*	74	97	/~t/	*are**'nt***
119	43	/s/	*-s*	81	84	/nADR/	*a**nother***
122	43	/9dont/	*I don't*	88	76	/anD6/	*on the*
123	43	/6fon/	*a phone*	102	65	/sr9t/	*it**sright***
131	40	/#Doz/	*are those*	103	65	/pr/	***pr**etty*
132	39	/hQmEni/	*how many*	109	61	/6dOgi/	*a doggy*
136	38	/s6/	*it'**sa***	111	59	/W*zD/	***whre'sth**e*
138	38	/k6bu/	*pee**kaboo***	117	57	/Iti/	*pr**etty***
141	37	/9TINk/	*I think*	118	57	/b9b9/	*bye bye*
152	33	/gEn/	*a**gain***	120	55	/snat/	*it**snot***
172	28	/9si/	*I see*	122	55	/6dOg/	*a dog*
174	27	/sIn/	***Cin**dy*	127	53	/WAtIzIt/	*what is it*
175	27	/sAm/	***some**body*	136	48	/sD6/	*what'**sthe***

Table 9.5: Most frequent errors made by the PUWSM and the LM. For each sequence mistakenly identified as a word, the number of times the model suggested the sequence as a word (freq), the rank of the sequence, and, for the undersegmentation errors, the orthographic form of the word sequence, otherwise an indication of the morpheme, or an example case where the error occurred is given in the columns labeled 'example'. For oversegmentation errors, a dash '-' at the beginning indicates a frequent morpheme, and boldface marking indicates the approximate part-word identified by the model.

The qualitative analysis presented in this section supports some of the previous findings. In particular, the model presented in this thesis can be improved by improving the phonotactics and lexicon components. Before finalizing the overview, the next section will discuss some of the possible improvements to the learning method used in the models presented in this thesis.

9.6 The learning method

All the models in this thesis follow a simple unsupervised method for making boundary decisions, and another simple method for combining the indications obtained from multiple boundary-detection measures. For the first purpose, the local changes in a measure's values are used. If a measure provides a stronger indication for the current position compared to the neighboring positions, it is taken as a boundary decision. In other words, a boundary decision is given at the positions where the indication of a boundary peaks. The measures were combined using majority voting. In a nutshell, for each position in an utterance, a number of boundary indications are collected, and if a majority of the indications is positive, the combined decision is also positive. The models assign weights to each boundary indicator, and these weights are updated based on their agreement with the majority after every positive or negative boundary decision. An indication, or a measure that agrees with the majority all the time gets a full vote, and the weight of an indication that behaves randomly is set to zero, causing its further decisions of it to be ignored.

As demonstrated by the performance scores presented so far, these simple mechanisms worked well. The overall performance of the combined model is competitive with the state of the art segmentation models in the literature. However, there are a number of points where the learning method can be improved.

For this work, the attractiveness of the majority voting algorithm has been its simplicity. The method provides a simple, but effective way of combining different quantities. When dealing with a large number of indications with varying values coming from different distributions, making yes-or-no decisions simplifies the combination process. However, it also levels the differences between strong indications and weak indications. This is bad for a good model of cue combination for segmentation, since there may be cases where a single strong indication, such as a long pause, is enough for the decision.

Two possibilities are considered as future extensions. One of the possibilities is a weighted linear, or log-linear combination strategy. The log-linear models that are becoming increasingly common in the computational linguistics literature are, in principle, attractive here as well because of the exponential-like distributions some of the measures follow (see Section 6.1). Another, possibly better, option is to use a Bayesian cue combination method (e.g., Kording et al., 2007). However both of these methods are considerably more complex than majority voting, and they are typically trained using supervised systems, and/or batch algorithms that require large amounts of input at once. Nevertheless, especially Bayesian cue combination may be another future improvement for the models presented here. Attractiveness of the Bayesian cue combination is two-fold. First, it allows the model to use the strengths the indications to be modeled. Second gives a natural way to distinguish between the prior information the learner has about a cue (i.e., how well a cue indicates a word boundary), and the reliability of the particular measurement of the cue (i.e., how reliable is the current observation).

Another possible improvement to the learning system may come from a more structured model, for example a hierarchical model. Each cue described in the previous chapters is composed of multiple sub-measures. For example, the predictability cue uses both entropy and mutual information. The combined models combine these sub-measures in a flat manner, without paying attention to which cue a particular sub-measure comes from. A hierarchical model would get a single indication from a single cue, and combine the cues together according to a set of weights that are cue specific. A model structured this way may be able to capture environmental effects on a certain cue better. For example, if the stress cue is not reliable (as in the output of a poorly constructed speech synthesizer, or a non-native speaker), a combined model could reduce the weight of the cue once it is known to be unreliable given another variable (e.g., speech synthesizer).

Besides the method of combining multiple information sources, an important aspect of the computational models described here is that they are unsupervised. The overall language acquisition process might be considered somewhat supervised,[4] in contrast to the assumptions we have consistently upheld in this thesis. Even though the learner does not get feedback for abstract concepts such as grammar rules or word boundaries, the learner's communication with his/her environment and the efficiency of the processing are dependent on learner's success in using the language correctly. Since during learning segmentation the learner has to learn one of these abstract concepts, without the knowledge of what sort of feedback we can get from communication failures, an unsupervised method is a better match for modeling this task.

The learning strategies presented in this thesis are unsupervised. In addition, they do not have any free parameters, except the magical number three: in all cases where phoneme n-grams are combined, the n-grams of size one to three are used. Section 6.2 investigated the effects of changing the length of the phoneme n-grams. However, in general, it seems the phoneme n-grams are most useful for sizes three to four, after which their usefulness starts to diminish slowly. A possible explanation for using n-gram sizes up to three, as Cohen et al. (2007) note, is that this may be related optimal processing capabilities of humans (see also Miller, 1956). The other explanation is that it can be learned from the input. Since increasing n-gram sizes further does not improve the results, even if the learner considers longer n-gram sizes as well at the beginning, an effective weight update mechanism would eliminate the ones that are not useful.

The aim of computational simulations described in this thesis is to model human performance in learning segmentation. Consequently, good segmentation performance is not the only aim. The models developed in this study try to be faithful to what we know about the human performance in segmentation.

An important aspect of segmentation by humans is that they use a set of cues for segmenting input utterances. The segmentation strategy described here does the same, it combines a number of cues known to be used by children during segmentation, and the framework is easily extensible to include more cues. As discussed above the possible methods for cue combinations have not not fully explored in this study. However, we do not yet have enough data about human performance to prefer one

[4] Even though the term may upset many linguists, language acquisition may fit better in the learning framework called *reinforcement learning* in the machine learning literature. The main difference to *supervised learning* is that learner is active, it gets feedback for its actions, but the feedback can be delayed.

particular method of combination to another.

Human sentence processing is known to be incremental and predictive. Following human segmentation and sentence processing, the models segment the input utterances in an incremental way, without waiting for the end of the input. In a sense, the lexicon-based segmentation described in Section 8.1 adds a predictive component by favoring boundaries that start at known word beginnings.

One last issue that requires additional notice here is the input. The input used in these simulations are realistic in the sense that they are samples from real child-directed speech. However, the simulations reported here take a phonemically transcribed speech as input. Furthermore, the transcriptions represent a word consistently the same wherever it appears. In real-world speech, the tokens of the same words sound at least slightly different depending on many variables, such as the context of the word, or noise in the environment. The transcriptions remove these differences, making the task of the learner easier. The same is also true for the stress marking used in this research. On the other hand, transcribed speech also removes a set of cues that helps discovering word boundaries. This relatively unrealistic input representation is used for practical reasons: we have neither corpora that encodes all relevant aspects of speech nor a standardized way of encoding these aspects. There have been attempts to introduce variability in the input by randomly degrading the quality of the input utterances, but, the representativeness of these approaches for the variation in actual speech is questionable (see Section 5.1 for a discussion).

9.7 Summary

In a number of incremental steps spread over the preceding three chapters, this thesis has described, the first incremental and multiple-cue combination model of segmentation that performs competitively with the state of the art segmentation models that use a language modeling strategy (such as the LM introduced in Section 5.5). In the present chapter I have provided a summary of what was presented before, brought the results obtained previously into a unified scheme, analyzed a number of issues that were delayed during the presentation of the individual models, and suggested possible future improvements to the models described here. Now it is time to have the general summary of thesis and to conclude.

10 Conclusions

I would have written far less, but I did not have the time.

Blaise Pascal

This study has mainly developed computational models of language acquisition and tested these models using computational simulations on realistic samples that children receive during language acquisition. The first two chapters of this thesis surveyed the broader field of language acquisition, focusing mainly on the formal methods of studying language acquisition. The particular aspect of language acquisition that is at the focus of this study is segmentation. After an introduction to the segmentation problem, the remainder of the thesis described a general strategy for segmentation, and reported results from the computational simulations of models following this strategy.

Chapter 2 introduced a central debate in the field of language acquisition, the nature–nurture debate, and a number of influential theories of language acquisition in relation to this debate. The discussion in this chapter led to the conclusion that based on the evidence available from language acquisition, we are far from concluding this age-old debate. Furthermore, most of the positions on both sides of the debate seems to be fuzzy, and very difficult or impossible to conclude. More importantly, the benefit of placing this debate at the center of the research agenda is questionable, and often, seems to be unfruitful, or even counterproductive.

In Chapter 3, the formal modeling practice for language acquisition research was introduced, and its strengths and weaknesses were identified. This chapter also provided a review of relevant studies from computational learning theory. Some common misconceptions in the language acquisition literature stemming from results of certain studies in computational learning theory were pointed out. The formal, analytical approach typically used in the computational learning theory literature is one of the approaches to studying computational models of learning. Another common method often employed in the study of cognitive processes is to use computational simulations. I argued that the approach taken in this thesis, the computational simulations, sometimes provide easier modeling opportunities than are available to formal analysis methods typically used in the field of computational learning theory. The input to the language learner is an example where computational simulations allow more realistic and more straightforward modeling options.

Most studies outlined in these two chapters focus on learning syntax. Children's acquisition of rules governing formation of sentences is indeed an interesting subject. However, learning syntax, as Miller (1996, p.238) puts it, is only 'slightly more amazing' than learning new words. The focus of this study is one of the first steps children need to take for learning words. The question of interest is: how do children extract lexical units, e.g., words, from a continuous stream of speech sounds without knowing which sound sequences are words? This problem, the segmentation problem, and what we know from developmental psycholinguistics about how children deal with this problem was reviewed in Chapter 4. In a nutshell, children seem to use several cues that are partial, noisy, overlapping, and sometimes conflicting indications to find word boundaries and words in continuous speech. These cues include predictability statistics, phonotactics, lexical knowledge, and lexical stress.

After discussing the segmentation problem from a computational perspective, the related work in this subfield was reviewed. Chapter 5 presented issues regarding evaluating the computational models of segmentation, pointing out a few confusions in interpreting the results of these models, and described two new error measures that may serve to assess the performance of segmentation models better. Next, a reference model was described. The reference model shares a common strategy, the 'language modeling' strategy with most successful computational models of segmentation. However, it does not model the child language acquisition process closely.

The following three chapters presented computational models of segmentation that are compatible with what we know about child language acquisition. These chapters define models that use the cues listed above, leading to a combined model which uses all the cues. Chapter 9 provided an overview and comparison of the models developed in preceding three chapters.

In the development of the framework that was shared by all these models, particular care has been taken that the work be compatible with what we know about the segmentation of speech by children from psycholinguistic research. First, as the psycholinguistic studies reviewed in Chapter 4 indicate, adults and children use multiple cues in segmentation task. Following this fact about human language acquisition, the model integrates multiple cues. Furthermore, it starts segmenting input utterances with language-general cues, and learns to use cues that are useful only after learning words of the input language.

Second, the segmentation decisions are incremental. Unlike computational models that require the complete utterance (or the complete corpus) to be presented before deciding for the best segmentation, the models presented here decide on boundaries while processing the utterance from left to right. In accordance with the results of some of the psycholinguistic studies, a certain amount of right context is used. However, the models presented here do not exhaustively search for best segmentation. The incremental nature of the segmentation algorithm also indicates that the computational resources required by the models are modest.

A third aspect, which can further be improved, is the input. As in most psychologically motivated computational models of segmentation, all simulations are run on phonemically transcribed child directed speech. The phonemic transcription necessarily introduces some idealizations, and it removes some of the cues available in the speech sound. However, these limitations are practical. The model can trivially be extended to make use of more realistic input.

The results of the simulations are encouraging. The segmentation performance, measured using performance scores used in the related literature indicates that the combined model is competitive with the state of the art segmentation models without similar levels of fidelity to what is known about child language acquisition.

As far as I can determine, the combined model presented here is the first model following the child acquisition process as faithfully as it does, with this level of segmentation performance. Besides the improvements in the performance, another advantage of the model presented in this thesis in comparison to previous models, such as connectionist systems, is that it uses an explicitly specified statistical model of learning. As a result, it allows easier interpretations of what the model is learning, and easier extensions where necessary.

The segmentation model presented in this thesis demonstrates a way to achieve good segmentation performance using more plausible segmentation strategies. However, this is only the beginning. As discussed in Chapter 9 at depth, there are many ways to improve the model. Of these improvements, two of them stand out. The first one is to use better cue combination mechanisms, and the second one is the use of more realistic input.

Bibliography

Abney, Steven. 1996. Statistical Methods and Linguistics. In Judith Klavans and Philip Resnik (eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, Chapter 1, Cambridge, MA: The MIT Press.

Aitchison, Jean. 1994. *Words in the mind: an introduction to the mental lexicon*. Malden, MA: Blackwell.

Akmajian, Adrian, Demers, Richard A., Farmer, Ann K. and Harnish, Robert M. 2010. *Linguistics: An Introduction to Language and Communication*. Cambridge, MA: The MIT Press, sixth edition.

Al-Shalabi, Riyad, Kannan, Ghassan, Hilat, Iyad, Ababneh, Ahmad and Al-Zubi, Ahmad. 2005. Experiments with the Successor Variety Algorithm Using the Cutoff and Entropy Methods. *Information Technology Journal* 4(1).

Allen, Joe and Christiansen, Morten H. 1996. Integrating Multiple Cues in Word Segmentation: A Connectionist Model using Hints. In *Proceedings of the Eighteenth Annual Cognitive Science Society Conference*, pages 370–375.

Ando, Rie Kubota and Lee, Lillian. 2003. Mostly-Unsupervised Statistical Segmentation of Japanese Kanji Sequences. *Natural Language Engineering* 9(2), 127–149.

Angluin, Dana. 1982. Inference of Reversible Languages. *Journal of the Association for Computing Machinery* 29(3), 741–765.

Angluin, Dana. 1987. Learning regular sets from queries and counterexamples. *Information and Computation* 75(2), 87–106.

Angluin, Dana. 1988a. Identifying Languages From Stochastic Examples. Technical Report YALE/DCS/TR614, Yale University, Department of Computer Science.

Angluin, Dana. 1988b. Queries and Concept Learning. *Journal Machine Learning* 2(4), 319–342.

Aslin, Richard N. 1993. Segmentation Of Fluent Speech Into Words: Learning Models And The Role Of Maternal Input. In B. De Boysson-Bardies, Scania de Schonen, Peter Jusczyk, Peter MacNeilage and John Morton (eds.), *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*, pages 305–315, Kluwer Academic Publishers.

Aslin, Richard N., Saffran, Jenny R. and Newport, Elissa L. 1998. Computation of Conditional Probability Statistics by 8-month-old Infants. *Psychological Science*

9(4), 321–324.

Aslin, Richard N., Woodward, Julide Z., LaMendola, Nicholas P. and Bever, Thomas G. 1996. Models of Word Segmentation in Fluent Maternal Speech to Infants. In James L. Morgan Katherine Demuth (ed.), *Signal to Syntax: Bootstrapping From Speech to Grammar in Early Acquisition*, Chapter 8, pages 117–134, Lawrence Erlbaum Associates.

Baker, Mark C. 2001. *The Atoms of Language: The Mind's Hidden Rules of Grammar*. New York: Basic Books.

Bannard, Colin and Matthews, Danielle. 2008. Stored Word Sequences in Language Learning. *Psychological Science* 19(3), 241–248.

Batchelder, Eleanor Olds. 2002. Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition* 83, 167–206.

Bernstein Ratner, Nan. 1987. The phonology of parent-child speech. In K. Nelson and A. van Kleeck (eds.), *Children's language*, volume 6, pages 159–174, Hillsdale, NJ: Erlbaum.

Bijeljac-Babic, Ranka, Bertoncini, Josiane and Mehler, Jacques. 1993. How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology* 29(4), 711–721.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Blanchard, Daniel, Heinz, Jeffrey and Golinkoff, Roberta. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language* 37(Special Issue 03), 487–511.

Bloom, Lois. 1976. *One Word at a Time: The Use of Single Word Utterances Before Syntax*. Mouton de Gruyter.

Bloom, Paul. 2000. *How Children Learn the Meanings of Words*. Cambridge, MA: The MIT Press.

Bloom, Paul and Markson, Lori. 1998. Capacities underlying word learning. *Trends in Cognitive Sciences* 2(2), 67–73.

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.

Boeckx, Cedric. 2009. *Language in Cognition: Uncovering Mental Structures and the Rules Behind Them*. Oxford: Wiley-Blackwell.

Boland, Philip J. 1989. Majority Systems and the Condorcet Jury Theorem. *Journal of the Royal Statistical Society. Series D (The Statistician)* 38(3), 181–189.

Bordag, Stefan. 2005. Unsupervised knowledge-free morpheme boundary detection. In *The Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Bordag, Stefan. 2007. Unsupervised and Knowledge-free Morpheme Segmentation and Analysis. In *The Working Notes for the CLEF Workshop 2007*.

Bortfeld, Heather, Morgan, James L., Golinkoff, Roberta Michnick and Rathbun, Karen. 2005. Mommy and Me: Familiar Names Help Launch Babies Into Speech-Stream Segmentation. *Psychological Science* 16, 298–304.

Box, George E. P. and Draper, Norman R. 1986. *Empirical Model-Building and Response Surfaces*. New York, USA: John Wiley & Sons, Inc.

Brent, Michael R. 1996. Advances in the computational study of language acquisition. *Cognition* 61, 1–38.

Brent, Michael R. 1999a. An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning* 34(1-3), 71–105.

Brent, Michael R. 1999b. Speech segmentation and word discovery: a computational perspective. *Trends in Cognitive Sciences* 3(8), 294–301.

Brent, Michael R. and Cartwright, Timothy A. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61, 93–125.

Brent, Michael R. and Siskind, Jeffrey Mark. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition* 81, B33–B44.

Bresnan, Joan. 2001. *Lexical Functional Syntax*. Blackwell.

Bresnan, Joan, Kaplan, Ronald M., Peters, Stanley and Zaenen, Annie. 1982. Cross-Serial Dependencies in Dutch. *Linguistic Inquiry* 13(4), 613–635.

Brown, Roger. 1973. *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Cairns, Paul, Shillcock, Richard, Chater, Nick and Levy, Joe. 1994. Modelling the acquisition of lexical segmentation. In *Proceedings of the 26th Child Language Research Forum*, University of Chicago Press.

Carey, Susan. 1978. The child as word learner. In Morris Halle, Joan Bresnan and George A. Miller (eds.), *Linguistic theory and psychological reality*, Chapter 8, pages 264–293, Cambridge, MA: The MIT Press.

Carnegie Mellon University. 1998. The Carnegie Mellon University Pronouncing Dictionary. `http://http://www.speech.cs.cmu.edu/cgi-bin/cmudict`, version 0.6d.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, Noam. 1959a. On certain formal properties of grammars. *Information and Control* 2(2), 137–167.

Chomsky, Noam. 1959b. A Review of B. F. Skinner's Verbal Behavior. *Language* 35(1), 26–58.

Chomsky, Noam. 1965. *Aspects of Theory of Syntax*. Cambridge, MA: The MIT Press.

Chomsky, Noam. 1980. *Rules and Representations*. New York: Columbia University Press.

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht, NL: Foris Publications.

Chouinard, Michelle M. and Clark, Eve V. 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language* 30(3), 637–669.

Christiansen, Morten H. and Allen, Joseph. 1997. Coping with variation in speech segmentation. In C. Heycock A. Sorace and R. Shillcock (eds.), *Proceedings of GALA 1997: Language Acquisition: Knowledge Representation and Processing*, page 327–332, University of Edinburgh Press.

Christiansen, Morten H., Allen, Joseph and Seidenberg, Mark S. 1998. Learning to Segment Speech Using Multiple Cues: A Connectionist Model. *Language and Cognitive Processes* 13(2), 221–268.

Christiansen, Morten H., Conway, Christopher M. and Curtin, Suzanne. 2005. Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. In J.W. Minett and W.S.-Y. Wang (eds.), *Language acquisition, change and emergence: Essays in evolutionary linguistics*, Chapter 5, pages 205–249, Hong Kong: City University of Hong Kong Press.

Christophe, Anne, Dupoux, Emmanuel, Bertoncini, Josiane and Mehler, Jacques. 1994. Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *The Journal of the Acoustical Society of America* 95(3), 1570–1580.

Church, Kenneth W. 1987. Phonological parsing and lexical retrieval. *Cognition* 25(1-2), 53–69.

Clark, Alexander. 2010. Efficient, correct, unsupervised learning of context-sensitive languages. In *Proceedings of CoNLL*, Association for Computational Linguistics.

Clark, Alexander and Eyraud, Rémi. 2006. Learning Auxiliary Fronting with Grammatical Inference. In *Proceedings of CoNLL*, pages 125–132, New York.

Clark, Alexander and Eyraud, Rémi. 2007. Polynomial Identification in the Limit of Substitutable Context-free Languages. *Journal of Machine Learning Resesarch* 8, 1725–1745.

Clark, Alexander and Lappin, Shalom. 2011a. Computational learning theory and language acquisition. In R. Kempson, N. Asherw and T. Fernando (eds.), *Handbook of Philosophy of Linguistics*, Elsevier, forthcoming.

Clark, Alexander and Lappin, Shalom. 2011b. *Linguistic Nativism and the Poverty of the Stimulus*. Oxford: Wiley-Blackwell.

Clark, Alexander and Thollard, Franck. 2004. Partially Distribution-Free Learning of Regular Languages from Positive Samples. In *Proceedings of COLING*, pages 85–91.

Clark, Eve V. 1993. *The Lexicon in Acquisition*. Cambridge Studies in Linguistics, Cambridge, UK: Cambridge University Press.

Clark, Stephen and Curran, James R. 2003. Log-Linear Models for Wide-Coverage CCG Parsing. In *Proceedings of the SIGDAT Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pages 97–104.

Cohen, Paul, Adams, Niall and Heeringa, Brent. 2007. Voting experts: An unsupervised algorithm for segmenting sequences. *Intelligent Data Analysis* 11(6), 607–625.

Cole, Ronald A. 1973. Listening for mispronunciations: A measure of what we hear during speech. *Perception and Psychophysics* 13(1), 153–156.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph. D.thesis, University of Pennsylvania.

Çöltekin, Çağrı. 2010. Improving Successor Variety for Morphological Segmentation. In *Proceedings of the 20th Meeting of Computational Linguistics in the Netherlands*.

Coltheart, Max. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology* 33A(4), 497–505.

Costa Florêncio, Cristophe. 2003. *Learning Categorial Grammars*. Ph. D.thesis, Univesiteit Utrecht.

Cowie, Fiona. 2010. Innateness and Language. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Stanford University, summer 2010 edition.

Crain, Stephen and Nakayama, Mineharu. 1987. Structure Dependence in Grammar Formation. *Language* 63(3), 522–543.

Crain, Stephen and Pietroski, Paul. 2002. Why language acquisition is a snap. *The Linguistic Review* 19(1-2), 163–183.

Croft, William. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. New York: Oxford University Press.

Curtiss, Susan. 1977. *Genie: A Psycholinguistic Study of a Modern-Day "Wild Child"*. Perspectives in neurolinguistics and psycholinguistics, Academic Press.

Cutler, Anne. 1996. Prosody and the word boundary problem. In James L. Morgan Katherine Demuth (ed.), *Signal to Syntax: Bootstrapping From Speech to Grammar in Early Acquisition*, Chapter 6, pages 87–99, Lawrence Erlbaum Associates.

Cutler, Anne and Butterfield, Sally. 1990. Durational cues to word boundaries in clear speech. *Speech Communication* 9(5-6), 485–495.

Cutler, Anne and Butterfield, Sally. 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language* 31(2), 218–236.

Cutler, Anne and Carter, David M. 1987. The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language* 2(3-4), 133–142.

Cutler, Anne and Mehler, Jacques. 1993. The periodicity bias. *Journal of Phonetics* 21, 103–108.

Cutler, Anne, Mehler, Jacques, Norris, Dennis and Segui, Juan. 1986. The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language* 25(4), 385–400.

Dahan, Delphine and Brent, Michael R. 1999. On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General* 128(2), 165–185.

Dahan, Delphine and Magnuson, James S. 2006. Spoken Word Recognition. In *Handbook of Psycholinguistics*, Chapter 8, pages 249–283, Elsevier, second edition.

Dauer, Rebecca M. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* 11, 51–62.

Davis, Martin D., Sigal, Ron and Weyuker, Elaine J. 1994. *Computability, complexity, and languages: fundamentals of theoretical computer science*. Boston: Academic

Press, Harcourt, Brace.

Davis, Matthew Harold. 2006. *Lexical Segmentation in Spoken Word Recognition*. Ph. D.thesis, Birkbeck College, University of London.

de Marcken, Carl. 1996. Linguistic structure as composition and perturbation. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 335–341, Morristown, NJ, USA: Association for Computational Linguistics.

Déjean, Hervé. 1998. Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora. In *Workshop on Paradigms and Grounding in Natural Language Learning*, pages 295–299.

Demberg, Vera. 2007. A Language-Independent Unsupervised Model for Morphological Segmentation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 920–927, Prague, Czech Republic: Association for Computational Linguistics.

Demuth, Katherine, Culbertson, Jennifer and Alter, Jennifer. 2006. Word-minimality, Epenthesis and Coda Licensing in the Early Acquisition of English. *Language and Speech* 49(2), 137–173.

Dickersin, Kay. 1990. The Existence of Publication Bias and Risk Factors for Its Occurrence. *Journal of the American Medical Association* 263(10), 1385–1389.

Dilley, Laura C. and McAuley, J. Devin. 2008. Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language* 59(3), 294–311.

Dittmar, Miriam, Abbot-Smith, Kirsten, Lieven, Elena and Tomasello, Michael. 2008. German Children's Comprehension of Word Order and Case Marking in Causative Sentences. *Child Development* 79(4), 1152–1167.

Elman, Jeff L. 1990. Finding Structure in Time. *Cognitive Science* 14, 179–211.

Elman, Jeffrey L. 1991. Distributed Representations, Simple Recurrent Networks, And Grammatical Structure. *Machine Learning* 7(2-3).

Elman, Jeffrey L. 2005. Connectionist models of cognitive development: where next? *Trends in Cognitive Sciences* 9(3), 111–117.

Elman, Jeffrey L., Bates, Elizabeth A., Johnson, Mark H., Karmiloff-Smith, Annette, Parisi, Domenico and Plunkett, Kim. 1996. *Rethinking Innateness*. Cambridge, MA: The MIT Press.

Fenson, Larry, Dale, Philip S., Reznick, J. Steven, Bates, Elizabeth, Thal, Donna J. and Pethick, Stephen J. 1994. Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development* 59, with commentary by Michael Tomasello, Carolyn B. Mervis, Joan Stile.

Fernald, Anne, Taeschner, Traute, Dunn, Judy, Papouseka, Mechthild, de Boysson-Bardies, Bénédicte and Fukui, Ikuko. 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language* 16(3), 477–501.

Fiser, József and Aslin, Richard N. 2002. Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America* 99(24), 15822–15826.

Fleck, Margaret M. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL-08)*, pages 130–138.

Fodor, Janet Dean. 1998. Unambiguous Triggers. *Linguistic Inquiry* 29(1), 1–36.

Fougeron, Cecile and Keating, Patricia A. 1997. Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America* 101(6), 3728–3740.

Francis, W. Nelson and Kucera, Henry. 1979. Brown Corpus Manual. Technical Report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Frigg, Roman and Hartmann, Stephan. 2009. Models in Science. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Stanford University, summer 2009 edition.

Gambell, Timothy and Yang, Charles. 2006. Word segmentation: Quick but not dirty. Unpublished manuscript, available at `http://www.ling.upenn.edu/ycharles/papers/quick.pdf`.

Ganger, Jennifer and Brent, Michael R. 2004. Reexamining the Vocabulary Spurt. *Developmental Psychology* 40(4), 621–632.

Ganong, William F. 1980. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance* 6(1), 110–125.

Gervain, Judit and Mehler, Jacques. 2010. Speech Perception and Language Acquisition in the First Year of Life. *Annual Review of Psychology* 61(1), 191–218.

Ghahramani, Zoubin, Wolpert, Daniel M. and Jordan, Michael I. 1997. Computational Models of Sensorimotor Integration. In Pietro Morasso and Vittorio Sanguineti (eds.), *Self-Organization, Computational Maps and Motor Control*, volume 119 of *Advances in Psychology*, pages 117–147, North-Holland.

Gibson, Edward and Wexler, Kenneth. 1994. Triggers. *Linguistic Inquiry* 25(3), 407–454.

Gold, E. Mark. 1967. Language identification in the limit. *Information and Control* 10(5), 447–474.

Goldberg, Adele. 2006. *Constructions at Work: The Nature of Generalization in Language*. New York: Oxford University Press.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2), 153–198.

Goldsmith, John. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering* 12(04), 353–371.

Goldwater, Sharon. 2006. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph. D.thesis, Brown University.

Goldwater, Sharon, Griffiths, Thomas L. and Johnson, Mark. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112, 21–54.

Graf Estes, Katharine, Evans, Julia L., Alibali, Martha W. and Saffran, Jenny R. 2007. Can Infants Map Meaning to Newly Segmented Words? Statistical Segmentation and Word Learning. *Psychological Science* 18(3), 254–260.

Greenberg, J.H. and Jenkins, J.J. 1964. Studies in the psychological correlates of the sound system of American English. *Word* 20, 157–177.

Guasti, Maria Teresa. 2002. *Language Acquisition: The Growth of Grammar*. Cambridge, MA: The MIT Press.

Hafer, Margaret A. and Weiss, Stephen F. 1974. Word Segmentation by Letter Successor Varieties. *Information Storage and Retrieval* 10(11-12), 371–385.

Halevy, Alon, Norvig, Peter and Pereira, Fernando. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 24(2), 8–12.

Hansen, Lars Kai and Salamon, Peter. 1990. Neural Network Ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* 12, 993–1001.

Harrington, Jonathan, Palethorpe, Sallyanne and Watson, Catherine I. 2000. Does the Queen speak the Queen's English?. *Nature* 408(6815), 927.

Harris, John and Lindsey, Geoff. 1995. The elements of phonological representation. In Jacques Durand and Francis Katamba (eds.), *Frontiers of phonology: atoms, structures, derivations*, pages 34–79, Harlow, Essex: Longman.

Harris, Zellig. 1951. *Methods in Structural Linguistics*. University of Chicago Press.

Harris, Zellig S. 1955. From Phoneme to Morpheme. *Language* 31(2), 190–222.

Hauser, Marc D., Chomsky, Noam and Fitch, W. Tecumseh. 2002. The faculty of language: what is it, who has it, and how did it evolve? *Science* 298(5598), 1569–1579.

Hauser, Marc D., Newport, Elissa L. and Aslin, Richard N. 2001. Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition* 78(3), B53–B64.

Hendriks, Petra, Siekman, Irene, Smits, Erik-Jan and Spenader, Jennifer. 2007. Pronouns in competition: Predicting acquisition delays cross-linguistically. In Dagmar Bittner and Natalia Gagarina (eds.), *ZAS Papers in Linguistics, Volume 48 (Intersentential Pronominal Reference in Child and Adult Language. Proceedings of the Conference on Intersentential Pronominal Reference in Child and Adult Language)*, pages 75–101.

Higginson, Roy Patrick. 1985. *Fixing-assimilation in language acquisition*. Ph. D.thesis, Washington State University.

Hockema, Stephen A. 2006. Finding Words in Speech: An Investigation of American English. *Language Learning and Development* 2(2), 119–146.

Hollich, George, Newman, Rochelle S. and Jusczyk, Peter W. 2005. Infants' Use of Synchronized Visual Information to Separate Streams of Speech. *Child Development* 76(3), 598–613.

Hopcroft, John E., Motwahl, Rajeev and Ullman, Jeffrey D. 2001. *Introduction to automata theory, languages, and computation*. Addison-Wesley, second edition.

Horning, James Jay. 1969. *A study of grammatical inference*. Ph. D.thesis, Computer Science Department, Stanford University.

Huang, Jin Hu and Powers, David. 2003. Chinese Word Segmentation based on Contextual Entropy. In *Proceedings of Pacific Asia Conference on Language, Information and Computation*, pages 121–127.

Hubel, D. H. and Wiesel, T. N. 1970. The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *The journal of Physiology* 206, 419–436.

Jain, Sanjay, Osherson, Daniel, Royer, James S. and Sharma, Arun. 1999. *Systems That Learn: an introduction to learning theory*. Cambridge, MA: The MIT Press, second edition.

Johnson, Elizabeth K. and Jusczyk, Peter W. 2001. Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics. *Journal of Memory and Language* 44(4), 548–567.

Johnson, Elizabeth K. and Seidl, Amanda H. 2009. At 11 months, prosody still outranks statistics. *Developmental Science* 12(1), 131–141.

Jones, Peter E. 1995. Contradictions and Unanswered Questions in the Genie Case: A Fresh Look at the Linguistic Evidence. *Language and Communication* 15(3), 261–280.

Jurafsky, Daniel and Martin, James H. 2008. *Speech and Language Processing*. Prentice Hall, second edition.

Jusczyk, Peter W. 1999. How infants begin to extract words from speech. *Trends in Cognitive Sciences* 3(9), 323–328.

Jusczyk, Peter W., Friederici, Angela D., Wessels, Jeanine M. I., Svenkerud, Vigdis Y. and Jusczyk, Ann Marie. 1993. Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language* 32(3), 402–420.

Jusczyk, Peter W., Hohne, Elizabeth A. and Bauman, Angela. 1999a. Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics* 61(8), 1465–1476.

Jusczyk, Peter W., Houston, Derek M. and Newsome, Mary. 1999b. The Beginnings of Word Segmentation in English-Learning Infants. *Cognitive Psychology* 39, 159–207.

Jusczyk, Peter W., Kennedy, Lori J. and Jusczyk, Ann Marie. 1995. Young infants' retention of information about syllables. *Infant Behavior and Development* 18(1), 27–41.

Jusczyk, Peter W., Luce, Paul A. and Charles-Luce, Jan. 1994. Infants' Sensitivity to Phonotactic Patterns in the Native Language. *Journal of Memory and Language* 33(5), 630–645.

Jusczyk, Peter W. and Thompson, Elizabeth. 1978. Perception of a phonetic contrast in multisyllabic utterances by 2-month-old infants. *Perception and Psychophysics* 23(2), 105–109.

Kanazawa, Makoto. 1996. Identification in the limit of categorial grammars. *Journal of Logic, Language and Information* 5(2), 115–155.

Kearns, Michael J. and Vazirani, Umesh V. 1994. *An introduction to computational learning theory*. Cambridge, MA: The MIT Press.

Kempe, André. 1999. Experiments in Unsupervised Entropy-Based Corpus Segmentation. In *Proc. Workshop on Computational Natural Language Learning (CoNLL'99)*, pages 7–13, Bergen, Norway.

Kirkham, Natasha Z., Slemmer, Jonathan A. and Johnson, Scott P. 2002. Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition* 83(2), B35–B42.

Klein, Dan and Manning, Christopher D. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *In Proceedings of the 42nd Annual Meeting of the ACL*, pages 479–486.

Klein, Dan and Manning, Christopher D. 2005. Natural language grammar induction with a generative constituent-context model. *Pattern Recognition* 38(9), 1407–1419.

Knuth, Donald. 1997. *The Art of Computer Programming, Volume 3: Sorting and Searching*. Addison-Wesley, third edition.

Kording, Konrad P., Beierholm, Ulrik, Ma, Wei Ji, Quartz, Steven, Tenenbaum, Joshua B. and Shams, Ladan. 2007. Causal Inference in Multisensory Perception. *PLoS ONE* 2(9), e943.

Korman, Myron. 1984. Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First Language* 5, 44–45.

Krueger, Kai A. and Dayan, Peter. 2009. Flexible shaping: How learning in small steps helps. *Cognition* 110(3), 380–394.

Landy, Michael S., Maloney, Laurence T., Johnston, Elizabeth B. and Young, Mark. 1995. Measurement and Modeling of Depth Cue Combination: in Defense of Weak Fusion. *Vision Research* 35(3), 389–412.

Lappin, Shalom and Shieber, Stuart M. 2007. Machine Learning Theory and Practice as a Source of Insight into Universal Grammar. *Journal of Linguistics* 43(2), 393–427.

Lenneberg, Eric H. 1967. *Biological foundations of language*. New York: John Wiley and Sons.

Littlestone, Nick and Warmuth, Manfred K. 1994. The Weighted Majority Algorithm. *Information and Computation* 108(2), 212–261.

MacWhinney, Brian. 1982. Basic syntactic processes. In S. Kuczaj (ed.), *Language development (Vol. 1): Syntax and semantics*, pages 73–136, Hillsdale, NJ: Lawerence Erlbaum.

MacWhinney, Brian. 2010. Computational models of child language learning: an introduction. *Journal of Child Language* 37(Special Issue 03), 477–485.

MacWhinney, Brian and Snow, Catherine. 1985. The child language data exchange system. *Journal of Child Language* 12(2).

Marcus, Gary F. 1993. Negative evidence in language acquisition. *Cognition* 46(1), 53–85.

Marcus, Gary F., Pinker, Steven, Ullman, Michael, Hollander, Michelle, Rosen, T. John and Xu, Fei. 1992. *Overregularization in Language Acquisition*. University Of Chicago Press.

Markman, Ellen M. 1989. *Categorization and naming in children*. Cambridge, MA: The MIT Press.

Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Marslen-Wilson, William D. 1987. Functional parallelism in spoken word-recognition. *Cognition* 25(1-2), 71–102.

Marslen-Wilson, William D. and Welsh, Alan. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology* 10(1), 29–63.

Mattys, Sven L. 2004. Stress Versus Coarticulation: Toward an Integrated Approach to Explicit Speech Segmentation. *Journal of Experimental Psychology: Human Perception and Performance* 30(2), 397–408.

Mattys, Sven L., Melhorn, James F. and White, Laurence. 2007. Effects of Syntactic Expectations on Speech Segmentation. *Journal of Experimental Psychology: Human Perception and Performance* 33(4), 960–977.

Mattys, Sven L., White, Laurence and Melhorn, James F. 2005. Integration of Multiple Speech Segmentation Cues: A Hierarchical Framework. *Journal of Experimental Psychology: General* 134(4), 477–500.

Mazuka, Reiko. 2007. The rhythm-based prosodic bootstrapping hypothesis of early language acquisition: Does it work for learning for all languages? *Journal of the Linguistic Society of Japan* 132, 1–13.

McClelland, James L. and Elman, Jeffrey L. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18(1), 1–86.

McGilvray, James. 2006. On the Innateness of Language. In Robert J. Stainton (ed.), *Contemporary Debates in Cognitive Science*, Contemporary Debates in Philosophy, Chapter 5, pages 97–112, Oxford: Wiley-Blackwell.

Mehler, Jacques, Jusczyk, Peter, Lambertz, Ghislaine, Halsted, Nilofar, Bertoncini, Josiane and Amiel-Tison, Claudine. 1988. A precursor of language acquisition in young infants. *Cognition* 29(2), 143–178.

Miller, George A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63(2), 81–97.

Miller, George A. 1996. *The science of words*. New York: Scientific American Library.

Miller, George A. 2003. The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences* 7(3), 141–144.

Monaghan, Padraic and Christiansen, Morten H. 2010. Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *Journal of Child Language* 37(Special Issue 03), 545–564.

Moon, Christine, Cooper, Robin Panneton and Fifer, William P. 1993. Two-day-olds prefer their native language. *Infant Behavior and Development* 16(4), 495–500.

Narasimhamurthy, Anand. 2005. Theoretical Bounds of Majority Voting Performance for a Binary Classification Problem. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1988–1995.

Nazzi, Thierry, Iakimova, Galina, Bertoncini, Josiane, Frédonie, Séverine and Alcantara, Carmela. 2006. Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences. *Journal of Memory and Language* 54(3), 283–299.

Nespor, Marina and Vogel, Irene. 1986. *Prosodic Phonology*. Dordrecht, NL: Foris Publications.

Newmeyer, Frederick J. 2004. Against a parameter-setting approach to typological variation. *Linguistic Variation Yearbook* 4(1), 181–234.

Newmeyer, Frederick J. 2006. A rejoinder to ‘On the role of parameters in Universal Grammar: A reply to Newmeyer’ by Ian Roberts & Anders Holmberg, unpublished manuscript, lingBuzz/000248.

Newport, Elissa L. 1988. Constraints on learning and their role in language acquisition: Studies of the acquisition of American sign language. *Language Sciences* 10(1), 147–172.

Newport, Elissa L. 1990. Maturational constraints on language learning. *Cognitive Science* 14, 11–28.

Newport, Elissa L. 1993. Maturational constraints on language learning. In Paul Bloom (ed.), *Language Acquisition*, pages 543–560, Cambridge, MA: The MIT Press.

Newport, Elissa L. and Aslin, Richard N. 2004. Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology* 48(2), 127–162.

Nickerson, Raymond S. 1997. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2(2), 175–220.

Niyogi, Partha. 2006. *The Computational Nature of Language Learning and Evolution*. Cambridge, MA: The MIT Press.

Norris, Dennis. 1994. Shortlist: a connectionist model of continuous speech recognition. *Cognition* 52(3), 189–234.

Nowak, Martin A. and Komarova, Natalia L. 2001. Towards an evolutionary theory of language. *Trends in Cognitive Sciences* 5(7), 288–295.

Nowak, Martin A., Komarova, Natalia L. and Niyogi, Partha. 2001. Evolution of Universal Grammar. *Science* 291, 114–118.

Nowak, Martin A., Komarova, Natalia L. and Niyogi, Partha. 2002. Computational and evolutionary aspects of language. *Nature* 417(6889), 611.

Nowak, Martin A., Plotkin, Joshua B. and Jansen, Vincent A. A. 2000. The evolution of syntactic communication. *Nature* 404, 495–498.

Olivier, Donald C. 1968. *Stochastic grammars and language acquisition mechanisms*. Ph. D.thesis, Harvard University.

Omar, Margret K. 1973. *The acquisition of Egyptian Arabic as a native language*. The Hague, NL: Mouton.

Osherson, Daniel N., Stob, Michael and Weinstein, Scott. 1984. Learning theory and natural language. *Cognition* 17(1), 1–28.

Pelucchi, Bruna, Hay, Jessica F. and Saffran, Jenny R. 2009. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition* 113(2), 244–247.

Perruchet, Pierre and Desaulty, Stéphane. 2008. A role for backward transitional probabilities in word segmentation? *Memory and Cognition* 36(7), 1299–1305.

Perruchet, Pierre and Vinter, Annie. 1998. PARSER: A Model for Word Segmentation. *Journal of Memory and Language* 39(2), 246–263.

Pike, Kenneth L. 1945. *The intonation of American English*. Ann Arbor: University of Michigan Press.

Pinker, Steven. 1989. *Learnability and cognition: the acquisition of argument structure*. Cambridge, MA: The MIT Press.

Pinker, Steven. 1994. *The language instinct: the new science of language and mind*. Penguin Books.

Pinker, Steven. 1995. Language Acquisition. In Lila R. Gleitman (ed.), *An Invitation to Cognitive Science, Vol. 1: Language*, Chapter 6, pages 135–182, Cambridge, MA: The MIT Press.

Polka, Linda and Sundara, Megha. 2003. Word segmentation in monolingual and bilingual infant learners of English and French. In M. J. Solé, D. Recasens and J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 1021–1024.

Pollard, Carl and Sag, Ivan A. 1987. *Information-based Syntax and Semantics, volume 1: Fundamentals*. Stanford: CSLI Publications.

Pons, Ferran. 2006. The effects of distributional learning on rats' sensitivity to phonetic information. *Journal of Experimental Psychology: Animal Behavior Processes* 32(1), 97–101.

Prince, Alan and Smolensky, Paul. 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell.

Pullum, Geoffrey K. and Scholz, Barbara C. 2002. Empirical assessment of stimulus poverty arguments. *Linguistic Review* 19(1/2), 9.

Pullum, Geoffrey K. and Scholz, Barbara C. 2010. Recursion and the infinitude claim. In Harry van der Hulst (ed.), *Recursion in Human Language*, Studies in Generative Grammar, No. 104, pages 113–138, Berlin: Mouton de Gruyter.

Quine, Willard Van Orman. 1960. *Word and object*. Cambridge, MA: The MIT Press.

Ramus, Franck. 2002. Acoustic correlates of linguistic rhythm: Perspectives. In *Proceedings of Speech Prosody 2002*, pages 115–120.

Reznick, J. Steven and Goldfield, Beverly A. 1992. Rapid change in lexical development in comprehension and production. *Developmental Psychology* 28(3), 406–413.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14(5), 465–471.

Roach, Peter. 1982. On the distinction between "stress-timed" and "syllable-timed" languages. In David Crystal (ed.), *Linguistic controversies: Essays in Honour of F. R. Palmer*, pages 73–79, London: Arnold.

Rollins, Pamela R., Pan, Barbara A., Conti-Ramsden, Gina and Snow, Catherine E. 1994. Communicative skills in children with specific language impairments: A comparison with their language-matched siblings. *Journal of Communication Disorders* 27(2), 189–206.

Rubin, Philip, Turvey, M. T. and Gelder, Peter Van. 1976. Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception and Psychophysics* 19(5), 394–398.

Rumelhart, David E. and McClelland, James L. 1986. On learning the past tenses of English verbs. In James L. McClelland and David E. Rumelhart (eds.), *Parallel Distributed Processing Vol 2*, pages 216–271, Cambridge, MA: The MIT Press.

Rymer, Russ. 1994. *Genie: A Scientific Tragedy*. Harper Paperbacks.

Rytting, C. Anton, Brew, Chris and Fosler-Lussier, Eric. 2010. Segmenting words from natural speech: subsegmental variation in segmental cues. *Journal of Child Language* 37(Special Issue 03), 513–543.

Saffran, Jenny R., Aslin, Richard N. and Newport, Elissa L. 1996a. Statistical learning by 8-month old infants. *Science* 274(5294), 1926–1928.

Saffran, Jenny R., Johnson, Elizabeth K., Aslin, Richard N. and Newport, Elissa L. 1999. Statistical learning of tone sequences by human infants and adults. *Cognition* 70(1), 27–52.

Saffran, Jenny R., Newport, Elissa L. and Aslin, Richard N. 1996b. Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language* 35(4), 606–621.

Sampson, Geoffrey. 1999. *Educating Eve: the 'language instinct' debate*. Open linguistics series, Cassell.

Sanders, Lisa D. and Neville, Helen J. 2000. Lexical, Syntactic, and Stress-Pattern Cues for Speech Segmentation. *Journal of Speech, Language and Hearing Research* 43(6), 1301–1321.

Sapir, Edward. 1921. *Language: An introduction to the study of speech*. New York: Harcourt, Brace and company.

Saxton, Matthew. 2010. *Child Language: Acquisition and Development*. SAGE Publications.

Schmid, Monika S. 2009. On L1 attrition and the linguistic system. In Leah Roberts, Georges Daniel Véronique, Anna Nilsson and Marion Tellier (eds.), *EUROSLA Yearbook*, volume 9, pages 212–244, John Benjamins.

Scholz, Barbara C. and Pullum, Geoffrey K. 2006. Irrational nativist exuberance. In Robert Stainton (ed.), *Contemporary Debates in Cognitive Science*, pages 59–80, Oxford: Basil Blackwell.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 379–423, 623–656.

Shieber, Stuart M. 1985. Evidence against the context-freeness of natural language.

Linguistics and Philosophy 8, 333–343.

Shillcock, Richard. 1995. Lexical Hypotheses in Continuous Speech. In Gerry T. M. Altmann (ed.), *Cognitive Models of Speech Processing*, Cambridge, MA: The MIT Press.

Shinohara, T. 1994. Rich Classes Inferrable from Positive Data: Length-Bounded Elementary Formal Systems. *Information and Computation* 108(2), 175–186.

Shukla, Mohinish, Nespor, Marina and Mehler, Jacques. 2007. An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology* 54(1), 1–32.

Siskind, Jeffrey Mark. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition* 61, 39–91.

Slis, I. H. 1970. Articulatory measurements on voiced, voiceless and nasal consonants. *Phonetica* 21, 193–210.

Snow, Catherine E. 1972. Mothers' Speech to Children Learning Language. *Child Development* 43(2), 549–565.

Soderstrom, Melanie, Blossom, Megan, Foygel, Rina and Morgan, James L. 2008. Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language* 35, 869–902.

Soderstrom, Melanie, Seidl, Amanda, Nelson, Deborah G. Kemler and Jusczyk, Peter W. 2003. The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language* 49(2), 249–267.

Solan, Zach, Horn, David, Ruppin, Eytan and Edelman, Shimon. 2005. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences* 102(33), 11629–11634.

Solovay, Robert M. and Strassen, Volker. 1977. A fast Monte-Carlo test for primality. *SIAM Journal on Computing* 6(1), 84–85.

Steedman, Mark. 2000. *The Syntactic Process*. Cambridge, MA: The MIT Press.

Stein, Benno and Potthast, Martin. 2008. Putting Successor Variety Stemming to Work. In Reinhold Decker and Hans J. Lenz (eds.), *Advances in Data Analysis*, pages 367–374, Springer.

Stoianov, Ivelin and Nerbonne, John. 2000. Exploring Phonotactics with Simple Recurrent Networks. In Frank van Eynde, Ineke Schuurman and Ness Schelkens (eds.), *Proceedings of Computational Linguistics in the Netherlands 1999*, pages 51–67.

Suomi, Kari, McQueen, James M. and Cutler, Anne. 1997. Vowel Harmony and Speech Segmentation in Finnish. *Journal of Memory and Language* 36(3), 422–444.

Swingley, Daniel. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology* 50(1), 86–132.

Tanenhaus, Michael K., Spivey-Knowlton, Michael J., Eberhard, Kathleen M. and Sedivy, Julie C. 1995. Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science* 268(5217), 1632–1634.

Templin, Mildred C. 1957. *Certain Language Skills in Children: Their Development* *and Interrelationships*. University of Minnesota Press.

Thiessen, Erik D. and Saffran, Jenny R. 2003. When Cues Collide: Use of Stress and Statistical Cues to Word Boundaries by 7- to 9-Month-Old Infants,. *Developmental Psychology* 39(4), 706–716.

Thiessen, Erik D. and Saffran, Jenny R. 2004. Infants' Acquisition of Stress-Based Word Segmentation Strategies. In *BUCLD 28: Proceedings of the 28th annual Boston University Conference on Language Development*, pages 608–619.

Thiessen, Erik D. and Saffran, Jenny R. 2007. Learning to Learn: Infants' Acquisition of Stress-Based Strategies for Word Segmentation. *Language Learning and Development* 3(1), 73–100.

Thompson, Susan P. and Newport, Elissa L. 2007. Statistical Learning of Syntax: The Role of Transitional Probability. *Language Learning and Development* 3(1), 1–42.

Tomasello, Michael. 2000. The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences* 4(4), 156–163.

Tomasello, Michael. 2001. Perceiving intentions and learning words in the second year of life. In M. Bowerman and S. C. Levinson (eds.), *Language acquisition and conceptual development*, pages 132–158, Cambridge, UK: Cambridge University Press.

Tomasello, Michael. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Tomasello, Michael. 2009. The usage-based theory of language acquisition. In Edith L. Bavin (ed.), *The Cambridge Handbook of Child Language*, Cambridge, UK: Cambridge University Press.

Trask, R. Larry. 2002. Review of The Atoms of Language: The Mind's Hidden Rules of Grammar by Mark C. Baker. *The Human Nature Review* 2, 77–81.

Tribus, M. and McIrvine, E. C. 1971. Energy and information. *Scientific American* 224, 178–184.

Valiant, Leslie G. 1984. A theory of learnable. *Communications of ACM* 27(11), 1134–1142.

van der Sandt, Rob. 2005. Gavagai with Peppers. *Speculative Grammarian (SpecGram)* CL(3), SpecGram is a 'parody science' journal of linguistics.

van Kampen, Anja, Parmaksiz, Güliz, van de Vijver, Ruben and Höhle, Barbara. 2008. Metrical and Statistical Cues for Word Segmentation: The Use of Vowel Harmony and Word Stress as Cues to Word Boundaries by 6- and 9Month-old Turkish Learners. In Anna Gavarro and M. Joao Freitas (eds.), *Language Acquisition and Development: Proceedings of GALA 2007*, pages 313–324.

van Rij, Jacolien, van Rijn, Hedderik and Hendriks, Petra. 2010. Cognitive architectures and language acquisition: A case study in pronoun comprehension. *Journal of Child Language* 37(03), 731–766.

van Rijsbergen, C. J. 1979. *Information Retrieval*. Butterworth-Heinemann, second edition.

Vapnik, V. N. and Chervonenkis, A. Ya. 1971. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications* 16(2), 264–280.

Venkataraman, Anand. 2001. A Statistical Model for Word Discovery in Transcribed Speech. *Computational Linguistics* 27(3), 351–372.

Vroomen, Jean, van Zon, Monique and de Gelder, Beatrice. 1996. Cues to speech segmentation: Evidence from juncture misperceptions and word spotting. *Memory and Cognition* 24(6), 744–755.

Warren, Richard M. 1970. Perceptual Restoration of Missing Speech Sounds. *Science* 167(3917), 392–393.

Werker, Janet F. and Tees, Richard C. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7(1), 49–63.

Wightman, Colin W., Shattuck-Hufnagel, Stefanie, Ostendorf, Mari and Price, Patti J. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America* 91(3), 1707–1717.

Wilson, Michael. 1988. MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, and Computers* 20(1), 6–10.

Wolpert, David H. and MacReady, William G. 1997. No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation* 1(1), 67–82.

Xu, Fei and Tenenbaum, Joshua B. 2007. Word learning as Bayesian inference. *Psychological Review* 114(2), 245–272.

Yang, Charles. 2002. *Knowledge and Learning in Natural Language*. New York: Oxford University Press.

Yang, Charles D. 2004. Universal Grammar, statistics or both? *Trends in Cognitive Sciences* 8(10), 451–456.

Yao, Xuchen, Ma, Jianqiang, Duarte, Sergio and Çağrı Çöltekin. 2009. An Inference-rules based Categorial Grammar Learner for Simulating Language Acquisition. In *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning*, Tilburg.

Zhikov, Valentin, Takamura, Hiroya and Manabu, Okumura. 2010. An efficient algorithm for unsupervised word segmentation with branching entropy and MDL. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 832–842, Stroudsburg, PA, USA: Association for Computational Linguistics.

Zipf, George K. 1935/1965. *The Psychobiology of Language*. Cambridge, MA: The MIT Press.

Short summary

Segmenting continuous speech into lexical units is one of the early tasks an infant needs to tackle during language acquisition. This thesis investigates this particular problem, segmentation, by means of computational modeling and simulations.

The segmentation problem is more difficult than it may be appreciated at first sight. Children need to find words in a continuous stream of speech, with no knowledge of words to start with. Fortunately, experimental studies reveal that children and adults use a number of cues in the input and simple strategies that exploit these cues in order to segment the speech. More interestingly, some of these cues are language independent, allowing a learner to segment the continuous input before knowing any words.

Two major aspects set the models presented in this thesis apart from other computational models in the literature. First, the models presented here use simple local strategies—as opposed to global optimization— that rely on cues known to be used by children, namely, predictability statistics, phonotactics and lexical stress. Second, these cues are combined using an explicit cue-combination model which can easily be extended to include more cues.

The models are tested using real-world transcribed child-directed speech. The simulation results show that the performance of individual strategies are comparable to the state-of-the-art computational models of segmentation. Furthermore, combinations of individual cues provide a consistent increase in performance. The combined model performs on a par with the reference state-of-the-art model, while while employing only mechanisms more similar to those available to humans performing the same task.

The dissertation starts with a general introduction to the problem of language acquisition, the difficulties and disagreements in the field. No work in language acquisition can be complete without mentioning the central debate of the field between nativism and empiricism. It is rare for a work in language acquisition not to take a clear side on this debate. The philosophical debate is intriguing. However, as I argue in Chapter 2 and Chapter 3, there seem to be no scientific criteria that would decide in favor of the one or or the other side, at least not until more physiological evidence becomes available. Furthermore, although it is a common practice in the field to take one of these view points a priori, it is not necessarily fruitful. The problem of language acquisition is a fascinating research topic in itself without its implications for the nature–nurture debate. Like many other questions of cognitive sciences, understanding

language acquisition better may contribute to this bigger debate. However, forming research questions which assume one of the viewpoints of this unsettled debate runs the risk of committing the fallacy 'ignoratio elenchi', begging the question of whether nature or nurture is more important for language learning. This is unlikely to contribute to our scientific understanding of the language acquisition.

As in many fields of science, modeling, particularly computational modeling, is one of the ways to study some of the questions regarding language acquisition. Modeling aids us in understanding the natural phenomena better by (1) finding parallels between the natural phenomenon and the computational model, (2) testing hypotheses that are difficult or impossible to test directly, and (3) providing more insight into the problem by describing it in detail. Computational models can be studied either by mathematical analysis, or by simulations. Chapter 3 offers a closer look at these two methods. These methods are complementary. However, computational simulations allow easier modeling practices with regard to probabilistic aspects of the phenomenon to be modeled. In the case of language acquisition, for example, it is difficult to model input to the child analytically. However, it is relatively easier to model the input with a sample of child-directed speech in adult–child conversations.

After surveying the related psycholinguistic research on segmentation in Chapter 4, the computational problem of segmentation and the solutions offered in the literature, is discussed in Chapter 5. Besides providing a taxonomy of segmentation models and discussing the issues regarding evaluation of the models, Chapter 5 also defines a reference model similar to many state-of-the-art models.

Some aspects of a computational model of segmentation that are discussed in Chapter 5 include, (1) the way the input is modeled, e.g., whether syllables, phonemes, or phonetic features are assumed to be the basic units of the input stream, (2) whether the model guesses boundaries or lexical units, (3) whether the model requires a large amount of input (batch) at once, or if it segments as the learning proceeds (incrementally), and (4) whether it discovers the lexical units by dividing larger chunks, e.g., utterances, or by combining basic units, e.g., phonemes. These aspects, together with the resource requirements of the models are important considerations for their suitability as a psychologically plausible model of segmentation.

The evaluation of any cognitive model is a rather difficult task. Ideally, we expect human-like performance from our models. However, depending on the modeling aims, we do not have enough data on human performance to come up with quantitative evaluation methods in most cases. As a result, all else being equal, we prefer the models that perform better, and we test the performance against a common gold-standard. Fortunately, increasingly many recent models follow common quantitative measures of success (precision, recall and F-score) and common reference corpora. This makes the comparison of different models relatively easy. However, comparing the models' performance is still not trivial. There are a few cases where the performance figures can easily be misinterpreted, and there are cases where using different measures may reveal more information. Chapter 5 points out common pitfalls in comparing different models' performance. For example, it is easy to over-credit a batch model in comparison to an incremental model if we compare the results as they are typically presented in the literature.

Many state-of-the-art computational models of segmentation follow a strategy similar to well-studied 'language models' in the computational linguistics literature. Instead of using one of these models of segmentation as a reference, Chapter 5 defines a simple model following the same strategy. The model (called LM in this thesis) shows a similar performance to the state-of-the-art models in the literature. The LM sets a high standard of performance to achieve. Besides the LM, a particular random segmentation model common in the computational segmentation literature is redefined here. Both models serve as reference models for the other models presented in the thesis.

The LM and the related models perform well on the segmentation task, and they are typically defined as explicit statistical models that are easy to reason with. However, this particular strategy does not follow what we know about human processing closely. A number of connectionist models follow human performance more closely. However, these systems tend to perform poorly, and interpreting what connectionist models learn is generally difficult. It seems, we lack explicitly described segmentation models that follow what we know about the way humans segment continuous speech. This study tries to fill this gap by developing and testing explicit models that use the strategies believed to be used by humans in this task. Crucially, special emphasis is given to how these strategies combine to improve segmentation.

After setting the stage for the intended modeling exercise, following three chapters investigate techniques based on three different strategies. The first strategy discussed in depth is segmentation using predictability statistics. In a nutshell, we know from the psycholinguistic literature that statistical regularities between the consecutive phonemes or syllables is a cue used by infants in the segmentation task. This chapter starts by analyzing a set of measures commonly used for quantifying this notion, namely, transitional probability, mutual information, successor variety and entropy. The analysis reveals similarities and differences between these measures, but it also shows that all of these measures indicate something relevant to word boundaries. Furthermore, despite considerable overlap in what they measure, they each seem to include some relevant aspect of the input that is not covered by the others. This suggests that combining these measures may be more effective than picking the best of them. The rest of the chapter defines a simple algorithm to combine all of these measures, and shows (using simulations) that the combination indeed produces better results than using the individual measures alone. As well as investigating the usefulness of the predictability cue and different ways of quantifying predictability, this chapter also presents an example combination model that can be extended using different cues.

The simulations with the model based on predictability statistics indicate that the model performs consistently better than the random results. Although, the performance results are not as good as the reference model LM, they are not not too far behind either.

Another language-independent strategy is the use of the information gathered from utterance boundaries for finding lexical unit boundaries. The information from utterance boundaries is useful in two ways. First, utterance boundaries are also word boundaries. It is possible to learn common word beginnings and endings from the utterance beginnings and endings. Second, words in natural languages follow certain regularities, which may allow us to extract broader generalizations than we would if we only memorized the exact sequences that occur at utterance boundaries. Chapter 7 investigates the use of information extracted from utterance boundaries for segmentation. As demonstrated in this chapter, the analysis of real-world child-directed speech indicates that utterance boundaries are good predictors of word boundaries. The analysis in Chapter 7 also shows that paying attention to the sequences that occur at utterance boundaries not only allows one to detect words previously seen at utterance boundaries, but also some words that have never occurred at utterance boundaries. This indicates that a learner paying attention to utterance boundaries can learn generalizations about word structure as well as complete words that occur at utterance boundaries. These intuitions are also tested using computational simulations with a model similar to the model developed in Chapter 6. The model based on utterance boundaries alone performs even better than the model based on predictability. More importantly, the combination of both performs better than the individual models, getting closer to the performance of the reference model LM.

Once some lexical units are discovered using the strategies similar to the ones discussed above, we can use (language-specific) information contained in these units. Chapter 8 discusses segmentation strategies based on two such sources of information. First, a model that favors the use of already known words has been tested. The model, when combined with previous two models, provides a consistent but small improvement in the overall performance. The second language-specific cue investigated is lexical stress. The results of the simulations with lexical stress are more difficult to interpret. Adding lexical stress as a cue seems to have an adverse effect on the overall performance. However, if the performance through time is examined more carefully, the effect of stress towards the end of the learning phase is positive. The difficulties regarding interpreting the contribution of stress are mainly due to the stress coding in the corpora used. Unfortunately, the available stress marking was rather crude, and unlikely to be representative of real-world data. Both language-specific cues are studied less in depth relatively in this study which leaves quite some room for improvement. Nevertheless, the simulations in Chapter 8 demonstrate that once we start learning some lexical units, they may be helpful in finding other lexical units.

Chapter 9 gives a summary of the segmentation models presented in the thesis, provides further analysis and comparison of the models presented earlier and discusses possible future directions.

Samenvatting in het Nederlands

De segmentatie van continue spraak in lexicale eenheden is één van de eerste vaardigheden die een kind moet leren gedurende de taalverwerving. Dit proefschrift onderzoekt segmentatie met behulp van computationeel modelleren en computationele simulaties.

Segmentatie is moeilijker dan het op het eerste gezicht kan lijken. Kinderen moeten woorden vinden in een continue stroom van spraak, zonder kennis van woorden te hebben. Gelukkig laten experimentele studies zien dat kinderen en volwassen een aantal aanwijzingen uit de invoer gebruiken, alsmede simpele strategieën die gebruik maken van deze aanwijzingen, om spraak te segmenteren. Nog interessanter is dat een aantal van deze aanwijzingen taal-onafhankelijk zijn, waardoor een taalverwerver continue input kan segmenteren voordat het een enkel woord kent.

De modellen die in dit proefschrift voorgesteld worden, verschillen op twee belangrijke vlakken van modellen uit de literatuur. Ten eerste gebruiken ze lokale strategieën – in tegenstelling tot globale optimalisatie – die gebruik maken van aanwijzingen waarvan bekend is dat kinderen ze gebruiken, namelijk voorspelbaarheidsstatistieken, fonotactiek en lexicale beklemtoning. Ten tweede worden deze aanwijzingen gecombineerd met behulp van een expliciet aanwijzing-combinatie model, dat eenvoudig uitgebreid kan worden met meer aanwijzingen.

Deze modellen zijn getest met behulp van reële getranscribeerde kind-gerichte spraak. De resultaten van de simulaties laten zien dat de prestaties van de individuele strategieën vergelijkbaar zijn met state-of-the-art computationele modellen voor segmentatie. Daarnaast levert het combineren van individuele aanwijzingen een consistente verbetering in prestaties op. Het gecombineerde model presteert even goed als het state-of-the-art model dat als referentie gebruikt wordt, terwijl het alleen gebruik maakt van mechanismen die beter vergelijkbaar zijn met mechanismen die voorhanden zijn voor mensen die dezelfde taak verrichten.

Dit proefschrift vangt aan met een algemene introductie in het probleem van taalverwerving, de en beschrijft de moeilijkheden en geschillen in het vakgebied. Het komt nauwelijks voor dat een werk geen standpunt kiest in het welbekende nature–nurture debat. Ik beargumenteer echter in Hoofdstuk 2 en Hoofdstuk 3 dat er geen wetenschappelijke criteria zijn die het debat ten gunste van één van beide standpunten beslecht, tenminste niet totdat er meer fysiologisch bewijs beschikbaar komt. Ondanks het feit dat er in dit onderzoeksgebied gewoonlijk a priori stelling wordt gekozen voor één van deze standpunten, is dit niet per definitie nuttig.

Zoals in veel andere vakgebieden in de wetenschap kan modelleren, of specifiek computationeel modelleren, gebruikt worden om vragen met betrekking tot taalverwerving te bestuderen. Computationele modellen kunnen bestudeerd worden met behulp van wiskundige analyse of computationele simulaties. Hoofdstuk 3 beschrijft deze methodes, die complementair zijn, nader. Computationele simulaties maken het modelleren echter eenvoudiger ten aanzien van de probabilistische aspecten van het fenomeen dat gemodelleerd wordt.

Na het bestuderen van gerelateerd psycholinguistisch onderzoek naar segmentatie in Hoofdstuk 4, wordt in Hoofdstuk 5 het probleem van computationele segmentatie besproken, alsmede oplossingen voor dit probleem in de literatuur. Hoofdstuk 5 geeft naast een taxonomie van segmentatiemodellen en een beschrijving van de evaluatiemethoden een referentiemodel dat lijkt op veel van de state-of-the-art modellen.

Na het bediscussiëren van de modelleeroefening, onderzoeken de volgende drie hoofdstukken de verschillende strategieën. Hoofdstuk 6 geeft een grondige beschrijving van segmentatie met behulp van voorspelbaarheidsstatistieken. Deze strategie vereist geen initiële taal-specifieke kennis, en het is aangetoond dat mensen dezelfde strategie gebruiken in deze taak. Een andere taal-onafhankelijke strategie gebruikt informatie die verzameld is over uitingsgrenzen voor het vinden van grenzen in lexicale eenheden. Hoofdstuk 7 onderzoekt deze strategie grondig, en toont aan dat het combineren van deze twee strategieën betere prestaties oplevert dan de strategieën afzonderlijk.

Zodra enkele lexicale eenheden ontdekt zijn met behulp van de strategieën die hierboven zijn beschreven, kunnen we (taal-specifieke) informatie uit deze eenheden gebruiken. Hoofdstuk 8 beschrijft segmentatiestrategieën die gebaseerd zijn op twee van dergelijke informatiebronnen. Ten eerste, een model dat voorkeur geeft voor het gebruik van bekende woorden, ten tweede een model dat gebruik maakt van lexicale beklemtoning. Hoofdstuk 8 demonstreert dat, zodra een taalverwerver een aantal lexicale eenheden heeft geleerd, ze hulpvol zijn in het verdere leerproces.

Alvorens tot een conclusie te komen, geeft Hoofdstuk 9 een samenvatting van de segmentatiemodellen die in dit proefschrift besproken zijn, geeft het een verdere analyse en vergelijking van deze modellen, en speculeert het over mogelijk toekomstig onderzoek.

A The Corpora

Two corpora of child directed speech from the CHILDES database (MacWhinney and Snow, 1985) are use in this study.

The first corpus, the BR corpus, is collected by Bernstein Ratner (1987) and phonemically transcribed and processed by Brent and Cartwright (1996) and Brent (1999a). The symbols used while transcribing are listed in Table A.1. Brent (1999a) describes the corpus as follows:

> The speakers were nine mothers speaking freely to their children, whose ages averaged 18 months (range 13–21). In order to minimize the number of subjective judgments and the amount of labor required every word was transcribed the same way every time it occurred. Onomatopoeia (e.g., bang) and interjections (e.g., uh and oh) were removed for the following reasons: (1) They occur in isolation much more frequently than ordinary words, so they would have inflated performance scores; (2) their frequency is highly variable from speaker to speaker and transcriber to transcriber, so their presence would have increased the random variance in performance scores; and (3) there is no standard spelling or pronunciation for many of them, so we could not tell from the orthographic transcript what sound was actually uttered. The total corpus consisted of 9790 utterances, 33,387 words, and 95,809 phonemes. The average of 3.4 words per utterance is typical of spontaneous speech to young children. The average of 2.9 phonemes per word is not surprising for a transcription system like ours, where diphthongs, r-colored vowels (e.g. the "ar" of bar), and syllabic consonants (e.g., the second syllable of bottle) are each transcribed by a single symbol. These sounds are represented by two symbols in some transcription systems and sometimes more than two in English orthography.

This corpus has been used by many other computational studies of segmentation. The corpus is also distributed with the implementation of the models presented by Venkataraman (2001) and Goldwater et al. (2009). The copies of the corpus in these sources are identical, and the same copy has been used in this study without any modifications except for 12 boundary mismatches between segmentation of two words in the text version and phonemic transcriptions. The phonemic transcriptions of 10 instances of the word /`ebisi`/ ‘ABC’ and two instances of the word /`Enim%`/ ‘anymore’ have been modified to match the text version. In all cases, this resulted in removing boundaries in the instances of /`e bi si`/ and /`Eni m%`/. The modifications were motivated by matching the words exactly with the stress patterns in the MRC database, for which the standard spellings were used as a key. Regardless of the use of stress, the modified version of the BR corpus was used for all simulations reported in this thesis. The effect of this modification to the performance scores is insignificant.

For the experiments in Section 8.3, the BR corpus was annotated with stress information using a procedure similar to Christiansen et al. (1998). This process is described in detail in Section 8.3.2.

The second corpus used in this study was gathered from multiple sources in the CHILDES database. For all American English transcripts in the CHILDES database as of April 2011, the recording sessions where target children were less than a year of age were combined. The resulting corpus was a partial combination of the following sections of CHILDES: Brent (Brent and Siskind, 2001), Higginson (Higginson, 1985), Providence (Demuth et al., 2006), Rollins (Rollins et al., 1994, the section of the corpora for normally developing children) and Sodesrstrom (Soderstrom et al., 2008).

All child-directed utterances in these sessions are processed and converted to phonemic transcriptions following Brent (1999a). The resulting corpus contained 53,770 child directed utterances for 24 different children recorded in 171 sessions. The ages of children were between 0;6 and 0;11.29 (mean=9;11, sd=48 days). The ordering of the utterances in each session was kept intact, and the sessions were combined according to the age of the child from younger to older.

Consonants	
Symbol	Example
D	**th**e
G	**j**ump
L	bott**le**
M	rhyth**m**
N	si**ng**
S	**sh**ip
T	**th**in
W	**wh**en
Z	a**z**ure
b	**b**oy
c	**ch**ip
d	**d**og
f	**f**ox
g	**g**o
h	**h**at
k	**c**ut
l	**l**amp
m	**m**an
n	**n**et
p	**p**ipe
r	**r**un
s	**s**it
t	**t**oy
v	**v**iew
w	**w**e
y	**y**ou
z	**z**ip
~	butt**on**

(a) Consonants

Vowels	
Symbol	Example
&	th**a**t
6	**a**bout
7	bO**y**
9	fl**y**
A	b**u**t
E	b**e**t
I	b**i**t
O	l**aw**
Q	b**ou**t
U	p**u**t
a	h**o**t
e	b**ay**
i	b**ee**
o	b**oa**t
u	b**oo**t

(b) Vowels

Rhotic Vowels	
Symbol	Example
#	**ar**e
%	f**or**
(	h**ere**
)	l**ure**
*	h**air**
3	b**ir**d
R	butt**er**

(c) Vowels with ‘r’ (Rhotic Vowels)

Table A.1: The symbols used for phonemes in the BR corpus.

B Detailed Performance Results

This appendix lists the detailed performance results, including true positive, false positive and false negative values, and all relevant combinations of the models described in this thesis.

Table B.1 presents performance scores for all models described in this thesis and all possible combinations of these models on the BR corpus. First part of the table presents the scores calculated for the complete learning process, and the second part presents the results for the last 290 utterances.

Table B.2 presents the same scores where the larger child directed corpus described in Appendix A was used to collect prior statistics on phoneme n-grams. This table does not present the result including stress due to the fact that the larger corpora has not been marked for stress. The details of using prior statistics are described in Section 9.3.

The last set of results in Table B.3 presents the scores for the larger collection of child directed speech alone. As in Table B.2, the models requiring stress information could not be included in this table.

	model	BTP	BFP	BFN	WTP	WFP	WFN	LTP	LFP	LFN	BP	BR	BF	WP	WR	WF	LP	LR	LF	E_o	E_u
Complete BR corpus	P	21826	9524	1761	23422	17718	9955	660	1137	664	69.6	92.5	79.5	56.9	70.2	62.9	36.7	49.8	42.3	15.3	7.5
	U	20005	4140	3582	23936	9999	9441	886	1739	438	82.9	84.8	83.8	70.5	71.7	71.1	33.8	66.9	44.9	6.6	15.2
	W	16819	4895	6768	19097	12407	14280	631	2824	693	77.5	71.3	74.3	60.6	57.2	58.9	18.3	47.7	26.4	7.8	28.7
	S	1927	536	21660	3246	9007	30131	512	5731	812	78.2	8.2	14.8	26.5	9.7	14.2	8.2	38.7	13.5	0.9	91.8
	PU	21404	4501	2183	25848	9847	7529	864	1153	460	82.6	90.7	86.5	72.4	77.4	74.8	42.8	65.3	51.7	7.2	9.3
	PW	21864	10090	1723	22931	18813	10446	666	1027	658	68.4	92.7	78.7	54.9	68.7	61.1	39.3	50.3	44.1	16.2	7.3
	PS	1974	528	21613	3282	9010	30095	510	5722	814	78.9	8.4	15.1	26.7	9.8	14.4	8.2	38.5	13.5	0.8	91.6
	UW	20764	4943	2823	24606	10891	8771	857	1486	467	80.8	88.0	84.2	69.3	73.7	71.5	36.6	64.7	46.7	7.9	12.0
	US	2105	455	21482	3521	8829	29856	518	5649	806	82.2	8.9	16.1	28.5	10.5	15.4	8.4	39.1	13.8	0.7	91.1
	WS	1877	548	21710	3194	9021	30183	504	5711	820	77.4	8.0	14.4	26.1	9.6	14.0	8.1	38.1	13.4	0.9	92.0
	PUW	21669	4878	1918	26067	10270	7310	870	1062	454	81.6	91.9	86.4	71.7	78.1	74.8	45.0	65.7	53.4	7.8	8.1
	PUS	14005	720	9582	17292	7223	16085	758	3320	566	95.1	59.4	73.1	70.5	51.8	59.7	18.6	57.3	28.1	1.2	40.6
	PWS	2111	623	21476	3462	9062	29915	508	5659	816	77.2	8.9	16.0	27.6	10.4	15.1	8.2	38.4	13.6	1.0	91.1
	UWS	6175	495	17412	8349	8111	25028	552	4865	772	92.6	26.2	40.8	50.7	25.0	33.5	10.2	41.7	16.4	0.8	73.8
	PUWS	17863	1389	5724	22736	6306	10641	830	2268	494	92.8	75.7	83.4	78.3	68.1	72.8	26.8	62.7	37.5	2.2	24.3
	RM	11720	31134	11867	4537	48107	28840	515	6351	809	27.3	49.7	35.3	8.6	13.6	10.5	7.5	38.9	12.6	49.9	50.3
	LM	19512	3676	4075	23755	9223	9622	807	787	517	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3
Last 290 utterances	P	680	238	24	769	439	225	159	124	86	74.1	96.6	83.8	63.7	77.4	69.8	56.2	64.9	60.2	13.7	3.4
	U	657	127	47	803	271	191	189	90	56	83.8	93.3	88.3	74.8	80.8	77.7	67.7	77.1	72.1	7.3	6.7
	W	600	184	104	685	389	309	158	141	87	76.5	85.2	80.6	63.8	68.9	66.2	52.8	64.5	58.1	10.6	14.8
	S	46	15	658	95	256	899	46	222	199	75.4	6.5	12.0	27.1	9.6	14.1	17.2	18.8	17.9	0.9	93.5
	PU	681	115	23	855	231	139	186	73	59	85.6	96.7	90.8	78.7	86.0	82.2	71.8	75.9	73.8	6.6	3.3
	PW	687	270	17	758	489	236	152	115	93	71.8	97.6	82.7	60.8	76.3	67.6	56.9	62.0	59.4	15.6	2.4
	PS	48	13	656	98	253	896	47	220	198	78.7	6.8	12.5	27.9	9.9	14.6	17.6	19.2	18.4	0.8	93.2
	UW	668	142	36	802	298	192	180	95	65	82.5	94.9	88.2	72.9	80.7	76.6	65.5	73.5	69.2	8.2	5.1
	US	50	13	654	103	250	891	51	217	194	79.4	7.1	13.0	29.2	10.4	15.3	19.0	20.8	19.9	0.8	92.9
	WS	45	17	659	92	260	902	45	224	200	72.6	6.4	11.7	26.1	9.3	13.7	16.7	18.4	17.5	1.0	93.6
	PUW	684	137	20	840	271	154	185	80	60	83.3	97.2	89.7	75.6	84.5	79.8	69.8	75.5	72.5	7.9	2.8
	PUS	558	19	146	721	146	273	173	105	72	96.7	79.3	87.1	83.2	72.5	77.5	62.2	70.6	66.2	1.1	20.7
	PWS	50	19	654	95	264	899	50	223	195	72.5	7.1	12.9	26.5	9.6	14.0	18.3	20.4	19.3	1.1	92.9
	UWS	201	28	503	287	232	707	90	191	155	87.8	28.6	43.1	55.3	28.9	37.9	32.0	36.7	34.2	1.6	71.4
	PUWS	629	51	75	817	153	177	190	78	55	92.5	89.3	90.9	84.2	82.2	83.2	70.9	77.6	74.1	2.9	10.7
	LM	626	77	78	801	192	193	181	63	64	89.0	88.9	89.0	80.7	80.6	80.6	74.2	73.9	74.0	4.4	11.1

Table B.1: Performance scores for all possible combinations of the models described in Chapters 6, 7 and 8 and two reference models described in Chapter 5. The first block reports the performance or error measures for the complete BR corpus, and the second block of reports the results for the last 290 utterances. The measures are described in Chapter 5.

	model	BTP	BFP	BFN	WTP	WFP	WFN	LTP	LFP	LFN	BP	BR	BF	WP	WR	WF	LP	LR	LF	E_o	E_u
Complete BR corpus	P	22652	11211	935	23442	20211	9935	623	837	701	66.9	96.0	78.9	53.7	70.2	60.9	42.7	47.1	44.8	18.0	4.0
	U	22173	6049	1414	26031	11981	7346	852	718	472	78.6	94.0	85.6	68.5	78.0	72.9	54.3	64.4	58.9	9.7	6.0
	W	16819	4895	6768	19097	12407	14280	631	2824	693	77.5	71.3	74.3	60.6	57.2	58.9	18.3	47.7	26.4	7.8	28.7
	PU	22576	5766	1011	26793	11339	6584	829	619	495	79.7	95.7	86.9	70.3	80.3	74.9	57.3	62.6	59.8	9.2	4.3
	PW	22612	10901	975	23523	19780	9854	615	745	709	67.5	95.9	79.2	54.3	70.5	61.4	45.2	46.5	45.8	17.5	4.1
	UW	22352	6684	1235	25853	12973	7524	842	756	482	77.0	94.8	85.0	66.6	77.5	71.6	52.7	63.6	57.6	10.7	5.2
	PUW	22723	5864	864	27079	11298	6298	833	568	491	79.5	96.3	87.1	70.6	81.1	75.5	59.5	62.9	61.1	9.4	3.7
	RM	11720	31134	11867	4537	48107	28840	515	6351	809	27.3	49.7	35.3	8.6	13.6	10.5	7.5	38.9	12.6	49.9	50.3
	LM	19512	3676	4075	23755	9223	9622	807	787	517	84.1	82.7	83.4	72.0	71.2	71.6	50.6	61.0	55.3	5.9	17.3
Last 290 utterances	P	683	258	21	750	481	244	151	117	94	72.6	97.0	83.0	60.9	75.5	67.4	56.3	61.6	58.9	14.9	3.0
	U	678	145	26	825	288	169	190	76	55	82.4	96.3	88.8	74.1	83.0	78.3	71.4	77.6	74.4	8.4	3.7
	W	600	184	104	685	389	309	158	141	87	76.5	85.2	80.6	63.8	68.9	66.2	52.8	64.5	58.1	10.6	14.8
	PU	686	138	18	844	270	150	187	70	58	83.3	97.4	89.8	75.8	84.9	80.1	72.8	76.3	74.5	8.0	2.6
	PW	686	298	18	737	537	257	141	126	104	69.7	97.4	81.3	57.8	74.1	65.0	52.8	57.6	55.1	17.2	2.6
	UW	678	172	26	802	338	192	184	86	61	79.8	96.3	87.3	70.4	80.7	75.2	68.1	75.1	71.5	9.9	3.7
	PUW	690	146	14	851	275	143	186	73	59	82.5	98.0	89.6	75.6	85.6	80.3	71.8	75.9	73.8	8.4	2.0
	LM	626	77	78	801	192	193	181	63	64	89.0	88.9	89.0	80.7	80.6	80.6	74.2	73.9	74.0	4.4	11.1

Table B.2: The performance results with prior data. The detailed performance measures for all possible combinations of the models described in Chapters 6, 7 and 8 and two reference models described in Chapter 5. The first block reports the performance or error measures for the complete BR corpus, and the second block of reports the results for the last 290 utterances. The measures are described in Chapter 5.

	model	BTP	BFP	BFN	WTP	WFP	WFN	LTP	LFP	LFN	BP	BR	BF	WP	WR	WF	LP	LR	LF	E_o	E_u
The complete corpus	P	119280	43852	2237	136241	80661	39046	865	1577	675	73.1	98.2	83.8	62.8	77.7	69.5	35.4	56.2	43.4	13.8	1.8
	U	114303	23098	7214	142657	48514	32630	1172	2534	368	83.2	94.1	88.3	74.6	81.4	77.9	31.6	76.1	44.7	7.3	5.9
	W	105525	99348	15992	87428	171215	87859	503	2698	1037	51.5	86.8	64.7	33.8	49.9	40.3	15.7	32.7	21.2	31.3	13.2
	PU	118055	20454	3462	151184	41095	24103	1097	1383	443	85.2	97.2	90.8	78.6	86.2	82.3	44.2	71.2	54.6	6.4	2.8
	PW	119672	49608	1845	132194	90856	43093	844	1406	696	70.7	98.5	82.3	59.3	75.4	66.4	37.5	54.8	44.5	15.6	1.5
	UW	116259	27369	5258	142003	55395	33284	1101	2088	439	80.9	95.7	87.7	71.9	81.0	76.2	34.5	71.5	46.6	8.6	4.3
	PUW	119106	22391	2411	151739	43528	23548	1085	1249	455	84.2	98.0	90.6	77.7	86.6	81.9	46.5	70.5	56.0	7.1	2.0
	RM	11720	31134	11867	4537	48107	28840	515	6351	809	27.3	49.7	35.3	8.6	13.6	10.5	7.5	38.9	12.6	49.9	50.3
	LM	113141	14188	8376	148114	32985	27173	1017	1008	523	88.9	93.1	90.9	81.8	84.5	83.1	50.2	66.0	57.1	4.5	6.9
Last 270 utterances	P	837	393	29	802	698	334	145	132	97	68.0	96.7	79.9	53.5	70.6	60.8	52.3	59.9	55.9	18.5	3.3
	U	852	248	14	900	470	236	174	79	68	77.5	98.4	86.7	65.7	79.2	71.8	68.8	71.9	70.3	11.7	1.6
	W	768	682	98	551	1169	585	84	186	158	53.0	88.7	66.3	32.0	48.5	38.6	31.1	34.7	32.8	32.1	11.3
	PU	855	208	11	945	388	191	183	69	59	80.4	98.7	88.6	70.9	83.2	76.5	72.6	75.6	74.1	9.8	1.3
	PW	854	419	12	817	726	319	141	118	101	67.1	98.6	79.9	52.9	71.9	61.0	54.4	58.3	56.3	19.7	1.4
	UW	857	235	9	900	462	236	175	77	67	78.5	99.0	87.5	66.1	79.2	72.1	69.4	72.3	70.9	11.1	1.0
	PUW	861	215	5	957	389	179	180	71	62	80.0	99.4	88.7	71.1	84.2	77.1	71.7	74.4	73.0	10.1	0.6
	LM	814	102	52	941	245	195	184	50	58	88.9	94.0	91.4	79.3	82.8	81.1	78.6	76.0	77.3	4.8	6.0

Table B.3: The performance for the larger child-directed speech corpus (see Appendix A for a description of the corpus). The detailed performance measures for all possible combinations of the models described in Chapters 6, 7 and 8 and two reference models described in Chapter 5. The first block reports the performance or error measures for the complete corpus, and the second block of reports the results for the last 270 utterances. The measures are described in Chapter 5.

Groningen Dissertations in Linguistics (GRODIL)

1. Henriëtte de Swart (1991). *Adverbs of Quantification: A Generalized Quantifier Approach.*
2. Eric Hoekstra (1991). *Licensing Conditions on Phrase Structure.*
3. Dicky Gilbers (1992). *Phonological Networks. A Theory of Segment Representation.*
4. Helen de Hoop (1992). *Case Configuration and Noun Phrase Interpretation.*
5. Gosse Bouma (1993). *Nonmonotonicity and Categorial Unification Grammar.*
6. Peter Blok (1993). *The Interpretation of Focus: an epistemic approach to pragmatics.*
7. Roelien Bastiaanse (1993). *Studies in Aphasia.*
8. Bert Bos (1993). *Rapid User Interface Development with the Script Language Gist.*
9. Wim Kosmeijer (1993). *Barriers and Licensing.*
10. Jan-Wouter Zwart (1993). *Dutch Syntax: A Minimalist Approach.*
11. Mark Kas (1993). *Essays on Boolean Functions and Negative Polarity.*
12. Ton van der Wouden (1994). *Negative Contexts.*
13. Joop Houtman (1994). *Coordination and Constituency: A Study in Categorial Grammar.*
14. Petra Hendriks (1995). *Comparatives and Categorial Grammar.*
15. Maarten de Wind (1995). *Inversion in French.*
16. Jelly Julia de Jong (1996). *The Case of Bound Pronouns in Peripheral Romance.*
17. Sjoukje van der Wal (1996). *Negative Polarity Items and Negation: Tandem Acquisition.*
18. Anastasia Giannakidou (1997). *The Landscape of Polarity Items.*
19. Karen Lattewitz (1997). *Adjacency in Dutch and German.*
20. Edith Kaan (1997). *Processing Subject-Object Ambiguities in Dutch.*
21. Henny Klein (1997). *Adverbs of Degree in Dutch.*
22. Leonie Bosveld-de Smet (1998). *On Mass and Plural Quantification: The Case of French 'des'/'du'-NPs.*
23. Rita Landeweerd (1998). *Discourse Semantics of Perspective and Temporal Structure.*
24. Mettina Veenstra (1998). *Formalizing the Minimalist Program.*
25. Roel Jonkers (1998). *Comprehension and Production of Verbs in Aphasic Speakers.*
26. Erik F. Tjong Kim Sang (1998). *Machine Learning of Phonotactics.*
27. Paulien Rijkhoek (1998). *On Degree Phrases and Result Clauses.*
28. Jan de Jong (1999). *Specific Language Impairment in Dutch: Inflectional Morphology and Argument Structure.*
29. H. Wee (1999). *Definite Focus.*
30. Eun-Hee Lee (2000). *Dynamic and Stative Information in Temporal Reasoning: Korean Tense and Aspect in Discourse.*
31. Ivilin Stoianov (2001). *Connectionist Lexical Processing.*
32. Klarien van der Linde (2001). *Sonority Substitutions.*
33. Monique Lamers (2001). *Sentence Processing: Using Syntactic, Semantic, and Thematic Information.*
34. Shalom Zuckerman (2001). *The Acquisition of "Optional" Movement.*

35. Rob Koeling (2001). *Dialogue-Based Disambiguation: Using Dialogue Status to Improve Speech Understanding.*
36. Esther Ruigendijk (2002). *Case Assignment in Agrammatism: a Cross-linguistic Study.*
37. Tony Mullen (2002). *An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection.*
38. Nanette Bienfait (2002). *Grammatica-onderwijs aan allochtone jongeren.*
39. Dirk-Bart den Ouden (2002). *Phonology in Aphasia: Syllables and Segments in Level-specific Deficits.*
40. Rienk Withaar (2002). *The Role of the Phonological Loop in Sentence Comprehension.*
41. Kim Sauter (2002). *Transfer and Access to Universal Grammar in Adult Second Language Acquisition.*
42. Laura Sabourin (2003). *Grammatical Gender and Second Language Processing: An ERP Study.*
43. Hein van Schie (2003). *Visual Semantics.*
44. Lilia Schürcks-Grozeva (2003). *Binding and Bulgarian.*
45. Stasinos Konstantopoulos (2003). *Using ILP to Learn Local Linguistic Structures.*
46. Wilbert Heeringa (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.*
47. Wouter Jansen (2004). *Laryngeal Contrast and Phonetic Voicing: A Laboratory Phonology Approach to English, Hungarian and Dutch.*
48. Judith Rispens (2004). *Syntactic and Phonological Processing in Developmental Dyslexia.*
49. Danielle Bougaïré (2004). *L'approche communicative des campagnes de sensibilisation en santé publique au Burkina Faso: les cas de la planification familiale, du sida et de l'excision.*
50. Tanja Gaustad (2004). *Linguistic Knowledge and Word Sense Disambiguation.*
51. Susanne Schoof (2004). *An HPSG Account of Nonfinite Verbal Complements in Latin.*
52. M. Begoña Villada Moirón (2005). *Data-driven identification of fixed expressions and their modifiability.*
53. Robbert Prins (2005). *Finite-State Pre-Processing for Natural Language Analysis.*
54. Leonoor van der Beek (2005). *Topics in Corpus-Based Dutch Syntax.*
55. Keiko Yoshioka (2005). *Linguistic and gestural introduction and tracking of referents in L1 and L2 discourse.*
56. Sible Andringa (2005). *Form-focused instruction and the development of second language proficiency.*
57. Joanneke Prenger (2005). *Taal telt! Een onderzoek naar de rol van taalvaardigheid en tekstbegrip in het realistisch wiskundeonderwijs.*
58. Neslihan Kansu-Yetkiner (2006). *Blood, Shame and Fear: Self-Presentation Strategies of Turkish Women's Talk about their Health and Sexuality.*
59. Mónika Z. Zempléni (2006). *Functional imaging of the hemispheric contribution to language processing.*
60. Maartje Schreuder (2006). *Prosodic Processes in Language and Music.*
61. Hidetoshi Shiraishi (2006). *Topics in Nivkh Phonology.*
62. Tamás Biró (2006). *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing.*
63. Dieuwke de Goede (2006). *Verbs in Spoken Sentence Processing: Unraveling the Activation Pattern of the Matrix Verb.*
64. Eleonora Rossi (2007). *Clitic production in Italian agrammatism.*
65. Holger Hopp (2007). *Ultimate Attainment at the Interfaces in Second Language Acquisition: Grammar and Processing.*
66. Gerlof Bouma (2008). *Starting a Sentence in Dutch: A corpus study of subject- and object-fronting.*
67. Julia Klitsch (2008). *Open your eyes and listen carefully. Auditory and audiovisual speech perception and the McGurk effect in Dutch speakers with and without aphasia.*
68. Janneke ter Beek (2008). *Restructuring and Infinitival Complements in Dutch.*
69. Jori Mur (2008). *Off-line Answer Extraction for Question Answering.*

70. Lonneke van der Plas (2008). *Automatic Lexico-Semantic Acquisition for Question Answering.*
71. Arjen Versloot (2008). *Mechanisms of Language Change: Vowel reduction in 15th century West Frisian.*
72. Ismail Fahmi (2009). *Automatic term and Relation Extraction for Medical Question Answering System.*
73. Tuba Yarbay Duman (2009). *Turkish Agrammatic Aphasia: Word Order, Time Reference and Case.*
74. Maria Trofimova (2009). *Case Assignment by Prepositions in Russian Aphasia.*
75. Rasmus Steinkrauss (2009). *Frequency and Function in WH Question Acquisition. A Usage-Based Case Study of German L1 Acquisition.*
76. Marjolein Deunk (2009). *Discourse Practices in Preschool. Young Children's Participation in Everyday Classroom Activities.*
77. Sake Jager (2009). *Towards ICT-Integrated Language Learning: Developing an Implementation Framework in terms of Pedagogy, Technology and Environment.*
78. Francisco Dellatorre Borges (2010). *Parse Selection with Support Vector Machines.*
79. Geoffrey Andogah (2010). *Geographically Constrained Information Retrieval.*
80. Jacqueline van Kruiningen (2010). *Onderwijsontwerp als conversatie. Probleemoplossing in interprofessioneel overleg.*
81. Robert G. Shackleton (2010). *Quantitative Assessment of English-American Speech Relationships.*
82. Tim Van de Cruys (2010). *Mining for Meaning: The Extraction of Lexico-semantic Knowledge from Text.*
83. Therese Leinonen (2010). *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects.*
84. Erik-Jan Smits (2010). *Acquiring Quantification. How Children Use Semantics and Pragmatics to Constrain Meaning.*
85. Tal Caspi (2010). *A Dynamic Perspective on Second Language Development.*
86. Teodora Mehotcheva (2010). *After the fiesta is over. Foreign language attrition of Spanish in Dutch and German Erasmus Students.*
87. Xiaoyan Xu (2010). *English language attrition and retention in Chinese and Dutch university students.*
88. Jelena Prokić (2010). *Families and Resemblances.*
89. Radek Šimík (2011). *Modal existential wh-constructions.*
90. Katrien Colman (2011). *Behavioral and neuroimaging studies on language processing in Dutch speakers with Parkinson's disease.*
91. Siti Mina Tamah (2011). *A Study on Student Interaction in the Implementation of the Jigsaw Technique in Language Teaching.*
92. Aletta Kwant (2011). *Geraakt door prentenboeken. Effecten van het gebruik van prentenboeken op de sociaal-emotionele ontwikkeling van kleuters.*
93. Marlies Kluck (2011). *Sentence amalgamation.*
94. Anja Schüppert (2011). *Origin of asymmetry: Mutual intelligibility of spoken Danish and Swedish.*
95. Peter Nabende (2011). *Applying Dynamic Bayesian Networks in Transliteration Detection and Generation.*
96. Barbara Plank (2011). *Domain Adaptation for Parsing.*
97. Çağrı Çöltekin (2011). *Catching Words in a Stream of Speech: Computational simulations of segmenting transcribed child-directed speech.*
98. Dörte Hessler (2011). *Audiovisual Processing in Aphasic and Non-Brain-Damaged Listeners: The Whole is More than the Sum of its Parts.*

GRODIL
Secretary of the Department of General Linguistics
Postbus 716
9700 AS Groningen
The Netherlands