NORTHWESTERN UNIVERSITY

Word Segmentation, Word Recognition, and Word Learning: A Computational Model of First Language Acquisition

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Linguistics

By

Robert Daland

June 2009

ABSTRACT

Word Segmentation, Word Recognition, and Word Learning: A Computational Model of First Language Acquisition

Robert Daland

Many word boundaries are not marked acoustically in fluent speech (Lehiste, 1960), a fact that is immediately apparent from listening to speech in an unfamiliar language, and which poses a special problem for infants. The acquisition literature shows that infants begin to segment speech (identify word boundaries) between 6 and 10.5 months (Saffran, Aslin, & Newport, 1996; Jusczyk, Hohne, & Baumann, 1999; Jusczyk, Houston, & Newsome, 1999; Mattys & Jusczyk, 2001; Bortfeld, Morgan, Golinkoff, & Rathbun, 2005) although they possess minuscule receptive vocabularies at this age (Dale & Fenson, 1996). Thus, word segmentation largely appears before and supports word learning (Aslin, Woodward, LaMendola, & Bever, 1996; van de Weijer, 1998; Brent & Siskind, 2001; Davis, 2004), rather than the other way around. These results raise several further questions. How do infants begin to find word boundaries in speech when they don't know most of the words they hear? How are word segmentation, word recognition, and word learning linked in development? I propose DiBS – x *Di*phone-*B*ased *S*egmentation – as a computational model of word segmentation. The core idea of DiBS is to recover word boundaries in speech based on the immediate phonotactic context, by estimating the probabilities of a word boundary within every possible sequence of two speech sounds (diphone, e.g. [ba]). As a proof of concept, a supervised DiBS model is tested on English and Russian data, yielding a

consistent pattern of high accuracy with some undersegmentation. Next, a learning theory is developed, by which DiBS can be estimated from information that is observable to infants, including the distribution of speech sounds at phrase edges and any words they have managed to learn; these models achieve superior segmentation relative to other prelexical statistical proposals such as segmentation based on Saffran et al's (1996) transitional probability. Finally, this learning model is integrated with a model of lexical access and word-learning to form a full bootstrapping model, which achieves a relatively high degree of success in word segmentation, but only partial success in word learning. The successes and failures of this model are discussed, as they highlight the need for additional research on wordform learning.

ACKNOWLEDGEMENTS

This dissertation is dedicated to my grandfather and namesake, who would have been so proud.

Above all I wish to thank Janet Pierrehumbert. This dissertation began from her – it sprang from a project in her Methods class with Matt Goldrick. This dissertation grew with her – in every stage Janet has provided invaluable intellectual and emotional support. I have learned so many thing from Janet that I can't really put into words. My biggest hope is this dissertation will make her proud.

In addition I wish to thank Matt Goldrick and Jessica Maye who have each in their own way made me the researcher that I am. Jessica introduced me to the unbelievably fascinating questions of language acquisition, supported my early years of research in her laboratory, and has always told me the little things that you normally have to learn the hard way. Just as importantly, Jessica looks out for the human side in her students. Matt has been a tireless advocate for graduate students and for me in particular; he taught me how to be a sober and cautious researcher, and has given patiently of his time to help me learn to write. Just as importantly for my future, he taught me how to attract grad students: “I've got this great idea, and I think it will only take 8 hours of your time...”

Andrea Sims – I hope for and look forward to years of collaborations.

I would also like to acknowledge a huge intellectual debt to researchers who have made outstanding contributions to our knowledge of infant language development, in particular Richard

Aslin, Peter Jusczyk, and Dan Swingley.

Other teachers who have inspired me include Dot Doyle, Jackie Dusenbury, Dan Teague, Sandy Sanderson, Tony Stewart, Natalie Dohrmann, Nick Halpern, Jeff Lidz, Ann Bradlow and Lis Elliott. Especial thanks to Stefan Kaufmann for supporting my interest in computational linguistics.

I wish to thank the following people for their friendship and emotional support: Katia Bower, Matt Berends, Reinout de Bock, Fabio Braggion, Helge Braun, Adam Daland, Edwina Daland, Virginia Daland, Will Daland, Scottie Gottbreht, Dottie Hall, Jeannie Jacob, Ronnie Kinner, Alanna Krause, Meredith Larson, Quinton Maynard, Nattalia Paterson, Rade Radjenovic, Lipika Samal, Maho Taguchi, Matt Weinberg, Jennie Zhao, and last alphabetically but not in my heart, Xiaoju Zheng. A great big thank you to the members of my class, all other linguistics graduate students, and other members of the Northwestern language research community.

Finally, I wish to think the Cognitive Science Program, Janet Pierrehumbert, and The Graduate School for support during my 2nd, 5th, and 6th years respectively. This support enabled me to make this dissertation.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

Simply to speak and understand a language is a cognitive and social achievement of astounding complexity. Even more astounding is the process of *learning* to speak and understand a novel language. And what is most astounding of all is the amazing rapidity with which every typically-developing child learns the language(s) they are exposed to.

By the time they are 3 or 4, children command all fundamental communicative functions of language – requesting, answering, ordering, stating, and so forth. This fact is illustrated by the following child utterances from the CHILDES child language database (MacWhinney, 2000):

(1)	requesting:	'Is Daddy with you?'	(Ross, 2;7, MacWhinney corpus 21a1.cha)
	answering:	'yeah I pway baskpots'	(Ross, 2;7, MacWhinney corpus 21a1.cha)
	stating:	'I go down in the water'	(Ross, 2;7, MacWhinney corpus 21a1.cha)
	ordering:	'Now watch'	(Ross, 2;7, MacWhinney corpus 21a1.cha)

This knowledge extends to fine-grained aspects of language, such as subtle aspects of syntax. To give but one example, 4-year-old English learners interpret linguistically sophisticated sentences like (1) in the same way as adults:

(2) Miss Piggy wanted to drive every car that Kermit did.

Interpretation 1: *did* = *drove*

Interpretation 2: *did* = *wanted to drive* (Syrett & Lidz, 2005)

That is, like adults, they access either interpretation in a context in which it is true, and reject both interpretations in contexts in which they are false. Research on (first) language acquisition attempts to explain when, how, and why children acquire various aspects of their native language.

One of the first steps in language acquisition is word segmentation. Word segmentation is the perceptual process whereby fluent listeners hear speech as a sequence of word-sized units[1]. Word segmentation is evidently a perceptual process, because unlike written English – which contains spaces betweens words – speech usually does not signal word boundaries with pauses or other unambiguous acoustic cues (Lehiste, 1960). Word segmentation is necessary for listeners to perceive unrecognized words in their input; thus, word segmentation paves the way for infants to proceed from learning "low-level" properties of their language (such as its phonetic categories) to "higher-level" properties (such as its syntax).

This dissertation takes up the question of how children acquire word segmentation. That is, it aims to discover what linguistic knowledge infants command that helps them to segment speech, and how they come to possess that knowledge. The methodology I will adopt is computational modeling – computer programs that implement specific theoretical proposals about the acquisition of word segmentation, and analyses of their performance. Since the goal of this dissertation is to explain the acquisition of word segmentation cross-linguistically, it is important to test these proposals not just on English data, but on other language data. This

1 A precise, coherent, and cross-linguistically satisfactory definition of 'word' is a topic of considerable linguistic research and controversy (Bauer, 1983). Thus I will defer detailed discussion of this topic to a later section. I will note here that for the purposes of this dissertation, 'word' will be operationally defined as a contiguous orthographic sequence, delimited on both sides by orthographic word boundaries such as spaces or sentential punctuation.

dissertation takes a nontrivial step in this direction by testing most of its models in parallel on both English and Russian, which required deriving a large and comparable Russian phonetic corpus (also nontrivial!).

Terminology

This dissertation draws from research in a number of distinct fields, with differing and occasionally idiosyncratic terminology. Therefore it is worth a few moments at the beginning to clarify the terminology used in this dissertation. A specific technical sense will be used for the following:

infant. A child between the ages of 6 and 10.5 months of age. (As reviewed in detail below, the developmental literature suggests that infants have little or no word segmentation prior to 6 months, and quite sophisticated segmentation abilities by 10.5 months of age.) When reviewing a study, infant will further indicate an English-learning infant unless specifically noted otherwise, as the majority of segmentation studies have been conducted in English.

phoneme. A phoneme is a cognitive unit corresponding to a collection of simultaneous phonological features that is realized in speech as a consonant or vowel. Phonemes are abstract cognitive units, whose realization as a speech sound depends on various factors, such as the prosodic position in which they occur. For example, the words *pat* and *tap* both contain the phonemes /p/, /æ/, and /t/ in different orders (Hyman, 1975)

phone. A phone is a low-level perceptual category, and is therefore conditioned on its prosodic position (see review in Pierrehumbert, 2002). For example, in American English the

phoneme /t/ is typically realized as an aspirated stop in foot-initial position (e.g. *Tom*), as an unaspirated stop in a syllable-initial [st] cluster, and as a flap intervocalically in foot-medial position (*butter*); foot-finally it alternates between an unreleased alveolar stop and a glottal stop. Each of these is a distinct phone category.

segment. The terms phoneme and phone distinguish levels of abstractness, so in practice there is a strong correspondence between the two. I will use segment whenever I wish to refer to a speech unit but do not wish to take a stance regarding level of abstractness on the part of the listener.

*phonotactics***.** Phonotactics refers to probabilistic and categorical constraints on phonological structures and sequences, including syllable structure and stress (Chomsky & Halle, 1965; Jusczyk, Luce, & Charles-Luce, 1994; Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999; Hayes & Wilson, 2008). In this dissertation I will focus primarily on segmental sequences and their distribution within and across words. For example, the sequence [pd] rarely occurs word-internally.

prosodic word. A prosodic word is a prosodic constituent that generally consists of a single content word together with associated function words/morphemes, e.g. determiners and/or inflectional morphology (Selkirk, 1984; Nespor & Vogel, 1986). For example, *gotta* and *catamaran* are both prosodic words with nuclear accents on the first syllable.

morphosyntactic word. A morphosyntactic word is a morphosyntactic constituent which indicates a lexical or syntactic meaning (Bauer, 2003). For example, *gotta* could be analyzed as containing two morphosyntactic words, the semi-auxiliary *get* and the infinitival marker *to*. In

general, a prosodic word may contain one or more morphosyntactic words, but a morphosyntactic word is not realized across multiple prosodic words (Selkirk, 1984; McCarthy & Prince, 1986/1996).

phrase. Intonational phrases tend to contain between four and seven syllables, are typically marked by phrase-final lengthening and phrase-initial strengthening, and usually consist of one intonation contour with a potential pitch discontinuity at the boundary (Christophe, Gout, Peperkamp, & Morgan, 2003; Beckman & Pierrehumbert, 1986; Shattuck-Hufnagel & Turk, 1996). I assume phrases contain a contiguous sequence of prosodic words and are demarcated on both edges by language-universal acoustic boundary cues, such as a syllable-length or greater pause, phrase-final lengthening[2], and phrase-initial fortition[3].

In addition to these technical terms, I will use the following acronyms:

*DiBS (**Di**phone-**B**ased **S**egmentation).* The theory of prelexical word segmentation developed and tested extensively throughout this dissertation/

*MLDT (**M**aximum **L**ikelihood **D**ecision **T**hreshold).* In a signal detection (Green & Swets, 1966) scenario, the value of an observable statistic (e.g. the probability of a word boundary given the observed diphone) may be compatible with both categories (e.g. boundary, no boundary), but in general, one will be more likely than the other. The MLDT is the point at which the crossover

2 The claim that phrase-final lengthening is universal is based on the fact that it has been found in all languages tested, including Brazilian Portugese (Barbosa, 2002), Dutch (de Pijper & Sanderman, 1994), Japanese (Fisher & Tokura, 1996), French (Rietveld, 1980), and English (Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992).

3 Similarly, phrase-initial fortition has been found in all languages tested, including Korean (Cho & Keating, 2001), French (Fougeron, 2001), Taiwanese, and English (Keating, Cho, Fougeron, & Hsu, 2003; Pierrehumbert & Talkin, 1991).

occurs; i.e. if the statistic is greater than MLDT, the best decision is to report the signal as present because that is more likely than the alternative.

*ROC (**R**eceiver **O**perating **C**haracteristic).* A plot used in signal detection theory (Green & Swets, 1966) to illustrate the sensitivity of a detector as a function of detection threshold.

*BNC (**B**ritish **N**ational **C**orpus).* A large corpus of British English.

*RNC (**R**ussian **N**ational **C**orpus).* A large corpus of Russian.

Infant Word Segmentation as a Distinct Research Problem

Infants, unlike adults, do not know very many words. But, like adults, they must solve a number of related but distinct sub-problems in speech processing, in particular *word recognition*, *word segmentation*, and *word learning*. Infants' lack of a substantial vocabulary has implications for their performance on all these sub-problems; in particular, it makes them much harder.

Infants cannot recognize most of the words they encounter. Thus, unlike adults, they are not always or even usually able to use words they recognize as anchors to segment adjacent words. In fact, there are well-tested adult models, such as TRACE (McClelland & Elman, 1986) and Shortlist B (Norris & McQueen, 2008) which plausibly explain word segmentation in adults as an epiphenomenon of word recognition: “Find the word, and the boundaries come for free.” In other words, word recognition can be of substantial benefit to word segmentation if many words are known already.

However, learning words is difficult if one is not able to segment already. This is because most novel word types are not presented to infants in isolation. In fact, only about 9% of novel

types are presented in isolation (Brent & Siskind, 2001), even when caregivers are explicitly instructed to teach new words to their children (Aslin et al, 1996). Thus, the vast majority of word types that infants eventually learn must have been segmented out from a multi-word utterance. In other words, word learning appears to require word segmentation.

If word segmentation is driven (only) by word recognition, as models such as TRACE presuppose, we are forced ineluctably to the conclusion that word segmentation also requires word learning. That is because word recognition requires knowing some words, and knowing some words requires having learned them. In other words, under this assumption, not only does word learning require word segmentation, but word segmentation requires word learning. For this reason, word segmentation and word learning are referred to as a *bootstrapping* problem – two related problems, each of whose solution appears to require having solved the other already.

Fortunately, there is a way out of this logical problem – infants appear to be able to segment even in the absence of substantial word knowledge. In other words, segmentation is possible even in the absence of word recognition. This conclusion is warranted by two kinds of facts. First, a wide range of laboratory studies, reviewed below, converge on the conclusion that infants begin to evince substantial word segmentation abilities between 6 and 10.5 months of age. Second, mothers' reports reveal that during this time window, infants know (understand the meanings of) between 10 and 40 words on average. Thus, the literature on infant speech perception suggests that segmentation is largely a prelexical process: “Find the boundaries, and the word comes for free.”

The problem, then, is how infants infer word boundaries without knowing (very many)

words. A major step forward came with the seminal study by Saffran, Aslin, & Newport (1996), who showed that infants use prelexical statistics in the speech stream to segment it into word-sized units. Subsequent work in this line of research – of which this dissertation is a part – has been aimed at discovering which statistics and representations it is that infants use to discover word boundaries.

There is still something of a logical problem with this formulation, however. Essentially, infants are looking for the statistical signature of word boundaries. This statistical signature arises from the fact that languages have regularities in the sound sequences that make up words. For example, /h/ occurs quite frequently in English word-initially (*who*, *how*, *human*, *hat*, etc..), but may not occur foot-internally, so the occurrence of /h/ is a strong indicator of a word onset, even though it can occur word-medially *(mahogany*, *vehicular).* A listener who is equipped knowledge of these regularities, known as *lexical phonotactics*, can make generalizations about which sound sequences are likely to constitute words. But how can an infant who does not possess a sizable lexicon possibly estimate the statistical signature of word boundaries?

To summarize, word segmentation is hard for infants because they don't know very many words. One consequence is that they cannot usually segment speech by recognizing words, as adults appear to do. The simplest remedy to this problem would be to learn more words, but since caregivers do not typically present novel words to children in isolation, it appears that they must be able to segment in order to learn more words. Thus, infants must make use of some kind of prelexical cue to segment. And again, because they do not know very many words, they are not able to estimate the statistical signature of a word boundary from the words they know. In every

case, the lack of specific word knowledge impedes word segmentation. Infant word segmentation is a more difficult problem than adult segmentation, since the infants know less than adults, but must learn more about the words they hear. To address this problem productively, it will prove helpful to discuss what a 'word' is.

What is a Word?

Morphologists are in general agreement that there is no single, cross-linguistically applicable, and unproblematic definition of word (Bauer, 1988; Lieber & Stekauer, 2009). This is an area of active and ongoing research, so it is beyond the the scope of this dissertation to review the state of the field; thus here I simply give a brief overview.

Two conceptually distinct but related 'clines' can be discerned in the morphological literature: productivity, and number of words. Productivity refers to the frequency with which a process applies to create new sequences; for example *-ness* can be suffixed to nearly any English noun or adjective, whereas new words suffixed with *-th* almost never appear (Baayen, 2003). Number of words refers to the number of words in a sequence; for example *cat* clearly consists of one word and *middle management* clearly consists of two words. I will show below that these clines are related but dissociable. That is, it is more often the case that relatively unproductive processes result in sequences that have most or all of the properties of single words, and it is more often the case that highly productive processes tend to result in sequences that have most or all of the properties of multi-word sequences; but there are examples at all four extremes. Fig 1.1 may serve to illustrate this claim:

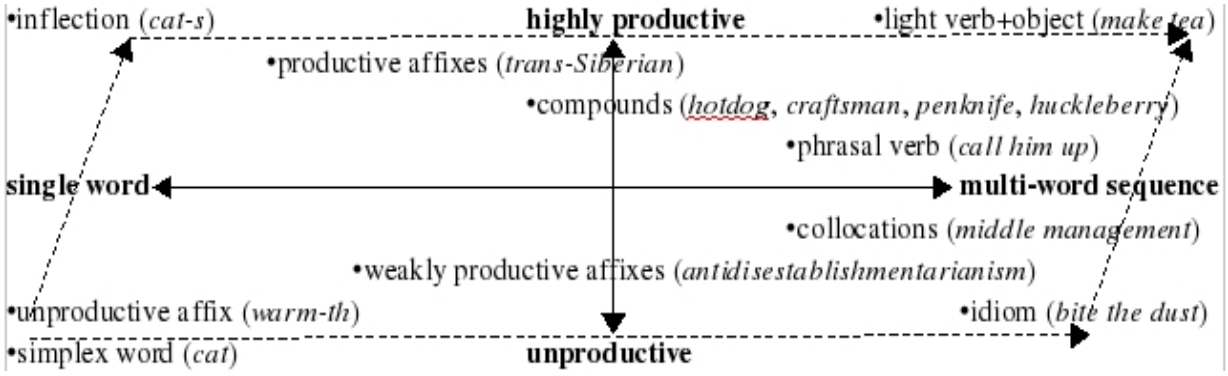


Fig 1.1: Dissociable productivity and number-of-words clines

Note that the relative positions of items in this diagram are not intended as strong claims about relative productivity or number of words; they are intended to illustrate dissociability of the two clines, as described in more detail below.

Productivity Cline

The productivity cline has most recently been the subject of intensive research by Hay and colleagues (Baayen, 2003; Hay, 2003; Hay & Baayen, 2002; Hay, Pierrehumbert, & Beckman, 2004; Hay, 2007), focusing especially on complex words, which can be provisionally defined as consisting of 1 free morpheme and 1 or more affixes (but also including cases in which an affix is affixed to a morpheme that does not occur in isolation, e.g. *obtuse*, *infer*). These researchers have identified a host of properties that correlate with each other, given below with examples that illustrate the contrast:

Property	**complex/decomposable**	**less decomposable**
junctural phonotactics	*obtuse*	*obdurate*
semantic transparency	*disarmament*	*government*
base/derived relative frequency	*softly* (*soft* > *softly*)	*swiftly* (*swift* < *swiftly*)
phonetic force of affix	*unmetalled* (220 ms)	*unfortunately* (60 ms)
affix productivity	*-ness*	*-ism*

Table 1.1: Productivity cline properties (Hay, 2003; Hay & Baayen, 2002; Hay, 2007)

Complexity, or decomposability, is typically assessed with an experimental task, in which participants are asked to rate the complexity of a form (after a pre-test with clear-cut examples such as *cat* ~ simplex and *cat-like* ~ moderately complex) (Hay, 2003). Junctural phonotactics refers to how typical a sequence is across word junctures (Pierrehumbert, 2001; Hay, 2003; Hockema, 2006); it is the measure which is precisely formalized in the theory of Diphone-Based Segmentation (DiBS) presented in Chapters 2-4. Semantic transparency, also called compositionality, refers to the extent to which the meaning of a whole can be predicted from the meaning of its parts and the nature of their combination. For example, *happiness* can be transparently explained as the abstract property of being happy; *crystal clear* has the literal meaning of 'as clear as crystal', from which the more typical meaning of 'very clear' can be easily derived from pragmatic principles; *wireless* has the literal meaning of 'without wires', from which the more typical meaning 'able to communicate without physical contact' might be derived pragmatically but certainly less easily; and finally *confuse*, whose typical meaning bears no transparent relationship to the meanings of *con-* 'with, together' or *fuse* 'meld, join'. Semantic transparency can be assessed using dictionaries under the assumption that forms with

noncompositional meanings are more likely to be listed separately and/or with a greater number of distinct senses (Hay, 2003). Phonetic force of the affix refers to vowel duration of the single suffix *un-* in Hay (2007); it is called phonetic force here because these results presumably will generalize along the lines of articulatory effort. Affix productivity refers specifically to the category- and hapax[4]-conditioned measures discussed in Baayen (2003) and tested extensively in Hay & Baayen (2002). The most interesting conclusion from these studies is not that morphological processes may vary along these dimensions – that was well-known before (cf. Bauer, 1988) – what is interesting about these studies is the incredible predictive power each measure has for the others (Hay & Baayen, 2002, Table 3).

Number-of-Words Cline

A cline can be discerned in the number of words, as well, as exemplified in the following table:

simplex	**complex**	**compound**	**collocation**	**fixed phrase**
dog	doggish	doghouse	dog-tired	Dog is man's best friend.
cat	cat-like	catseye	feline grace	Cats have nine lives.
dish	dishless	dishwasher	satellite dish	Revenge is a dish best served cold.

Table 1.2: Number-of-words cline examples

At one end of the cline, simplex items are canonical single words, with no discernible subparts. At the other end of the cline, fixed phrases clearly contain more than one such simplex item,

4 A *hapax* type is a word which occurs only once in a given corpus (Baayen, 2001).

possibly with additional subparts. In the middle of the cline lie compounds, which differ from simplex words in possessing two discernible sub-parts that otherwise can occur as simplex/complex words, but nonetheless share many properties of single words. Complex words are between simplex words and compounds, in that they usually contain a single free morpheme (like simplex words) and additional parts (like compounds), but unlike compounds, the additional parts are bound affixes, i.e. parts that may not occur without a word to affix to. Collocations are between compounds and fixed phrases in that they typically consist of multiple free morphemes, but are generally more separable than compounds, in ways that will be made more precise below.

There are a variety of properties in the literature that separate this cline (defined below):

	simplex	complex	compound	collocation	fixed phrase
Isolable sub-parts	0	1	2-	2+	2+
Single stress/accent	1	1	indef.	indef.	>1
Uninterruptable	yes	usually	usually	sometimes	sometimes
Minimal movable unit	yes	yes	usually	sometimes	rarely
Minimal elidable unit	yes	yes	usually	sometimes	rarely

Table 1.3: Number-of-words cline properties

An isolable unit is a unit that can be pronounced in isolation. For example, *dog* can be pronounced in isolation, so it is isolable; in contrast, *-ish* is not isolable. Thus, *dog* has no proper sub-parts that are isolable, where *doggish* has one proper sub-part that is isolable. Note that function words such as *the* and *he* are not isolable (except under narrow focus, under which nearly every linguistic unit is isolable); thus this criterion does not distinguish between bound

affixes and functional items that are otherwise analyzed as free.

A more specifically phonological criterion is that simplex and complex words typically bear a single accent. For example, *cat* and *catty* both have a single stress on the leftmost syllable, whereas *nine lives* has two stresses in my idiolect. This criterion is somewhat problematic for compounds and collocations, where intuitions appear to vary considerably. For example, Lieber & Stekauer (2009) cite *apple pie* as right-stressed and *apple cake* as left-stressed, although it might be more satisfactory to count this as a prominence contrast (Ladd, 1996); indeed, there is a lively literature which attempts to predict the location of prominence in compounds and other collocations (Giegerich, 2004; Jones, 1969; Liberman & Sproat, 1982; Marchand, 1960; Olsen, 2000; Plag, 2006; Sampson, 1980), including some literature documenting contextual variability in prominence assignment (Bauer, 1983; Kingdon, 1958; Roach, 1983; Stekauer, Valera, & Diaz, 2007). From a phonological standpoint it is appealing to analyze compounds such as *doghouse*, *firetruck*, *penknife*, and *raspberry/cranberry/huckleberry* in the same way as collocations, since outside of these compounds, the less-prominent vowel typically occurs under primary lexical stress. Moreover, it is possible that multi-word phrases often have multiple pitch accents that could help to decompose the speech stream in ways not explored here (Pierrehumbert, p.c.). But I do not wish to take a strong stance on these issues at present. The relevant point is that while simplex and complex words clearly possess a single accent and fixed phrases clearly possess multiple accents, the facts are decidedly less clear for compounds and collocations, and merit further theoretical and empirical research – a point to which I will return in Chapter 5.

Various syntactic criteria have also been proposed to distinguish the number-of-words

cline. In particular, simplex/complex words can be distinguished from many multi-word sequences by the inability to insert intervening material. For example, it is not grammatical to put any material inside the simplex verb *call* without destroying it (**ca-boyfriend-ll*), whereas the phrasal verb *call up* can take its object between the verb and the particle (*call my boyfriend up*). This metric is not absolutely diagnostic in English, as the process of expletive-infixation in English can split morphemes, for example *absofuckinglutely*, as it operates on the phonological domain of the foot rather than a morphological domain (McCarthy, 1982). Moreover, while collocations typically allow intervening material, it often destroys any noncompositional meaning, e.g. *bread and butter science* means 'everyday, normal, prototypical science' whereas *bread and creamy butter science* is not really interpretable without a special discourse setting that renders some contextual support to an appropriate metaphorical meaning for *creamy butter*.

Finally, simplex and complex words can be diagnosed as minimal syntactic constituents for various operations. For example, sub-parts of a word cannot be fronted: *Bagels, I like* ('I don't like those other things but I like bagels') but **-s, I like bagel* (*'I don't like just one bagel, but I like multiple bagels'). No such restriction applies to collocations, e.g. *Middle is the kind of management he is*. Similarly, sub-parts of a word cannot be elided, but parts of collocations can be elided:

(6)	Q:	What level of management is he?	Q:	Did you encode it?
	A:	Middle.	A:	*No, de

Relatedly, one-substitution is not generally possible for compounds, e.g. *the black one* can refer to a black bird, but **the blackone* cannot refer to a blackbird. But even this test has issues, e.g. *He wanted a riding horse, as neither of the carriage ones would suffice* when *riding horse* otherwise behaves like a compound (Bauer, 1988).

In summary, there is a cline from unambiguously simplex to unambiguously multi-word sequences. There are a variety of phonological, semantic, and syntactic criteria which tend to separate the cline at varying points, and with varying degrees of reliability; in practice it is easy to distinguish one word from two in many cases, but compounds and collocations are a gray area. With this background in hand, I now turn to what is known about how infants segment words.

Developmental trajectory of word segmentation

Broadly speaking, the acquisition literature shows that infants acquire impressive word segmentation abilities between 6 and 10.5 months of age. This section reviews this developmental literature in detail. However, it begins with a brief overview of infant methodology, so that it is clear how the infant data were collected.

Methodology

Because infants are not capable of (or interested in) following directions, infant perception must be assessed indirectly. Typically, the researcher attempts to manipulate and/or measure the infant's attention; the vast majority of infant studies on word segmentation use a *looking time* paradigm (Fantz, 1964) Looking time paradigms are based on the assumption that

infants gaze at objects they are attending to, so that gaze duration is a proxy for duration of attention.

As a consequence, the test phase of a word segmentation experiment always involves a visual stimulus and a set of auditory stimuli. A typical physical setup is shown in Fig. 1.2:

Fig. 1.2: Laboratory apparatus for infant language development studies

The infant is seated in an infant-appropriate chair or caregiver's lap, facing toward a visual display, with mounted speakers on either side. (The location of the speakers and screen may vary with the paradigm).

A typical test phase will involve a single visual stimulus and two different auditory stimuli. For example, the visual stimulus might be a picture of a flower, and the auditory stimuli might be the words *bin* and *din*. In the simplest case, the experimenter would assess the infant's preference for these with separate test trials, in which each auditory stimulus is repeated with short pauses between repetitions. That is, in one test trial, the infant would see the flower and hear *bin.. bin.. bin*; in another test trial the infant would see the same flower and hear *din.. din..*

din. If the infant exhibits a difference in gaze duration between these two test trials, it is likely due to a preference for one auditory stimulus over the other (since the visual stimulus is the same).

The most common variation on this basic paradigm is known as a familiarization-preference test. In this variant, the infant is first familiarized to a passage containing a target word or set of words in fluent speech. Then, during the test phase, the infant is assessed for preference of the target versus some competitor, e.g. a part-word which is consistent with an incorrect segmentation of the familiarization passage. Other variants of this basic paradigm involve habituation and/or conditioning (e.g. Werker & Tees, 1984). In a habituation paradigm, the infant is first habituated to one auditory stimulus, then presented with the alternative stimulus; habituation is especially suitable for determining whether infants can discriminate two stimuli, as a strong recovery in looking time occurs when the change is detected. In a conditioning paradigm, the infant is conditioned to turn their head upon detecting something in the auditory stimulus such as a change; conditioning is accomplished by means of toy which lights up and emits sounds when the target is presented. A concrete example is given below, in the form of a description of Saffran, Aslin, & Newport's (1996) seminal segmentation study.

The experimenters in Saffran et al. (1996) constructed an artificial language made up of 4 words: *pabiku*, *tibudo*, *golatu*, and *daropi*. Two-minute passages were constructed by concatenating tokens of these words, subject to the constraint that the same word was not repeated directly after itself. These passages were synthesized with flat prosody to eliminate normal boundary cues such as stress, pitch reset, and phrase-final lengthening, yielding strings

like the following:

(7a) ..*golatudaropitibudopabikugolatupabikutibudodaropitibudodaropipabikugolatudaropi*..

For the reader's convenience, the string below uses both text formatting (underlining, etc...) and colors to indicate the appropriate segmentation:

(7b) ..*golatu*daropi*tibudo*pabiku*golatu*pabiku*tibudo*daropi*tibudo*daropi*pabiku*golatu*daropi*..

At test, infants were presented with two types of stimuli: words and part-words. Words were actual words in the target language. Part-words were sequences consisting of the end of one word and the beginning of another, e.g. *tupabi* from *gola**tu** + **pabi**ku*. Thus, infants were exposed to both words and part-words during familiarization, and the test phase was designed to determine whether they attended differently to these two classes of stimuli. In fact, 8-month-old infants reliably distinguished the words from the part-words. Saffran et al. (1996) concluded that infants were segmenting words based on statistical properties of the speech stream (since there were no other properties, such as prosody, that differentiated words from part-words).

Having outlined the general pattern in infant segmentation studies, I will generally omit discussion of the methodology, except where it is especially relevant, with the aim of maintaining focus on the theoretical import of the findings.

Trajectory

Already by 4.5 months infants recognize the sound pattern of their own name (Mandel, Jusczyk, & Pisoni, 1995). Specifically, the researchers found that infants preferred their own name to foils that were matched and mismatched in stress template (e.g. Johnny prefers *Johnny* to *Bobby*). Strictly speaking, this study is not evidence for word segmentation at 4.5 months. This is because infants do not actually need to segment anything to succeed at the task. The test phase presents the name in isolation, so the test phase does not require segmentation. And since adults are likely to frequently call their infant's name, infants are likely to hear their name in isolation frequently. Thus, this study shows that infants recognize sound sequences that they are likely to have heard in isolation. And since it is the earliest reported effect, we may regard 4.5 month as the lower cutoff; prior to this age there is no evidence of segmentation at all.

The earliest age at which *bona fide* segmentation is reported is 6 months. Bortfeld, Morgan, Golinkoff, and Rathbun (2005) showed that infants used their own name to segment the following word. Specifically, they exposed infants to sequences like *[name]'s bike* and *[foil]'s cup*, where [name] refers to the infants own name and [foil] refers to some other (stress-template-matched) name. At test, infants were assessed for preference of the targets in isolation. Infants preferred the familiar words that co-occurred with their own name to the alternate words that co-occurred with the foil, showing that they were able to segment words that occur after their own name. However, infants did not prefer the alternate words to control words they were not familiarized to, suggesting that they did not segment the alternate words. In other words, the presence of a familiar word (their name) facilitates segmentation of the following word. The

utility of this segmentation strategy is somewhat limited, however: although infants are likely to hear their name quite frequently, it is unlikely to co-occur with most of the novel words they occur.

In contrast, English-learning infants have learned to use a slightly more general cue by 7.5 months – stress (Jusczyk, Cutler, & Redanz, 1993). Because the dominant stress pattern of English words is for the primary stress to fall on the initial syllable, infants can do fairly well by positing word boundaries at the onsets of stressed syllables. (Specifically, Cutler & Carter (1987) showed that 90% of the content words in a spoken corpus were stress-initial.) Jusczyk, Houston, & Newsome (1999) showed that 7.5-month-olds do exactly that, using stress to segment novel words from fluent passages. Crucially, they found that infants exhibit this metrical strategy even when it yields the incorrect segmentation. For example, when infants are faced with a novel noun such as *guiTAR* that violates the general strong-weak stress pattern, they appear to treat *TAR* as the onset of a novel word. Thus, infants segmentation abilities are no longer limited to the item-specific recognition evident at 6 months; by 7.5 months they have acquired the generalization that English words exhibit a strong-weak stress pattern and exploit it for word segmentation.

By 8 months, infants make use of some kind of sublexical statistic such as transitional probability[5] to segment novel words from an artificial language (Saffran et al., 1996). Subsequent research using the same paradigm has shown that such statistical cues are not heavily weighted in comparison to other linguistic cues, such as coarticulation[6] and stress (Johnson & Jusczyk, 2001).

5 Transitional probability $p(x \rightarrow y)$ is the conditional probability of observing y as the next event, given that the current event is x.

6 By coarticulation, the authors meant fine phonetic variation that would not distinguish phone categories, such as the natural acoustic effect of a consonant on a vowel in the preceding syllable. They did not study coarse variation such as place assimilation, which would distinguish phone categories.

While these results raise important questions about the nature of the cues that infants really attend to, they do not directly contribute to our understanding of segmentation. This is because in normal usage, linguistic cues rarely compete with one another, rather exhibiting "confluences across levels" (Pierrehumbert, 2003) which generally line up to suggest the correct parse.

By 9 months, the statistical cues that infants make use of include native-language phonotactics (Friederici & Wessels, 1993). This is the earliest that infants have been shown to know the difference between native and non-native phonotactics (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993), and in fact, they are sufficiently sensitive as to prefer high-frequency sequences over low-frequency sequences (Jusczyk, Luce, & Charles-Luce, 1994), a preference which they do not exhibit at 6 months.

Of special relevance for this dissertation are a sequence of studies by Sven Mattys and colleagues, demonstrating how infants' burgeoning knowledge of phonotactics is recruited for word segmentation. The starting point for this research is the observation that most diphones – sequences of two segments – exhibit highly constrained distributions; more specifically, a diphone normally occurs either within a word, or across a word boundary, but not both (Hockema, 2006). For example, the sequence [tp] occur almost exclusively across word boundaries, whereas the sequence [ba] primarily occurs word-internally. For this reason, this dissertation may refer to word-internal diphones as WI and word-spanning diphones as WS.

Mattys, Jusczyk, Luce, & Morgan (1999) exposed infants to CVC.CVC nonwords, manipulating both the medial C.C cluster and the stress pattern. They found that infants' preference was modulated most strongly by stress pattern, but was also sensitive to diphone status

(WI or WS). Since raw diphone frequency was controlled, this indicates that infants are sensitive to the relative frequency with which diphones span word boundaries specifically.

In a follow-up study, Mattys & Jusczyk (2001) showed that infants use this diphone cue to segment novel words from an utterance-medial position. They did this by familiarizing infants to passages in which a novel target word (*gaffe* or *tove*) was embedded in a context consisting of two novel words. For a word in the WS condition, the context was chosen so that the diphones at both edges of the target were word-spanning. For a word in the WI condition, the context was chosen so that the diphones at both edges of the target were word-internal. In other words, the WS context facilitated segmenting the target out as a novel word, whereas the WI context inhibited segmenting the target out as a novel word. Example stimuli in English orthography are given below in Table 1.4:

target	**WS**	**WI**
gaffe	... bean gaffe hold ...	... fang gaffe tine ...
tove	... brave tove trusts ...	... gruff tove knows ...

Table 1.4: Stimuli in Mattys & Jusczyk (2001)

Infants were divided into two groups; one group heard *tove* in the WS context and *gaffe* in the WI context, and the other group heard *tove* in the WI context and *gaffe* in the WS context. Infants consistently preferred the target word that was embedded in the favorable WS context to the target that was embedded in the unfavorable WI context, and they did not differ in preference between the target in the unfavorable WI context and control words to which they were not

familiarized. Thus, this study showed that infants exploit phonotactic cues at word edges to segment novel words from an unfamiliar, phrase-medial context.

By 10.5 months of age, infants are able to exploit ever more phonotactic information. In particular, they are able to use allophonic variation[7] which cues word boundaries, appropriately segmenting *night rate* as two words but recognizing *nitrate* as a single word (Jusczyk, Hohne, & Baumann, 1999). In addition, they have learned to integrate phonotactic cues with prosodic cues, correctly segmenting weak-strong nouns such as *guitar* where 7.5 months mis-segment this word because of its atypical stress pattern (Jusczyk et al, 1999b).

In summary, infants undergo a rapid developmental shift between 6 and 10.5 months of age. At 6 months of age, infants have just learned to use their own names as anchors for segmenting the following words. By 7.5 months of age, they exploit the dominant strong-weak stress pattern of English to posit word boundaries at the onsets of stressed syllables. Between 8 and 10.5 months of age, infants exhibit increasing command of the phonotactics of their language, learning which sound sequences are frequent in their language and which are likely to signal the presence of a word boundary. Crucially for this dissertation, by 9 months of age, they pass the 'acid test' of segmenting novel words from novel contexts using their knowledge of word-spanning diphones. Thus, by this age phonotactic knowledge has become an important component of how infants' segment speech prelexically; this knowledge is from generalizations

7 The term 'allophonic variation' is used here as it is standardly used in linguistic theory (Hyman, 1975), to indicate forms which in the adult grammar originate from the same phoneme(s), but differ in their surface realization owing to prosodic or other factors. In this case, *nitrate* and *night rate* both contain a /tr/ sequence, but the phonetic realizations differ because the two consonants are syllabified together in the onset in *nitrate*, but split into different syllables in *night rate*, with attendant phonetic consequences. The use of the term 'allophonic variation' here is not intended to imply that infants actually know that these two sequences are phonemically identical; merely that whatever causes this variation, infants use it.

over the phonological structure of their language, which is how infants can segment speech even when they do not recognize all of the words they hear.

Theoretical Models

In recent years, there has been a flurry of research on word segmentation, from a variety of different theoretical perspectives. For theoretical convenience, I have classified existing models into four categories: *connectionist*, *coherence-based*, *bayesian-joint*, and *phonotactic*. I will analyze each of these classes in turn, describing its major properties, the insights it has produced, and limitations. Before this, however, I will describe the properties I see as important for a model of word segmentation.

Desiderata

Before discussing particular models or classes of models, it may be useful to consider the properties that we desire in a model of word segmentation. The overarching criterion I will adopt is that the model be *cognitively plausible*, i.e. that it incorporates reasonable assumptions about the kinds of input, representations, and computations that listeners are able to perform. This general criterion can be broken down into several more specific properties, whose names I have italicized for later reference.

A fundamental property of human language acquisition is that it is largely *unsupervised*, meaning that children must infer the correct solution on their own, rather than being told by a caregiver. For example, in word segmentation, adults do not normally indicate the position of the

word boundaries in what they said (e.g. by snapping their fingers between words). All the models I consider below are unsupervised,[8] thus I do not discuss this property further.

Another basic property of a cognitively plausible model is that it explains the target behavior (word segmentation) in some baseline case; in other words, the model must be testable. In fact, it is desirable that model not only be testable in principle, but that it actually be computationally *implemented*. By implemented, I mean that a working piece of software exists which implements a segmentation proposal in a paper. This provides other researchers with a way to falsify the model, by testing its performance on language data. It further provides for fair comparisons with other published implementations.

Beyond the basic ability to test a model, there is a further criterion of what kinds of language data it has been tested on. As will be evident from the discussion below, the great majority of published models have only been tested on English data, or artificial languages with English-like phonological properties. Thus it is unclear whether the success of these models owes to specific properties of English, or whether the model is of language-general applicability. This is of especial concern for English since it is relatively unusual both in its impoverished inflectional morphology and its highly complex phonotactics, both of which could plausibly affect the segmentation models discussed below. Since the research problem in this dissertation is the acquisition of word segmentation, which infants of all languages must solve, only models of the latter, language-general class are candidates for a general solution. The best way to

8 Some of the models I consider treat phrase boundaries as word boundaries, and utilize this information for word boundary inference elsewhere. This could be regarded as supervised data, since the algorithm is being given some input/output pairs in which the presence of the "output" (a word boundary) is certain. However, I do not regard this as supervised learning, since the adult does not *tell* the child that phrase boundaries are also word boundaries. Rather, this is an inference that the child appears to make on their own (Soderstrom, Kemler-Nelson, & Jusczyk, 2005).

determine this is to test the model on *cross-linguistic* data, i.e. language data from any languages other than English.

As discussed above, word segmentation and word learning are not independent problems. Thus, a full developmental account of segmentation should also include some account of *lexical* acquisition, i.e. word learning. One strategy that computationally-minded researchers have taken is to treat these two problems as a *joint* optimization problem, in which both word segmentation and word learning are accomplished by the same algorithm. The alternative strategy is to treat these as *related* but logically distinct problems, and then to specify the relationship between lexicon and segmentation.

Finally, a cognitively plausible model should be *incremental* since human language acquisition is. A batch model is one which segments an entire input corpus all at once. In contrast, an incremental model develops in response to language input, accepting input at some child-realistic scale, and then modifying its internal representations and segmentation processes according to its language exposure. Of course, some batch models can in principle be converted to incremental models simply by feeding in input in smaller chunks. Thus, the operational criterion for whether a model is incremental or not is whether the model requires batch input, as described in a publication.

Connectionist Models

The first connectionist model that bears on word segmentation is Elman's (1990) seminal paper describing the Simple Recurrent Network (SRN). Since the SRN architecture is used in

nearly all of the models reviewed in this section, it bears describing. An SRN is similar to a standard 3-layer feedforward network, consisting of an *input* layer, a *hidden* layer, and an *output* layer. Information at the input layer (typically a binary vector) is communicated to the hidden layer via connections which stretch (multiply) the input and then squash it (via the logistic function); in exactly the same manner, information at the hidden layer is transmitted to the output layer. An SRN differs from a standard feedforward net, however, in that it possesses an additional *context* layer which serves as a short-term memory. It does this by copying the hidden layer activations from the previous time step and transmitting them to the hidden layer in the current time step[9] (see Figure 1.3 for illustration)

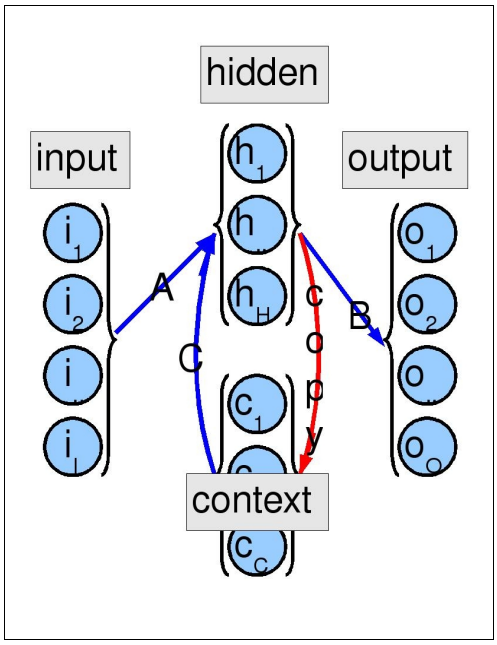


Fig 1.3: Architecture of Simple Recurrent Network

9 Since the hidden layer at the previous time step itself had access to the hidden layer from the time step before that, the hidden layer can in principle encode events and contexts over many time units (Botvinick & Plaut, 2006).

Like feedforward networks, the SRN is most commonly trained using a variant of the backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986). Backpropagation trains the network to minimize the difference between its output and the correct output, in effect "giving" the network the correct answer. Thus, backpropagation is normally regarded as a supervised training algorithm, and therefore not appropriate for modeling language acquisition. This issue is circumvented in the SRN by making the output be a prediction about the upcoming input, i.e. a *prediction* task. Since this information actually does become available to the infant, the prediction task can be regarded as an unsupervised training algorithm.

Elman (1990) trained an SRN with sequences of letters generated by connected words from an artificial (miniature) language. The crucial finding from this paper was that prediction error was strongly correlated with word position. More specifically, prediction error tended to decrease across a word, so that a sharp increase in prediction error was strongly correlated with a word boundary.

Following this line of research, Aslin et al (1996) trained a feedforward network (like an SRN, but without the "memory" of a context layer) whose input consisted of trigrams of phonetic feature bundles, together with an "utterance boundary" indicator (i.e. spatial representation of time). As per the prediction task, the network's task was to predict the value of the utterance boundary indicator. The researchers found that the utterance boundary unit fired not only at utterance boundaries, but also at utterance-medial word boundaries. In other words, this study showed unequivocally that *phonological information at the end of utterance boundaries is also*

informative for detecting word boundaries. I will return to this point in Chapter 4.

Cairns, Shillcock, Chater, and Levy (1997) focused on the role of phonotactics. They showed using an ideal-observer model (equivalent to the baseline model in Chapter 2) that diphones are an excellent segmentation cue. However, they were unable to train their SRN to take advantage of this cue in an unsupervised way. Thus, this study highlighted both the potential utility of the diphone cue, and the learnability problem of how infants could access that utility.

In a related line of research, Christiansen, Allen, and Seidenberg (1998) trained an SRN on the prediction task with a spoken corpus. Unlike previous models, this SRN was supposed to predict both the utterance boundary marker and the upcoming phoneme. To be more precise, the researchers considered a variety of conditions in which more or less corpus information was supplied to the model. For example, they contrasted cases in which the main stress was either supplied or withheld from the network. In every case, they found that the more cues the model received in its input, the better its all-around performance. Although perhaps not surprising when framed this way, it is also the case that the more cues the model gets, the more cues it must predict, and thus, the more difficult the overall task becomes. In other words, even though combining cues in some sense represents a computational burden, the SRN model nonetheless exhibits better performance when it is given access to a wider range of cues. To the extent that the SRN mimics human performance, it gives insight onto why humans are so good at segmenting speech.

The last contribution to the connectionist approach to word segmentation is Davis (2004), although this model could more properly be described as word recognition. As in the earlier

connectionist models, Davis exposed an SRN to a phonetic sequence. However, the nature of the prediction task was quite different. In this case, the network was asked to predict the semantic content of the utterance. The semantic content was coded using a localist representation in which each output node stood for a word from a mini-lexicon of 20 words. Thus, the phrase GIRL HIT BOY would be coded by activating the GIRL, HIT, and BOY nodes. By the end of training the network recognized an extremely high proportion of words as they were spoken, indicating that it had also segmented and learned them.

These are major papers in what I am calling the connectionist approach to word segmentation. In terms of the desiderata I described above, connectionist models are inherently *implemented* and *incremental*. To my knowledge, no such model has been tested on cross-linguistic data, although in principle this is possible. Most of the connectionist models described above do not have an explicit lexicon, and so do not address the related problem of word learning. However, Davis (2004) and arguably Elman (1990) achieved joint segmentation and lexicon discovery for artificial languages with small, closed vocabularies.

The major advantage of this approach is evident from the review above: each study illustrates that segmentation that some cue or set of cues is useful for word segmentation with a fairly general-purpose learning algorithm. For example, the Aslin et al (1996) study shows that utterance-boundary distributions are informative. Thus, these studies have made important contributions to the acquisition of word segmentation by demonstrating what aspects of the signal are informative.

One major limitation of the connectionist approach is that it is not usually clear *why* a given result obtains. For example, while the Aslin et al (1996) study shows that utterance-boundary distributions are informative for word segmentation, it is not clear what specific information the network exploits. For example, one kind of information that the model might extract is the probability distribution for word boundaries given the current input segment. The behavior of Aslin et al (1996)'s model is broadly consistent with the interpretation that the hidden layer encodes this distribution. But it is also broadly consistent with a variety of alternative interpretations, for example, that it encodes one of the coherence-based statistics reviewed in more detail below. This issue is not specific to Aslin et al.'s (1996) network, but is endemic to the connectionist approach, and follows from the fact that connectionist networks 'are high-dimensional systems and consequently difficult to study using traditional techniques' (Elman, 1990, p. 208). In other words, the hidden layer representations are opaque, standing in a non-transparent relationship both to input and output representations.

A second limitation to the connectionist approach is the question of how segmentation is related to lexical acquisition. Of the models described above, most do not have an explicit lexicon, so it is not clear what kind of relationship such models predict. The absence of a lexicon in these models is not an intrinsic property of connectionist models. That is, it is logically possible to extend these models to somehow incorporate a lexicon. But there is not one obvious right way forward in this line: a number of important properties would need to be worked out, such as how the budding lexicon impacts existing segmentation, and how a new word is learned.

The exception to this is Davis (2004) (and arguably Elman, 1990), in which word

segmentation is driven by word recognition. In these models, lexical representations are *de facto* stored in the weights connecting the various layers, which are the same weights that instantiate the other forms of processing that the network exhibits. Because these connections weights have a limited capacity, and since both of these networks are trained on a mini-lexicon consisting of not more than 30 words, it is unclear whether the same results would "scale up" to an open vocabulary such as infants are exposed to.

One reason to believe these results would not scale up is training time. Davis' (2004) model required 500,000 training sequences of 2-4 tokens each to learn 20 words. To put this in perspective, that's approximately *75,000 exposures per word*, whereas laboratory studies show that 14-month-olds can form-meaning associations from about 10 exposures spread out over a few minutes (Booth & Waxman, 2003), and adults are 80% successful at learning form-meaning associations with 7 exposures (Storkel, Armbruster, & Hogan, 2006). Thus, there are important differences in the amount and quality of data that yield word-learning in humans versus in existing connectionist models.

In summary, the connectionist approach is excellent for determining whether something is possible, but it offers little insight into why or how. Connectionist research on word segmentation has made several important contributions to our understanding of infant word segmentation, in particular the observations that *utterance boundaries and phonotactics are highly informative*. However, owing to the opacity of hidden layers in connectionist networks, it is not really clear what properties of utterance boundaries or what specific phonotactic cues are being used. A related issue is the role of a lexicon in word segmentation. Most connectionist models do not

have a lexicon, and while there is no reason in principle they cannot be hooked up to a lexicon, the nature of the processing relationship is not *a priori* clear and must be worked out. In existing models that do acquire a lexicon, the lexical representations are opaque, and unlikely to scale up to the open-vocabulary conditions faced by infants.

Coherence-Based Models

Responding in part to the issue of opaque representations in the connectionist approach, a number of researchers have proposed particular statistics that infants may attend to which could help them "get off the ground" in word segmentation. For example, Saffran et al (1996) proposed that infants posit word boundaries at points in the speech stream which exhibit low *transitional probability*, i.e. in which the following segment is especially unlikely to follow the current segment. The intuition underlying this and related proposals is that words are coherent units, so a statistic that measures local coherence should be greater within words than across word boundaries. The (forward) transitional probability is just such a statistic.

All of the coherence-based measures I will discuss can be defined with reference to the *unigram* and *bigram* distributions, so I will begin by defining those. The unigram frequency $f(x)$ of a linguistic unit x in some corpus is the number of times that x occurs in the corpus. For example, in the "corpus" consisting of the preceding sentence, the letter x occurred twice, so its unigram frequency is 2. Analogously, the bigram frequency $f(xy)$ of a 2-unit sequence xy is the number of times that x occurs followed immediately by y. For example, in the corpus consisting of the previous sentence, the sequence "qu" occurred twice so its bigram frequency is also 2. The

unigram and bigram distributions are the putative probability distributions that generate the unigram and bigram frequency counts in a corpus. In practice, these distributions are estimated by relative frequency, i.e. dividing the observed counts by the total frequency mass in the corpus:

$$p(x_i) = f(x_i) / \sum_j f(x_j) \qquad \text{(unigram distribution)} \qquad (1.1)$$

$$p(xy_i) = f(xy_i) / \sum_j f(xy_j) \qquad \text{(bigram distribution)} \qquad (1.2)$$

Thus, the terms *p(x)* and *p(xy)* will be referred to as the unigram and bigram probabilities, respectively.

The coherence-based statistics which have been proposed include:

forward transitional probability $FTP(x,y) = p(xy)/p(x)$ (1.3)

Saffran et al (1996), Aslin, Saffran, & Newport (1998)

pointwise mutual information $PMI(x,y) = \ln (p(xy)/(p(x)p(y)))$ (1.4)

Rytting (2004), Swingley (2005)

raw diphone probability $RDP(x,y) = p(xy)$ (1.5)

Cairns et al. (1997), Hay (2003)

Saffran and colleagues report that the forward transitional probability is motivated by the work of

Harris (1955), who reportedly used a variant of it morpheme segmentation. It is simply the likelihood of the next element, given the current one. Swingley (2005) proposed pointwise mutual information, which is similar to forward transitional probability, except that it conditions on the likelihood of both segments rather than just the initial segment; it could be thought of as a measure of the association between the two phones. Hay (2003) observed that diphones which were extremely rare were overwhelmingly likely to span word boundaries, and proposed that raw diphone probability could be used as a fallback cue when no better information was available.

Swingley's (2005) model stands apart from the other coherence-based models on nearly all the desiderata defined above: it has been *implemented* and tested on *cross-linguistic* data, and it addresses the relationship between segmentation and *lexical* acquisition as *related* problems, although as currently implemented it is not an incremental model. Thus, it worth reviewing this model in more detail.

Swingley's model tabulates syllable co-occurrence statistics and postulates words based on a combination of high frequency and mutual information. Strictly speaking, it is intended as a model of lexical acquisition rather than word segmentation proper, but I have classed it with the coherence-based models since it can be used to segment speech as well. The model works by tabulating syllabic unigram, bigram, and trigram frequencies, as well as bigram mutual information. “Likely word” percentiles thresholds are then defined as a function of number of syllables. For example, likely bisyllabic words might be defined as those whose mutual information is above the 70th percentile and whose bigram frequency is above the same unigram frequency percentile (Swingley considers a continuum of parameters in which the mutual

information criterion is linked to the unigram frequency criterion.) Longer likely words were favored by removing the sub-words they contained; for example if *dangerous* were deemed a likely trisyllable it would explain the wordlikeness of its subparts *danger* and *gerous*, so these would be removed from the likely bisyllable list.

In the best case (the 70th percentile criterion in the previous paragraph) Swingley reports about 80% of the likely words corresponded to actual words, in both English and Dutch. Of these likely words, about 150-200 (Dutch-English) were monosyllabic, about 30-60 were bisyllabic, and 2-6 were trisyllabic. Furthermore, he reports that many of the errors on bisyllables were pairs of frequently occuring words.

One of the most interesting aspects of this study was the predominance of the trochaic stress pattern in words extracted by the model. As reviewed earlier in this chapter, by 7.5 months of age English-learning infants exhibit a trochaic bias, preferentially positing word boundaries at the onset of stressed syllables. Swingley's model is provocative in illustrating that even an imperfect word learning mechanism can explain the observed metrical segmentation bias in English learning children (Jusczyk et al, 1999b) as a statistical generalization over the burgeoning lexicon (Pierrehumbert, 2001), rather than as an innately specified trochaic bias (Cutler & Norris, 1988).

With respect to this point, there are a number of issues in need of clarification and future research. For example, as reviewed above, the best evidence that we currently possess suggests that infants at this age possess a vocabulary of around 40 words (Dale & Fenson, 1996), probably too small to justify any strong generalizations about stress patterns which might be useful for

word segmentation. This immediately raises the question of how many words is *not* too small to justify this kind of statistical generalization.[10]

There is an additional question, which is whether infants at this age might 'know' significantly more words than the estimated 40 above. The best evidence we have is from caregivers' reports about the words their infant *understands*. There are two reasons why this is likely to be an underestimate. The first reason is that the infant may understand a word, but it happens that no situation has arisen in which the caregiver saw and remembered positive evidence that the infant understood it (e.g. by looking at the ball when the mother says *ball*). The second reason is that an infant may 'know' a wordform (in the sense of recognizing it as a distinct unit of their language) without understanding its meaning. For example, it is possible, even likely, that English-learning infants know that the determiners *the* and *a* are words of English, without understanding the subtle meaning contrast between these two words. Unfortunately, further consideration of these issues is outside the scope of this dissertation, so for the present purposes, I will assume that caregiver's reports (Dale & Fenson, 1996) are correct.

The general advantage of Swingley's (2005) model and other coherence-based model is that it is founded on a well-established principle of gestalt psychology: listeners group perceptual bits together based on some kind of perceptual coherence (Kohler, 1967). Thus, coherence-based models enjoy a greater degree of *a priori* psycholinguistic plausibility than other classes discussed here. The model is simple, intuitive, and makes generally reasonable learning

10 I cannot help but speculate on this point, in the hopes that an interested reader will take this question up for their own research project. One way of framing this problem would be to estimate the probability of a word boundary relative to the position of a stress, e.g. the probability of a word boundary immediately following a stress. Bayes' theorem provides a way to estimate this from the opposite conditional probability, which can be estimated from the budding lexicon via a generative model.

assumptions. Furthermore, unlike connectionist models, in which the internal representational structure is opaque, it is quite easy to "open up the hood" of a coherence-based model and investigate what it knows.

There are three disadvantage of coherence-based models. First, they are not probabilistically principled. In the case of Swingley's model in particular, the word discovery procedure includes a variety of *ad hoc* heuristics, notably equating the frequency scale of syllable bigrams with syllable unigrams, equating mutual information with these frequencies by mapping everything to percentiles, and then requiring the same mutual information percentile threshold as the frequency threshold. If the model is to be interpreted as a true model of what infants do, these would have to be counted as innate bits of the child's knowledge – along with the ultimate percentile criterion that the child adopts. Second, and more generally, coherence-based approaches attempts to recover hidden structure (word boundaries) using an incidental statistic, rather than modeling the desired structure explicitly. It only stands to reason that these models would achieve a higher level of success if they tried to find word boundaries by looking for them, rather than by looking for "coherence".

Finally, with two exceptions, coherence-based approaches have not been satisfactorily implemented. The first exception is Swingley (2005), as extensively discussed above. The other exception is Cairns et al. (1997) who implemented the raw diphone probability proposal, and found generally very poor discrimination of word boundaries from non-boundaries.[11] These

11 Yang (2004) implemented Saffran et al's (1996) proposal, but adopted an extremely ungenerous interpretation of its assumption. For instance, it posited word boundaries at local (token) minima in utterances. As Yang himself points out, this makes it impossible for listeners to segment monosyllabic words, since two adjacent boundaries cannot both be minima. As monosyllabic words form the vast majority of tokens in child-directed English, this implementation dooms the Saffran proposal to failure before it begins.

issues serve to highlight why it is important to implement a model and publish a working implementation: the lack of rigorous and fair tests of coherence-based approaches is a major impediment to evaluating these proposals on a fair playing field with the other models discussed here.

Bayesian Joint-Lexical Models

The common thread underlying what I will refer to as Bayesian joint-lexical models is a probabilistically principled framework for jointly segmenting a corpus and discovering the words in it. The general idea in these models is to specify a probability distribution over lexicons given a corpus, and then to select the *maximum a posteriori* (MAP) lexicon, i.e. the most likely lexicon given both the observed data and any prior biases the learner may have as to likely lexicons.

The first such model was Brent & Cartwright (1996), which was cast in the framework of information theory. More specifically, Brent & Cartwright described a model that operates according to the Minimum Description Length (MDL) principle. The information-theoretic basis for this formulation is that an (unsegmented) corpus can be represented (segmented) as a sequence of codes (words) from a codebook (lexicon), in which the "cost" of a particular representation is simply the cost of the codebook plus the cost of generating the corpus from that codebook. The MDL principle states that the optimal code is one which allows for the smallest total cost, which can be interpreted as the "shortest" description since the cost of a codebook is simply its length and the cost of a segmentation is the length of the encoded corpus.

I classify this model as a Bayesian model, despite the apparent lack of an explicit prior

distribution, because Goldwater showed in her (2006) dissertation that it is equivalent to a fully Bayesian model. The argument is so well-stated that I simply copy it here:

> The relationship between MDL and Bayesian inference becomes clear when we consider results from information theory. In particular, information theory tells us that, under an optimal encoding, the length (in bits) of an encoded corpus will be exactly $-lg\ p(d \mid h)$, where d is the corpus and h is the codebook used to encode d. Therefore the optimal codebook $\hbar$ will be the one that satisfies
>
> $$\begin{aligned}\hbar &= \text{argmin}_h\ \text{len}(\text{encoding}_h(d)) + \text{len}(h) \\ &= \text{argmin}_h\ -\text{lg}\ p(d \mid h) + \text{len}(h) \\ &= \text{argmax}_h\ p(d \mid h) \cdot 2^{-\text{len}(h)}\end{aligned}$$
>
> In other words, MDL is simply MAP Bayesian inference with the assumption that the prior probability of a particular hypothesized grammar (the codebook) decreases exponentially with its description length (pp. 12-13)

In fact, I will not further discuss Brent & Cartwright (1996) or its successor Venkataraman (2001) since they were superseded by the work of Goldwater and her colleagues.

Goldwater describes a two-stage generative Bayesian model. This model crucially relies on the fact that for any given segmentation of a corpus d into a sequence of words $\mathbf{w}$, there is a corresponding lexicon Ω_w and frequency distribution Y_w. The workhorse of the model consists of prior distributions on these two distributions. Specifically, the *generator* P_ω assigns probabilities

to lexical types on the basis of their phonological form, and the *adaptor* P_y assigns probabilities to frequency distributions over lexical types. Thus, the language model is given by

$$p(\mathbf{w} \mid d) = P_\omega(\Omega_w) \cdot P_y(Y_w) \tag{1.6}$$

Goldwater then defines a *search procedure* to find the most likely segmentation under this language model. In other words, this model measures the probability of a segmentation by the prior probability of the lexicon it induces.

The hypothesis space for the search procedure consists of all possible word boundaries in the corpus. The search procedure uses Gibbs sampling with simulated annealing Markov Chain Monte Carlo (for a detailed exposition see Goldwater, 2006). What this means is that the algorithm iterates through every possible word boundary in the corpus and considers two alternatives: word boundary or no word boundary. Changes are accepted or rejected stochastically according to whether they improve the prior likelihood of the segmentation, which can be cleverly updated using only local information. Since the search procedure is not of central interest here, and Goldwater goes to considerable lengths to demonstrate that it does not impose any additional biases in the solutions identified, I omit any further description. The essential point is that the search procedure finds or approximates the most likely segmentation according to the model described above.

In the simplest model she describes, Goldwater uses a unigram model for the *generator* P_ω. This model assigns a hypergeometric distribution over word lengths and a uniform

distribution over all possible strings of a given length. Thus,

$$P_{\omega}(\omega = \phi_1\phi_2...\phi_n) = (1-p_{\#})^{n-1}p_{\#} * (|\Phi|^{-1})^n \quad (1.7)$$

where $p_{\#}$ is the probability of a word boundary and $|\Phi|$ is the number of phones (so that $|\Phi|^{-1}$ is the uniform probability of a phone).

In the simplest model, Goldwater uses a Chinese Restaurant Process (CRP(α)) as the *adaptor*. This is a probability distribution over partitions of the integers, but it can most easily be described by a stochastic process. Imagine a restaurant with an infinite supply of tables, each of which can seat an infinite number of customers. When a new customer comes in, they must choose a table. They can choose to sit at any of the occupied tables, or to sit at the next unoccupied table. Suppose further that the customer chooses occupied tables with a probability that is proportional to the number of customers it already has, and they choose new tables with a probability that is proportional to some constant α, known as the concentration parameter. For the case of language, 'customers' correspond to word tokens and 'tables' correspond to word types. Thus, words which are highly frequent are more likely to recur, but there is always some probability of a novel word. Formally, let:

k	type (index)
z_i	type index of the i^{th} token
$\mathbf{z}_{-i}$	sequence of types observed before the i^{th} token

$K(\mathbf{z}_{-i})$ number of types observed before the i^{th} token

n^{k}_{-i} frequency of k^{th} type before the i^{th} token observed

α concentration parameter (1.8)

The probability that the i^{th} token is a member of the kth type is

$$p(z_i = k \mid \mathbf{z}_{-i}) = f(k)/(i-1+\alpha) \quad (1.9)$$

where $f(k) = n^{k}_{-i}$ for $1 \leq k \leq K(\mathbf{z}_{-i})$ and α for $k=K(\mathbf{z}_{-i})+1$ (i.e. a new table). (The term $i-1+\alpha$ is simply the normalizing constant to make this a probability distribution.) The probability of a sequence is simply the product of the probabilities of each element, given the previous sequence:

$$\begin{aligned} CRP(\mathbf{z},\alpha) &= \prod_{i \leq n} p(z_i \mid \mathbf{z}_{-i}) \\ &= \Gamma(1+\alpha)/\Gamma(n+\alpha) \cdot \alpha^{K(\mathbf{z})-1} \cdot \prod_{k \leq K(\mathbf{z})} (n^{k}_{\mathbf{z}}-1)! \end{aligned} \quad (1.10)$$

where Γ is the generalized factorial (gamma) function. Thus, the Chinese Restaurant Process assigns probabilities to sequences of words based on the principle that words recur with probability that depends on their frequency.

Although the technical machinery of this model is fairly involved, the intuitions are simple: there is a prior distribution over wordforms (*generator*), and a prior distribution over word frequency distributions (*adaptor*), and the optimum segmentation is the one which

maximizes the joint probability of these two prior distributions. Wordforms are assigned a probability, in most of the experiments she reports, based on a geometric distribution over their length, i.e. words of length 3 are half as likely as words of length 4, which are half as likely as words of length 5, etc... The adaptor favors Zipfian distributions in which a few elements occur many times and many items occur few times (Baayen, 2001), as controlled by a hyperparameter α. The model's search procedure is designed to identify the segmentation which maximizes the prior probability of the corpus according to the product of these two prior distributions.

Goldwater and colleagues obtained two deep results with this model. First, she showed that the basic model described above (with limited modifications to handle corpora with utterance boundaries, i.e. unambiguous word boundaries) exhibited superior performance to that of then-extant models of the same class, namely Brent (1999) and Venkataraman (2001). More specifically, as shown in Table 1.5, she found that the search procedure used in these other models *does* introduce unintended biases. She found this by showing that the solution her own model found (**bold**) had higher probability than the optimum found by their own search procedure (underlined):

Model	True	None	Brent	Venkataraman	Goldwater
Brent	208.2	321.7	217.0	218.0	**189.8**
Venkataraman	204.5	90.9	210.7	210.8	**183.0**
Goldwater	222.4	393.6	231.2	231.6	**200.6**

Table 1.5: Scaled negative log-likelihood (score) of segmentations (columns) under different models (rows) (Table 5.3 from Goldwater, 2006, p. 119)

(In Table 1.5, *True* refers to the correct segmentation, *None* to the segmentation in which only utterance boundaries are marked, and *Brent/Venkataraman/Goldwater* to the optimum segmentation produced by the model's search procedure. Higher probability means lower negative log-likelihood.) This finding is important because it demonstrates how crucial the search procedure is Bayesian models of this type, in particular how it may introduce biases which are not explicitly included in the language model proper. More specifically, it illustrates that Brent's (1999) model and successors are flawed precisely because of this search bias.[12]

The other deep result is that undersegmentation is the empirical consequence for models which assume that a word is independent of the word preceding it. Although this independence assumption is obviously false, Goldwater showed that its falseness has consequences for word segmentation. This is because language possesses a large number of collocations, multi-word sequences that co-occur much more frequently than expected under independence; or in other words, they behave like the model expects words to do. Thus, the model assigns higher probability to segmentations in which collocations are segmented as a single word. In fact, Goldwater showed this by extending her model to a bigram (hierarchical Dirichlet process) model, which improved both precision and recall.

Another study in this general framework was conducted by Blanchard & Heinz (2008). In terms of Goldwater's (2006) description, their work could be described as enriching the generator. (The lexical generator that Goldwater adopted is phonologically primitive, assigning

12 The search procedures in Brent (1999) and Venkataraman (2001) make use of a number of approximations to allow for a dynamic programming approach. This is the sense in which the explicit prior differs from what the model does.

equal probabilities to any lexical type consisting of the same segments, e.g. it assigns equal probability to the real but not phonotactically equivalent words *acts*, *asked*, *axed*, *cats*, *cast*, *scat*, *stack*, *sacked*, *task*, *tacks*, the phonotactically licit non-word *atsk*, and various phonotactically ill-formed words such as [kstæ] and [ætks]). Blanchard & Heinz adapted the Brent's (1999) incremental model to bootstrap its lexicon and lexical phonotactics off each other, achieving generally superior performance relative to Goldwater (2006). Although this model presumably suffers from the same search issue as Brent (1999), it is nonetheless informative in demonstrating the utility and importance of phonotactics.

A related finding was reported by Johnson (2008) using Johnson, Griffith, & Goldwater's (2007) *adaptor grammar*. Essentially an adaptor grammar is a generalization of Goldwater's (2006) approach to probabilistic context-free grammars. In other words, an adaptor grammar specifies a non-parametric Bayesian model over hierarchical segmentations (trees) of an input corpus, rather than non-hierarchical (flat) segmentations. Thus, adaptor grammars provide for models of richer linguistic structure, in particular syllable structure; but in terms of desiderata, they are analogous to Goldwater's basic model.

In summary, all Bayesian joint-lexical models have a common Bayesian underpinning, are *implemented*, achieve lexical acquisition by simultaneously optimizing word segmentation and *lexical* acquisition (hence “Bayesian joint-lexical”), and were not tested on cross-linguistic data by their author.

Most Bayesian joint-lexical models are not incremental. One exception is Blanchard &

Heinz's (2008) model, which is adapted from the earlier incremental model of Brent (1999) model. As Goldwater (2006) demonstrated, in current-generation incremental models of this type, the search procedure imposes substantial biases on the solution, over and above the model's explicit priors. Thus, Bayesian joint-lexical models satisfy all of the desiderata, except perhaps being *incremental*.

Beyond merely satisfying the desiderata, the Bayesian joint-lexical approach is conceptually elegant. For example, the coherence-based approach attempts to identify word boundaries by modeling an incidental statistic, rather than by modeling word boundaries directly. In contrast, the Bayesian models offer a clear and principled probabilistic formulation of the word segmentation problem. To be more precise, the Bayesian joint-lexical models adopt an *ideal observer* approach, which describe the optimum solution. In other words, an ideal-observer model is not primarily focused on how infants actually solve a problem, but on best solution itself, given their capabilities.

It is worth dwelling on this point, since there are subtle differences between the ideal observer approach and what I will argue is cognitively plausible. The major rationale behind the ideal observer approach is that it cleanly separates the problem of defining an optimum solution from the processing strategies that infants might make use of to identify that solution. The utility of this distinction is apparent in the incisive comparisons it allows. The key example is Goldwater's (2006) bigram/independence comparison, which showed that *if* infants make the independence assumption, they are likely to undersegment – and in particular they are likely to segment collocations as whole lexical items. Even though this comparison does not tell us how

infants handle collocations, it clearly illustrates why collocations are a problem that need to be handled. Moreover, it formally explains why many existing models tend to exhibit undersegmentation, *and* it suggests one possible solution. These insights came from separating the optimum solution from the strategy that the infant should adopt to find that solution.

The ideal observer approach is also suited to certain contexts in which the engineering goal is to make maximal use of the data. For example, in natural language processing applications, labeled data is typically scarce and/or expensive to collect. Thus, unsupervised and semi-supervised methods are at a premium. One natural application would be developing lexicons for automatic speech recognition for languages in which electronic text resources are not available (Pierrehumbert, p.c.).

From this perspective, batch learning is somewhat less troubling than it otherwise would be. Batch-learning, in which the model processes all of its input in one go, is clearly not how infants acquire their language. But from an ideal-observer perspective, it doesn't matter how the model takes in its output, all that matters is obtaining the optimum solution. Thus, whether batch-learning is cognitively plausible or not is a question of perspective. The ideal-observer approach is not focused on what infants actually do, but on what can be learned about what they must do from the nature of the optimum solution. If one does not trouble oneself about how an infant reaches the optimum solution, it is perfectly cognitively plausible to employ a batch-learning method to obtain that solution. This dissertation, on the other hand, *does* concern itself with that infants actually do. This is why I listed the incremental property as a desideratum, and why it is not fully satisfying to simply accept claims to the effect that an equivalent incremental model

exists.

However, there is another troubling aspect of the ideal observer approach: arguably it is solving a different problem than people do. In these models, parsing decisions are always lexically mediated. That is, owing to the 1-1 relationship between word boundaries and words, placing word boundaries forces the model to learn or know a corresponding word. In contrast, infants are known to posit word boundaries on the basis of prelexical factors such as stress (Jusczyk et al, 1999b) and phonotactics (Mattys & Jusczyk, 2001), without recognizing the target beforehand, or instantaneously learning it.

Put another way, the word-learning facts pose serious questions about what is 'optimal'. Joint-lexical models have an effectively infinite memory and word-learning capacity. As a result, there is no need for such models to form generalizations about what cues and sequences are likely to signal a word boundary. Instead, the model can simply learn a new word on-the-fly to explain the word boundary; and if the initial guess turns out later to be probabilistically sub-optimal, the model can revise it later, unlearning the incorrect target and learning a new target or targets. In contrast, human listeners do seem to form generalizations about cues and sequences that indicate word boundaries, rather than learning every novel word they encounter (Storkel et al., 2006).

In summary, the Bayesian approach, as developed by Goldwater's (2006) thesis and subsequent publications, is a very promising avenue of research. Unlike both the connectionist and coherence-based approaches, it puts the problem on a firm probabilistic foundation, formalizing it as search for the maximum likelihood segmentation. This approach has yielded important insights already, such as the observation that collocations cause undersegmentation in

models which make improper independence assumptions. Moreover, the framework is both modular and extensible, so that interested researchers can easily modify the (publicly available) code to address their own research questions.

However, I have argued that existing Bayesian joint-lexical models are solving the wrong problem. Word boundaries can only be identified on the basis of word recognition/learning, which does not leave any room for phonological generalizations, such as the attested strategy of positing word boundaries before stressed syllables (Jusczyk et al, 1999b). Of course, it is logically possible to re-structure these models so as to incorporate phonological generalizations, and the Bayesian formulation is so elegant and computationally attractive that ultimately this may be the right theoretical road to take. But owing to this issue, I turn for the time being to the final class of models, which attempt to segment speech by drawing on phonotactic regularities.

Phonotactic Models

The common properties of phonotactic models are that they model boundaries explicitly using observable phonotactic statistics. Thus, these models are able to posit word boundaries in novel phonological material without adding it to their lexicon, crucially exhibiting phonological generalization. Moreover, all published phonotactic models have tested their model on non-English data. However, as we shall see they may be cognitively implausible in other ways.

The first phonotactic models is described in Xanthos (2004), who defines and integrates two phonotactic approaches which extend a basic *n*-phone model. The first builds on and formalizes Aslin et al's (1996) insight that utterance boundaries contain distributional

information that is informative for identifying utterance-medial word boundaries. Specifically, Xanthos defines the "boundary typicality" of a segment *n*-phone as the ratio of the probability of the *n*-phone utterance-initially (or -finally) to its context-free probability. Thus, *n*-phone which occur more often utterance-initially than in other positions will have an utterance-initial typicality greater than 1. The total boundary typicality of a sequence $w_1w_2...w_nw_{n+1}..w_{2n}$ is calculated by averaging the initial-typicality of $w_1w_2...w_n$ and the final-typicality of $w_{n+1}..w_{2n}$. A boundary is imposed whenever the utterance-typicality exceeds some threshold; Xanthos reports results for the "natural" threshold of 1. Xanthos' second phonotactic method formalizes the successor-count approach of Harris (1955). The (forward) successor count of an *n*-phone is defined as the number of segments which can follow the *n*-phone (i.e. have been observed to follow the *n*-phone in the model's previous input). Xanthos extends this to include the analogous predecessor (backward successor) count, and imposes boundaries at local maxima of the successor and predecessor counts. Xanthos reports results on a child-directed French corpus for *n* of up to 5, finding unsuprisingly that the combined model (either mechanism can posit boundaries) does better than either one alone.

The troubling aspect of this paper is its heuristicity; in other words, the lack of formal motivation for most of the design choices. To choose but two examples, the choice to impose word boundaries based on local minima of successor/predecessor counts is, as Xanthos himself admits, a fairly arbitrary property of his model; and, while the utterance typicality measure is clearly measuring something that is relevant to boundary detection, averaging the initial- and final- typicality scores to get a grand total is hardly a principled model of utterance occurrences.

Fleck (2008) addresses this latter issue, describing a model which explicitly estimates the probability of a boundary *p(b | l,r)* between its left *l* and right *r* contexts. This model again builds upon the insight of Aslin et al (1996) that utterance boundaries are informative for segmentation. Fleck formalizes this intuition by modeling utterance beginnings and ends as the initial-state left and right contexts. She further makes the following assumptions:

conditional independence given b $\quad p(r, l \mid b) = p(r \mid b) \cdot p(l \mid b)$

conditional independence given $\neg b$ $\quad p(r, l \mid \neg b) = p(r \mid \neg b) \cdot p(l \mid \neg b)$ (1.11)

The first assumption is similar to a unigram phone model of word types, except that the left and right contexts are not limited to single segments. The second is a phonological form of the independence assumption discussed by Goldwater (2006). Using these assumptions, Fleck derives a relatively straightforward estimate for *p(b | l,r)*. She then describes a learning algorithm which iteratively estimates the left- and right- contexts and their corresponding boundary probabilities. In addition, Fleck uses a morphological post-processing algorithm which distinguishes affixes from function words, and removes spurious boundaries around affixes. Like Xanthos (2004), Fleck uses *n*-phones with *n* up to 5, on which I will comment more below. Words in the corpus are then defined by their boundaries.

Like all previous models described above, Fleck runs her model on phonetic corpora generated by mapping textual corpora with a canonical phonetic transcription. However, Fleck goes beyond other work reported above in three ways. First, she runs the model on Spanish and

Arabic corpora (with canonical phonetic pronunciations). Second, she explicitly compares her model against Goldwater's (2006) model on the same datasets. Finally, she runs her model on the Buckeye corpus (Pitt, Dilley, Johnson, Kiesling, Raymond, Hume, & Fossler-Lussier, 2007) which includes natural conversational variation in pronunciation.

This has several advantages. Most immediately, Fleck finds that the phonotactic algorithm exhibits loosely comparable performance on all three languages, with the best performance on English, slightly worse performance on Arabic, and the worst performance on Spanish. Moreover, she finds that this relatively simple phonotactic model's performance is competitive with that of Goldwater (2006): the Goldwater model clearly exceeds Fleck's model in finding word boundary recall (presumably owing to the relaxation of the faulty word-independence assumption) but Fleck's model exceeds Goldwater's model in a variety of other measures and *on a variety of language data*, such as boundary precision and overall lexicon identification.

Finally, and perhaps most interestingly, Fleck finds that both models exhibit degraded performance on the Buckeye corpus, which contains natural phonetic variation such as simplified consonant clusters. However, the phonotactic model's performance is not degraded as severely as the Bayesian joint-optimization model. The likely reason for this is that Goldwater's model is designed to model a single pronunciation for each word; so the input crucially violates this assumption. In contrast, Fleck's model is primarily phonotactic, with the result that it can exploit phonological generalizations to posit word boundaries rather than relying purely on the assumption of an invariant realization of each word type. To the extent that word-boundary-relevant phonotactics are preserved under conversational reduction, it makes sense that a

phonotactic approach may fare better with conversationally reduced input than a Bayesian joint-lexical model. This fact reinforces the point initially raised in the previous section, that the joint-lexical models appear to be solving a different problem than infants do, owing to the lack of explicit generalizations with regard to segmentation cues.

These phonotactic models have made important contributions to our understanding of the acquisition of word segmentation. First and foremost, the work of Xanthos (2004) and Fleck (2008) suggests that the phonotactic approach to word segmentation, despite operating on language-specific representations, is a cross-linguistically robust strategy, giving loosely comparable results across different languages. Moreover, Fleck's (2008) data suggests that the phonotactic approach is very promising for coping with conversational reduction, presumably because many of the phonotactic cues that are most informative for signaling word boundaries are also least likely to be reduced.[13]

These phonotactic models, despite their advantages of being *implemented*, tested on *cross-linguistic* data, achieving *joint* optimization of word segmentation and lexicon discovery, and being somewhat *incremental*, nonetheless make somewhat cognitively implausible assumptions. In particular, Fleck's assumption of conditional independence of sounds within a word is strongly false, at least for unigrams; it implies, for example, that the word-initial sequence [st] is just as likely as the word-initial sequence [ts], which may be true in some languages but is

13 This conclusion must be regarded with some care, since most of the segments in the Buckeye corpus were identified by an automatic speech recognition engine. Although the entire corpus was hand-checked, it is clear that the automatic components of this process introduced certain biases. For example, over 90% of the tokens of the word *the* were realized with a high front vowel, which was the canonical pronunciation in the recognizer's lexicon; whereas in my own speech, *the* is typically realized with some reduction, i.e. not a canonical high vowel. If the recognizer similarly artificially preserved exactly those cues which are most useful for phonotactically-based segmentation, this result would be an artifact rather than a finding of genuine theoretical interest.

certainly false in general. This assumption is crucial to derive the probability of a boundary, and thus, while necessary for the model to function, is cognitively highly implausible.

Similarly, both Xanthos (2004) and Fleck (2008) models are built on an underlying database of 5-phone statistics, whereas there is no cognitive evidence that I am aware of that supports the claim that infants attend to *n*-phones for *n* higher than 2. In fact, Pierrehumbert (1994) showed that there is not sufficient data for learners to estimate correct statistics even for trigrams, except for the most frequent ones. Moreover, owing to the Zipfian distribution of linguistic events, the frequently occurring 4- and 5-grams are likely words themselves; and as already discussed at length above, there are only a few such words that infants actually know. In other words, the 5-gram model provides an implicit role for words in Fleck's model, which somewhat subverts the spirit of a phonotactic model.

To summarize, then, phonotactic models show considerable promise in explaining the prelexical segmentation abilities evinced by infants. The phonotactic approach has made important contributions, such as illustrating its robustness to cross-linguistic variation and natural conversational variation in pronunciation. However, existing models are either probabilistically unsound or make cognitively implausible assumptions, e.g. modeling infant memories with 5-phone models.

Summary

In summary, existing models of word segmentation can be divided into *connectionist*,

coherence-based, *Bayesian joint-lexical*, and *phonotactic* model. While there is some variation among published studies, there is substantial within-class consistency in whether models exhibit the desiderata of cognitive plausibility I discussed above: whether the model is *implemented*, tested on *cross-linguistic* data, provides for *lexical* acquisition (and if so, whether word learning is treated as a *joint* optimization problem or as a *related* but logically distinct problem), and finally, whether the model accepts *incremental* input, segmenting and learning as it goes. The preceding discussion of model classes and desiderata is summarized in Table 1.6, with additional comments in the final row:

	connectionist	coherence-based	Bayesian joint-lexical	phonotactic
implemented	✓	✓-	✓	✓
cross-linguistic				✓
lexical			✓ joint	✓related
incremental	✓		✓-	✓
other	*opaque hidden unit representations		*no phonlogical generalizations	**n*-phone

Table 1.6: Evaluation of word segmentation model properties

The general conclusion to be drawn from this review is that progress has been made on many fronts on the problem of word segmentation, but no model to date is fully cognitively plausible.

The two most promising avenues for research are the Bayesian joint-lexical and phonotactic models, since existing models satisfy most or all of the desiderata.

However, existing Bayesian joint-lexical models make the cognitively implausible assumption that every word is learned the first time it is encountered, an assumption which is currently built into the joint optimization strategy which is at the heart of these models. Phonotactic models, at least in principle, are not forced to this assumption by their architecture; however, existing models make other cognitively implausible assumptions, such as the assumption that children attend to and store 5-grams.

Thus, my goal in this dissertation is to develop a more cognitively plausible phonotactic model of word segmentation, and to develop a cognitively plausible account of the relationship between word segmentation and word learning.

Two-stage framework

Current theories of word recognition (Vitevitch & Luce, 1998; Pierrehumbert, 2003) posit two distinct levels of representation, a *sublexical* and a *lexical* level, with distinct attendant processes. Theories differ as to the precise nature of the representations in both levels, but are in general agreement that the lexical level involves representation of wordforms, whereas the sublexical level involves representations that compose wordforms. Theories accordingly differ as to the nature of sublexical processes, but are in general agreement that during *lexical access*, stored wordforms compete to explain the input.

One of the crucial pieces for evidence for this distinction is distinct and sometimes

opposing effects of sublexical probability and lexical neighborhood density. Sublexical probability (referred to as phonotactic probability in the psycholinguistics literature, but called *sublexical* here to distinguish it from *lexical* phonotactics) of a wordform refers to the probability of its sub-parts, often operationalized as the sum of position-specific and diphone probabilities (Vitevitch & Luce, 2004). Neighborhood density refers to the number of similar-sounding words, often operationalized by the number of words differing from the target by the insertion, removal, or mutation of 1 segment. For example, Luce & Pisoni (1998) found an inhibitory effect of neighborhood density on word recognition (lexical decision), consistent with the prediction that recognition is slower when there are more competitors. In contrast, Vitevitch, Luce, Charles-Luce, & Kemmerer (1997) found a facilitory effect of sublexical probability on the same task. Evidence for this distinction has been found in perception, production, recall, and learning (Frisch, Large, & Pisoni, 2000; Luce & Large, 2001; Luce & Pisoni, 1998; Storkel et al, 2006; Thorn & Frankish, 2005; Vitevitch, 1997; Vitevitch, 2002a; Vitevitch, 2002b; Vitevitch, Armbruster, & Chu, 2004; Vitevitch & Luce, 1998, 1999).

I submit that the developmental facts reviewed earlier in this chapter are another argument for this distinction. Recall that the developmental literature shows that by 10.5 months of age, typical infants know 10-40 words, but use an array of metrical and phonotactic cues to segment novel nonwords from unfamiliar contexts. These facts are inconsistent with the hypothesis that word recognition is the primary locus of infants' word segmentation: infants are clearly able to segment speech without recognizing all or most of the words they segment. These results can be explained by the assumption that the primary locus of infant word segmentation is sublexical.

Accordingly, I propose that segmentation is in fact the primary process associated with the sublexical level of representation. This proposal is essentially Pierrehumbert's (2001) “Fast Phonological Preprocessor (FPP)”, which “uses language- specific, but still general, prosodic and phonotactic patterns to chunk the speech stream on its way up to the lexical network. By integrating such information, the FPP imputes possible word boundaries to particular temporal locations in the speech signal.” The architecture I assume is given in Fig. 1.4 with an example below:

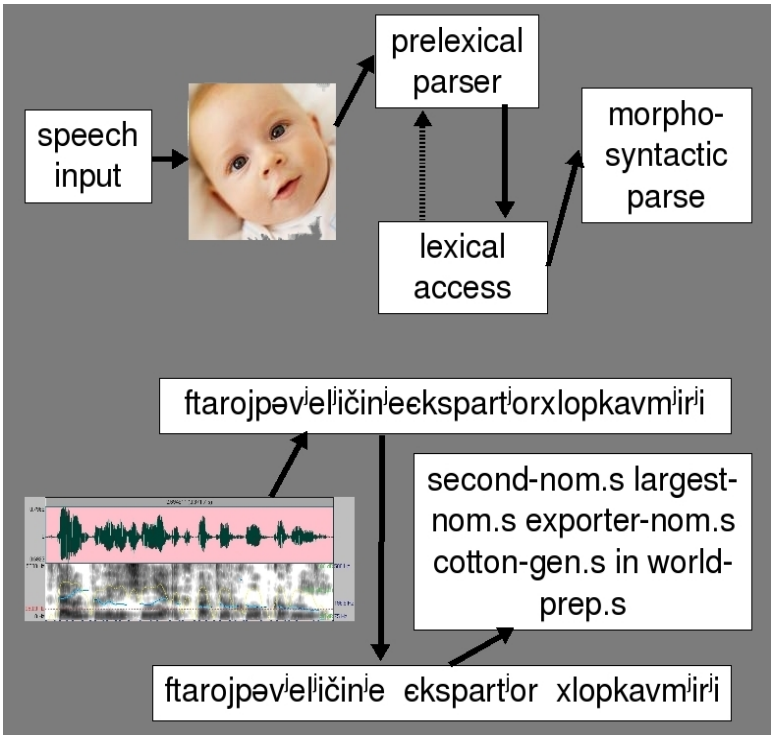


Fig. 1.4: Two-stage speech processing framework

In Fig. 1.4, the bottom half is an example of the general architecture outlined in the top half. An example of Russian input is given, first without the 'word boundaries' (spaces) to represent the

input to the prelexical parser, then with 'word boundaries' to represent the output of the prelexical parser, then with a morphosyntactic parse to represent the output of the lexical access mechanism.

I further assume a continuity theory of development, in which these processes are operative throughout the lifespan of a listener. Of course, as already pointed out, an infant without a sizable vocabulary is at a considerable disadvantage. In particular, the absence of a sizable vocabulary implies a greater dependency on sublexical processing than adults might exhibit. Or in other words, the segmentation abilities that infants exhibit are primarily due to sublexical segmentation.

Existing models of word segmentation have tended to focus on the problem itself, and the related problem of word learning, without setting out clear and explicit assumptions about the cognitive architecture that supports segmentation. As a result, previous research has not made a point of focusing on the cognitive implications that stem from the segmentation algorithm. In contrast, the two-stage architecture I adopt here allows for clear and relatively precise predictions about the nature of processing that must occur at each level. This can be seen, I will argue, by considering the range of possible error patterns that the sublexical segmentation mechanism makes.

Error patterns

The first observation that can be made is that adults in general understand one another, implying that segmentation errors are quite infrequent. Thus, it must be a property of the system

as a whole that any system-internal errors are somehow filtered. More precisely, any incorrect decisions introduced by prelexical segmentation must be corrected during lexical access. Therefore, the error pattern exhibited by the prelexical segmentation process defines the problem of lexical access. There are four logically possible error patterns: no errors, undersegmentation, oversegmentation, and over+undersegmentation. I will assume the first case, of no errors, is simply too much to hope for, and will not consider it further.

Undersegmentation is the pattern of error in which the segmentation mechanism conservatively identifies word boundaries. In this case, the mechanism does not find all word boundaries, but when it identifies a word boundary, it is rarely wrong. The cognitive implication from this pattern of errors is that the lexical access mechanism can generally rely on word boundaries discovered by the segmentation mechanism, but must discover some additional ones. In this case, the primary contribution of the lexical access mechanism to word segmentation is to further segment speech, presumably by matching stored lexical representations at onsets of the partially segmented signal, and positing additional boundaries for unmatched substrings.

In contrast, *oversegmentation* is the opposite pattern of errors, in which the segmentation mechanism aggressively identifies word boundaries. In this case, the algorithm finds virtually all of the word boundaries, but also falsely identifies many non-boundaries as boundaries. The cognitive implication from this pattern of errors is the lexical access mechanism can generally rely on *non*-boundaries discovered by the segmentation, but must filter the incorrect word boundaries. Then the primary contribution of the lexical access mechanism to word segmentation is to eliminate spuriously posited word boundaries, presumably by simple virtue of lexical

searches failing when they are initiated at false boundaries.

The final possible error pattern is *over+undersegmentation*, in which the segmentation mechanism both fails to identify a significant proportion of underlying boundaries, and also falsely identifies a significant proportion of non-boundaries as boundaries. The cognitive implication is that the lexical access mechanism cannot fully rely on any decision made the segmentation mechanism, and must be prepared to filter errors of both types. This implies a very limited role for the segmentation mechanism, at best mildly improving the speed and accuracy of the lexicon, and a correspondingly greater role for the lexical access mechanism.

This last pattern is unlikely to be the correct one, for two reasons. First, from the standpoint of computational design, it is inefficient to have a distinct processing level that does not make a meaningful and independent contribution to speech processing. This is not an absolute argument, as it is not inconceivable that human speech processing is inefficient in this particular way, but a growing body of literature suggests that the human speech processing system is in general exquisitely adapted to solve the problems it is faced with near-optimum efficiency (see Jurafsky, 2003, for a review).

Second, in this framework it must be the sublexical segmentation mechanism which explains most of infants' segmentation abilities. If infants truly exhibited *over+undersegmentation* in the course of word segmentation, it should be more apparent in the developmental literature. While there are examples in which infants fail to segment, the overwhelming majority of studies suggest that infants are quite good at word segmentation by 10.5 months. In fact, I am only aware of two negative results on word segmentation on natural

stimuli from the infant's native language. The first is the 7.5-month-olds in Jusczyk et al (1999b), who posit word boundaries at the onset of stress syllables even for iambic words like *guitar* – and this kind of error is corrected by 10.5 months as shown by the same study. The other negative results is from Mattys & Jusczyk (2001), who showed that infants failed to segment a novel word when it was embedded in a context which failed to support parsing it out as a separate word – arguably the correct behavior rather than evidence of improper segmentation. In other words, the impressive segmentation abilities of infants are not consistent with any segmentation mechanism that predicts a substantial proportion of both over- and under-segmentation errors.

Research questions

I have argued that existing models of word segmentation suffer from one or more cognitively implausible assumptions. As I see it, the most promising class of models are phonotactic, in part because the Bayesian joint-lexical models predict that infants will command a sizable vocabulary as they begin to exhibit word segmentation. Therefore in this dissertation I propose to develop a novel phonotactic model of word segmentation that I will refer to as *Diphone-Based Segmentation* (DiBS), based on the finding of Mattys & Jusczyk (2001) that infants attend to the relative frequency with which diphones span word boundaries versus occurring word-internally. Then the goal of this dissertation is test this proposal as fully as possible, and to generate a full developmental account of word segmentation and its relation to word learning.

The general research strategy I will employ is to build a computational model which *only*

makes use of the diphone cue. This model will be used to answer the following questions:

1) Proof of concept: Does diphone-based segmentation actually yield good segmentation?
2) Input robustness: How sensitive is the diphone-based segmentation to input assumptions?
3) Cross-linguistic robustness: Is diphone-based segmentation similar for different languages?
4) Learnability: How can infants estimate the appropriate diphone statistics?
5) Toward word-learning: How can diphone-based segmentation facilitate word-learning?

Contributions

This dissertation makes a number of theoretical and empirical contributions to our understanding of the acquisition of word segmentation. Perhaps most importantly, it develops a full learnability account for effective prelexical segmentation using phonotactic properties that are available on the surface, i.e. utterance-boundary distributions.

In addition, this dissertation is to my knowledge the first research on this topic which focuses on the implications of the segmentation error pattern for bootstrapping lexical acquisition in the context of an incremental model. In this vein, another important contribution is an explicit and fair comparison across a range of coherence-based approaches, and the resulting finding that without additional model structure, coherence-based approaches are not adequate to explain the acquisition facts. Finally, this dissertation extends support for the claim that the phonotactic approach to word segmentation is robust to language variation, by implementing a novel

phonotactic model and testing it on English and Russian language data, as well as English data with conversational reduction.

Beyond these theoretical contributions, this dissertation involved the creation of a large scale (~35 million word) Russian phonetic corpus from a text corpus, and software for converting Russian orthography to a phonological transcription and thence to a phonetic transcription. (The software for generating a phonological/phonetic transcript can be obtained by contacting me.)

Structure of the Dissertation

The remainder of the dissertation is structure as follows.

Chapter 2 ("English") begins by defining a baseline DiBS model which segments speech using the statistically optimal diphone statistics (supervised learning). Then, in Corpus Experiment I, the baseline model is trained and tested on a phonetic transcription of the British National Corpus. Corpus Experiment II examines the effects of abstractness of the input representation and conversational reduction processes by testing the baseline model on two different transcriptions of the Buckeye corpus (Pitt et al, 207), one with conversational reduction, and one with canonical transcriptions.

Chapter 3 ("Russian") replicates Corpus Experiment I with Russian data. It begins by describing relevant aspects of the Russian language. Next, it describes how a phonetic transcription of the Russian National Corpus (RNC) was generated. Finally, Corpus Experiment III applies the baseline DiBS model to the RNC.

Chapter 4 ("Learnability") addresses the question of how the optimal diphone statistics

might be learned in an unsupervised manner. Corpus Experiment IV begins by implementing a variety of coherence-based approaches, such as the forward transitional probability proposal of Saffran et al (1996); it is shown that these proposals achieve poorer segmentation than DiBS for every threshold. Then, a bootstrapping theory is developed by which the diphone statistics can be estimated either from a small lexicon such as an infant might possess ("Early Learner DiBS") or from raw utterance-boundary distributions, without any lexicon at all ("Prelexical DiBS"). Corpus Experiments V and VI test these bootstrapping models on the English and Russian corpora developed in previous chapters.

Chapter 5 ("Toward Word Learning") addresses the question of how infants might learn words from the output of the prelexical segmentation mechanism developed in the preceding chapter. It is argued that lexical access is the locus of word learning, and so a theory of lexical access is developed. Corpus Experiment VII tests the adult (best-case) scenario in which the baseline DiBS segmenter is combined with the proposed lexical access mechanism operating with a full lexicon; as well as related cases using the prelexical parser and/or no lexicon. A theory of word-learning is proposed, whereby learners add wordforms that they access sufficiently frequently and with sufficiently high confidence in the segmentation. Corpus Experiment VIII test the combined bootstrapping model, in which word segmentation and word learning are bootstrapped together. Finding that many spurious single-consonant words are learned, Corpus Experiment IX re-tests the bootstrapping model with a single word-learning constraint: a novel word must contain a vowel.

Chapter 6 ("Conclusions") highlights the concrete progress this dissertation makes toward

our understanding of the acquisition of word segmentation. It also discusses various issues of potential interest for follow-up.

CHAPTER 2: ENGLISH

Abstract

This chapter begins by giving formal definitions of the word segmentation problem and the baseline model referred to throughout the remainder of this dissertation. Next, it defines the elements of signal detection theory that are used in this dissertation to analyze results. Then, in Corpus Experiment I, the baseline model is applied to the British National Corpus, which is mapped to a phonetic representation using the CELEX pronouncing database. Since this method does not include pronunciation variation such as occurs in natural conversational processes, Corpus Experiment II applies the baseline model to two different versions of the Buckeye corpus, one in which every word is realized with a canonical pronunciation (as in Corpus Experiment I), and a phonetic transcription of the same corpus that includes conversational reduction processes. A common pattern of undersegmentation is found, and the cognitive implications for acquisition are discussed.

This chapter defines a baseline implementation of DiBS and tests it in two Corpus Experiments. The baseline model (hereafter referred to as baseline-DiBS) is a *supervised* model, in the sense that it is given access to word boundary locations during training. There are two motivation for beginning with a supervised model. First, its performance is an upper bound for unsupervised models. Thus, if baseline-DiBS does not achieve a promising level of segmentation, no unsupervised model will do better, and it could be concluded from these facts alone that DiBS is not a tenable model of infant segmentation. Second, when learnability is

considered in earnest in later chapters, it will prove useful to have a standard of comparison, and the baseline model results herein will serve admirably.

Recall that the core idea of a DiBS model is to posit word boundaries based on their probability given the surrounding context, i.e. *p(# | xy)*. In the baseline model, this value is simply calculated from the relative frequency in the training corpus:

$$\text{baseline-DiBS:} \qquad p(\# \mid xy) = f(\#, xy) / f(xy) \tag{2.1}$$

In terms of the desiderata identified in the previous chapter, baseline-DiBS is *implemented* and *incremental*. In future chapters, I will apply baseline-DiBS to *cross-linguistic* (Russian) and develop a theory relating it to *lexical* acquisition.

Baseline-DiBS as defined here has in principle been implemented already in Cairns et al (1997). However, the phonetic corpus in that study used a different transcription system; in addition, that study used a corpus of about 10,000 words, comparatively small by contemporary standards. Corpus Experiment I replicates and extends the Cairns et al (1997) results by implementing the same model on the corpus that will be used throughout this dissertation, the 100 million word British National Corpus. Before the experiment, I give a formal definition of the segmentation problem.

Formal definition of segmentation problem

Formally a phrase $\rho = (\omega, \phi, \pi, R)$ consists of a sequence of a sequence of words ω from

a lexicon Ω, their realization as a sequence of phone ϕ from an alphabet Φ , a partition[14] π of ϕ into wordforms, and the realization R which relates each words to their corresponding forms $R[\omega_i] = \phi_{\pi(i)} ... \phi_{\pi(i+1)}$, where $\pi(i)$ is the location of the i^{th} boundary in the partition. The notation $R[\cdot]$ is used to indicate that word realization is a random variable, i.e. without assuming that words are realized invariantly as some canonical sequence of phones.

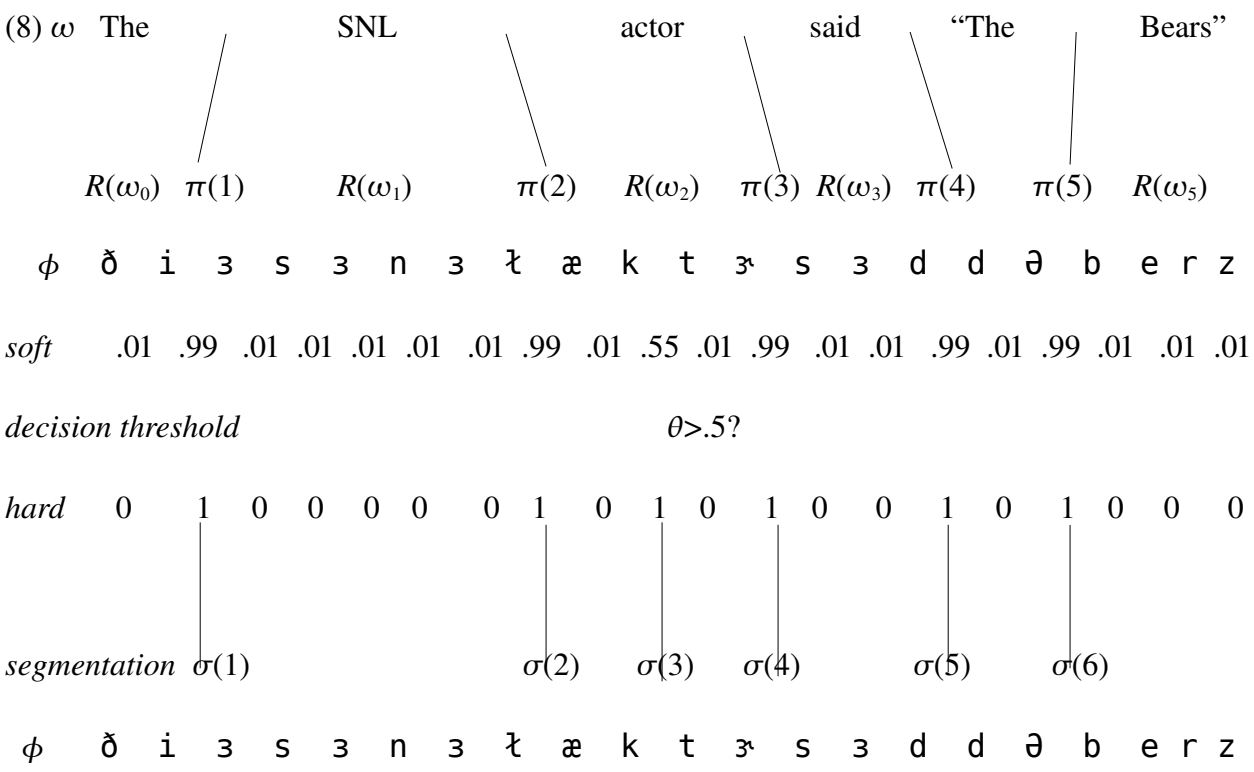
A *segmentation* σ of a phone sequence ϕ is a partition. Note that there is a 1-1 relationship between segmentations and wordforms, but not between segmentations and words themselves. This is for two reasons. First, a word may have multiple realizations as distinct wordforms. For example, the word *the* may be realized with an interdental fricative onset, or the fricative may be simplified to a stop. Second, the same wordform could be a realization of multiple different words. For example, *dear* and *deer* could be realized as the same wordform even though they are distinct words. A segmentation σ of the phone sequence of a phrase $\rho = (\omega, \phi, \pi, R)$ is a *true parse* of ϕ if and only if $\sigma = \pi$. (This is the formal device which distinguishes the problem of word segmentation from the problem of assigning wordforms to words.)

A *hard parser* is a function $f: \Phi^* \rightarrow 2^{\Phi^*}$ which assigns a segmentation to a phone sequence ϕ, and a *hard parse* is the output of a hard parser. A *parse distribution* is a function p: $2^{\Phi^*} \rightarrow I$ which assigns probabilities[15] to hard parses of a phone sequence ϕ. A *soft parser* is a function p: $\Phi^* \rightarrow \mathbb{R}^{\Phi^*}$ which assigns to each possible boundary in a phone sequence a statistic which corresponds monotonically to the likelihood of a word boundary, and a *soft parse* is the resulting output sequence. A *decision procedure* is a function $\Theta : \mathbb{R}^{\Phi^*} \rightarrow 2^{\Phi^*}$ which maps soft

14 A partition of a sequence is an exhaustive decomposition into components, e.g. *government* → *govern* + *ment*.
15 The notation I is used to refer to the unit interval [0,1], i.e. the domain of probabilities.

parses to hard parses; in general the decision procedure will simply map the statistic to a word boundary if it is above/below some threshold θ (and otherwise to the absence of a word boundary), in which case it will be called a *decision threshold*.

These formal definitions are illustrated in (8), which gives the true parse on top and incorrectly oversegments *actor* beneath:



Note that the partitions do not include the utterance boundaries. It is a matter of formal bookkeeping whether these are counted as part of the partition or not; but in practice, utterance-initial and -final word boundaries will not be scored, so they are omitted here. Note further that the same word *the* has two distinct realizations, as [ðiˌ] and as [də]. In practice, this will not

occur in most of the corpora used in this dissertation. However, since it can occur it is important to make room for it in this framework.

Informally, the segmentation problem is to recover the true parse, or failing that, a parse that is as "close" to the parse as possible. The appropriate way to characterize this will in theory depend on the kind of output the model produces. For example, a model which returns a full parse distribution could be characterized as close to the true parse if it assigns most of its probability mass to segmentations which share all or nearly the same boundaries as the true parse. In practice, however, the simplest case to evaluate is a hard parse. Even for models which assign a richer output than a hard parse (such as Fleck, 2008), it is convenient to map the output of the model to a hard parse. Thus, I define the segmentation problem as defining an algorithm f which accepts an input corpus C consisting of a sequence of phrases and returns a hard parse f(C) of that corpus.

This process puts different models on a level playing field, by providing for well-defined comparisons between model outputs even when the model assigns richer structure to the corpus than is represented in a hard parse. The tool I and most other researchers use to compare hard parses is *signal detection theory*, described in the next section.

Signal detection theory

Elements

Signal detection theory (Green and Swets, 1966) can be used to evaluate signal detectors under conditions of a binary signal in noise with repeated sampling. The fundamental idea is to

divide the world into four cases: whether the signal occurred (present/absent), and whether the detector reported a signal (detect/not). Signal detection theory is appropriate in cases, such as word segmentation, where the signal is relatively unlikely compared to its absence, because percent correct can be misleading in such cases (cf. Appendix 2A).

Of course, the ideal detector would report a signal whenever one occurs, and report no signal whenever the signal does not occur. But since noise is an inherent aspect of the detection process, there is some chance that the detector will be wrong. It can be wrong in two ways: by failing to report a signal when it occurred, and by wrongly reporting the signal when it did not occur. Similarly, the detector can be right in two ways: by reporting a signal when it occurs, and by failing to report a signal when it did not occur. These are the fundamental events of signal detection theory:

- *hit* signal present, detector reports it
- *miss* signal present, detector doesn't report it
- *false alarm* signal not present, detector reports it
- *correct rejection* signal not present, detector doesn't report it

One way to compare two detectors is to compare the number of hits, etc.. on the same sample.

However, it is often desirable to compare detectors independent of the precise sample size they were tested on. Thus, rather than comparing the raw counts above, detectors can be measured using the following rates:

- *hit rate* p(detect | signal) = hits / (hits + misses)
- *false alarm rate* p(detect | ~signal) = false alarms / (correct rejects + false alarms)
- *precision* p(signal | detect) = hits / (hits + false alarms)
- *accuracy* p(correct decision) = (hits + correct rejects) / all decisions

The hit rate, also called recall or true positive rate, indicates the probability of detecting a signal when one has occurred. Precision indicates the probability that a signal has actually occurred, given that it was detected. Although similar-sounding, these numbers reflect very different aspects of a detector's performance, as illustrated in the example in Appendix 2A. The false alarm rate is the probability of incorrectly detecting a signal when the signal is absent. Accuracy is the overall rate of correct decisions.

Receiver Operating Characteristic (ROC)

In general, detectors report the presence of a stimulus whenever some measurement exceeds (or falls below) a *decision threshold*. For example, a smoke detector might consist of a device that measures air clarity and a clarity threshold θ. Whenever air clarity drops below θ, the smoke detector starts making noise. It is useful to think of the threshold in terms of the sensitivity of the detector: when the detector is too sensitive, it will start going off every time the stove is turned on, but will at least reliably go off when there is a fire. Similarly, if the detector is not sensitive enough, it will never false alarm, but it may not go off even when there really is a

fire. The response of the detector across a range of decision thresholds is standardly summarized using a graph known as the Receiver Operating Characteristic, which plots the hit rate against the false alarm rate (Green & Swets, 1966). An example ROC curve is shown in Figure 2.1 for illustration:

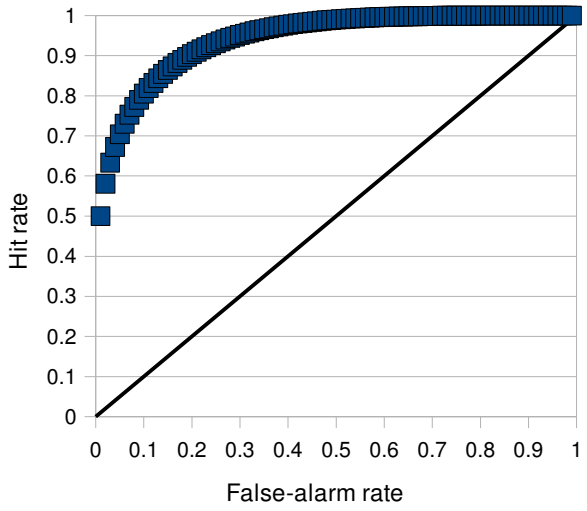


Fig. 2.1: Example Receiver Operating Characteristic (ROC) curve

In this example curve, the ROC is represented by the points on the curve, and the diagonal line represents a measure of *chance* performance. This is the rate of hits and false alarms that are expected by simply detecting the signal randomly with some probability, i.e. independently of whether the signal is there.

Informally, a detector is “bad” if the ROC curve stays close to the diagonal line, and it is “good” if it stays far away from the diagonal. The ideal detector would contact the upper left hand corner, i.e. achieving a perfect hit rate without every making any false alarms.

Threshold selection

Choosing a threshold is often a value-laden choice. In particular, it depends on the relative frequency and cost of each error type. For example, if misses are far more costly than false alarms, it makes sense to make the detector fairly sensitive, even if this results in a higher rate of false alarms. In some cases, it is possible to assess the cost of each error type in common units (e.g. monetary units), thereby defining an objective function with a unique optimum. However, it is often the case that the costs of each error type are incommensurable (see Appendix 2A for an example), or in the present case, depend in an as-yet-unknown way on the larger system in which the detector functions. In these cases, there is no well-defined criteria by which to distinguish one threshold as "optimal".

In the absence of such a clear prior criteria, the most principled approach is to select the decision threshold which minimizes the total number of errors. Equivalently this is the threshold which maximizes the likelihood of making a correct decision, and is therefore known as the *maximum likelihood decision threshold* (MLDT). In general, the MLDT will depend upon the detector. However, the MLDT is predictable for the class of *probabilistic detectors*, which are designed to report the probability of the signal they are designed to detect. In this case, the expected MLDT is θ=.5. That is, the MLDT is the threshold at which the detector reports the signal whenever the signal is more likely than its absence.

To anticipate briefly, this point will become important in Chapter 4, when I implement a variety of coherence-based models. The DiBS models developed here are all probabilistic

detectors in the sense above, and thus always have a predictable MLDT at θ=.5. In contrast, there is no way to determine the MLDT of coherence-based models in advance, so threshold selection must be regarded as a free parameter in evaluating such a model. Thus, there is a principled way to select the decision threshold in DiBS, but not in coherence-based models.

Evaluating parses

A segmentation algorithm f can be evaluated on a corpus C by treating the presence of a phrase-medial word boundary in C as the signal, and the presence of a corresponding boundary in f(C) as reporting the signal. For most of the models developed in this dissertation, the primary form of evaluation will be an ROC curve, together with qualitative analysis. These curves summarize the ability of the model to find word boundaries for a wide range of decision thresholds.

In some cases it will prove useful to compare models at a particular decision threshold. For example, the Bayesian joint-lexical models output hard parses rather than soft parses, so their performance cannot be summarized by an ROC curve. In such a case, the performance is typically summarized by reporting the *boundary recall* and *boundary precision*. In these cases, more detailed analysis is generally possible.

In particular, since the segmentation of a corpus uniquely determines the wordform tokens it contains, it is also possible to determine the *wordform recall* and *wordform precision*. This is done by treating the whole wordform in C as the signal, and interpreting the model as detecting the signal if f(C) contains word boundaries on both edges of the corresponding

wordform token, i.e. if the wordform token is correctly segmented.

Since these wordform measures are based on tokens, it is further possible to determine the *lexicon recall* and *lexicon precision*. This is done in existing models by treating the lexicon as the set of wordforms that occur in the true parse of C, and the estimated lexicon as the set of wordforms that occur in f(C). Then a type in the lexicon is treated as the signal, and the model is interpreted as detecting the signal if the estimated lexicon contains the same wordform. (Note that in existing models, the calculation of lexical recall and precision assumes a 1-1 relationship between wordforms tokens and lexical types.)

As discussed in Chapter 1, I consider word learning to be a related but separate problem from word segmentation. Thus, I do not report wordform or lexicon recall/precision in Chapters 2-4, where I consider the word segmentation problem specifically.

Formal definition of diphone-based segmentation

Recall that a segmentation algorithm f accepts an input corpus C of phrases $\rho = (\omega, \phi, \pi, R)$ and to each phone sequence ϕ assigns a segmentation σ. The algorithm f is *diphone-based* if and only if the presence/absence of a word boundary between the phones $\phi_{k-1}\phi_k$ depends only on ϕ_{k-1} and ϕ_k.

In practice, the segmentation algorithms developed in this dissertation will actually assign soft parses, which are then mapped to a hard parse using a decision threshold. In principle, the probabilistic information in the soft parse might be of considerable use to the downstream lexical access mechanism. In particular, hard decisions will lead to hard errors, whereas probabilistic

information might prevent unrecoverable errors. However, as remarked above, it is considerably simpler to evaluate whether a decision is correct or not than to evaluate a probability distribution over outputs. Thus, while it is consistent with the spirit of DiBS to pass a soft parse or other richer structure downstream, this implementation of DiBS outputs hard parses for the sake of easier evaluation.

The models defined in this dissertation are probabilistic detectors in the sense defined in the previous section. That is, they attempt to calculate the probability of a word boundary between two phones, given the phone identity. I will use the following notation to indicate this probability:

p(# | xy) probability of a word boundary in the middle of the sequence [*xy*] (2.2)

Thus, diphone-based segmentation refers to a segmentation algorithm which posits word boundaries in a phone sequence by modeling the probability of a word boundary between every pair of successive phones.

Baseline model

The baseline model is simply the statistically optimum diphone-based segmentation model; that is, the model which is equipped with the true underlying probability *p* of a word boundary between every diphone *xy* that occurs in the corpus. For a given corpus C, this probability is determined by the relative with which a word boundary occurs between x and y:

$$\text{baseline:} \qquad p_C(\# \mid xy) = f_C(\#, xy) / f_C(xy) \tag{2.3}$$

where *f(#, xy)* indicates the frequency with which an utterance-medial word boundary occurs between [*x*] and [*y*], and *f(xy)* indicates the total frequency with which the diphone [*xy*] occurs.

Note that the baseline model requires *supervised* learning, because to calculate the diphone statistics according to Eq. 2.1, the model must have access to the location of utterance-medial word boundaries, which is exactly what the infant is trying to estimate. Thus, as discussed in Chapter 1, the baseline is not an appropriate model for infant acquisition. Rather, it is an upper bound that describes the maximum level of performance that could be obtained by an *unsupervised* method. The next section describes Corpus Experiment I, which establishes this upper bound on the British National Corpus.

Corpus Experiment I: Baseline-DiBS on the BNC

The goal of Corpus Experiment I is to establish baseline performance for diphone-based segmentation, both to serve as a proof of concept for the diphone-based approach, and as a standard for future models. In the following subsections I describe the corpus (and rationale) and the baseline model's performance on it.

Corpus

Corpus Experiment I is conducted by running the baseline model on a phonetic

transcription derived from the British National Corpus (BNC, 2007)[16]. A brief description of the BNC from its website (http://www.natcorp.ox.ac.uk/corpus/index.xml) is given below:

> The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written...
>
> The written part of the BNC (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part (10%) consists of orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different age, region and social classes in a demographically balanced way) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins.

The paramount concerns in selecting the corpus for this experiment are that it be *comparable* to other cross-linguistic corpora, sufficiently *large* as to avoid data sparsity issues, and *representative* of speech.

16 Data cited herein has been extracted from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

The BNC admirably fulfills the first two of these three criteria. In terms of size, the BNC contains approximately 92,339,941 million words, about 10,000 times as large as the Bernstein-Ratner corpus of child-directed speech used in most of the other segmentation models reviewed in Chapter 1 (e.g. Brent & Cartwright, 1996). More data is better for a variety of reasons. Of special importance here is data sparsity: the well-known Zipfian distribution of language implies that most of the linguistic events of interest that occur are rare. It is perhaps underappreciated that data sparsity is a serious issue even in phonology, even for diphone models. For example, the Buckeye corpus (Pitt et al, 2007) currently comprises about 300,000 word tokens (30 times larger than the Bernstein-Ratner). There are approximately 2,700 diphone types in the Buckeye; of these, half occur less than 15 times, and 400 occur only once.

This is not just an implementation issue, but a real cognitive issue. No matter how large the input sample, data will be sparse, and the amount and kind of input data determines the scale of the data sparsity issue. Accordingly, it is important to select a corpus which is large enough to model the data sparsity problem faced by infants. Back-of-the-envelope calculations, summarized in Appendix 2B, suggest that (English-learning) infants hear somewhere between 5 and 10 million words in their first year. Thus, the BNC is of a sufficient size to provide for several years of input. In contrast, the Bernstein-Ratner corpus represents about a morning of input. Since the goal of Corpus Experiment I is to provide the best-case test of diphone-based segmentation, it is important to use as large a corpus as possible so as to minimize the sampling errors that arise from data sparsity.

In terms of comparability, a major goal of this dissertation is to test the phonotactic

models developed in on cross-linguistic data. The Russian National Corpus (RNC) was explicitly modeled after the BNC, meaning that these corpora are as similar as two corpora from different languages and cultures could reasonably be expected to be. For example, the range of genres represented is roughly equivalent. Nonetheless, there are cultural conventions which lead to real differences in these corpora as well. For example, Russian commas are completely syntactically determined in the modern language, e.g. obligatory before embedded clauses. English commas, though clearly sensitive to syntactic structure, are not completely deterministic in the same way, and at least in my own writing, appear to be directly sensitive to prosodic structure. Commas in particular are a substantive difference because I assume they signal phrase boundaries; in fact, comma placement appears to be sensitive to cultural/historical/stylistic factors (Pierrehumbert, p.c.; Truss, 2003). This kind of cultural variation in corpus properties, though somewhat regrettable, is unavoidable when comparing across languages. In short, although there are some differences, the BNC and RNC are quite comparable.

Unfortunately, the BNC is not especially representative of speech, especially of the speech that infants are exposed to. This is so for two reasons. First, the BNC is largely composed of written sources, which presumably contains a richer vocabulary and wider variety of syntactic constructions than everyday conversational speech. Second, and probably more significantly, the phonetic transcription method used here projects each orthographic word to a single, canonical phonetic realization. In other words, every word is pronounced the same way every time in the phonetic transcript, whereas conversational speech contains a variety of pronunciation variation, owing to various reduction and assimilatory processes (Johnson, 2004). In these two ways, the

input corpus used here differs substantially from what infants are actually exposed to.

Since the primary goal of Corpus Experiment I is not to model acquisition *per se*, but to serve as a proof of concept and standard for comparison, I deemed it more important to meet the size and comparability criterion than the representative criterion. (It was not possible to meet all three here as there is no large, freely available corpus which includes conversational reduction processes.)

Phonetic form

A phonetic transcription of the BNC was generated by mapping orthographic forms to the most frequent phonological form listed in the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995). For simplicity, stress was not represented directly, although it was represented indirectly through its segmental reflexes (e.g. presence/absence of vowel reduction).

Word-external punctuation (commas, periods, semicolons, etc...) was mapped to phrase boundaries. Word-internal punctuation was not treated as word boundaries, e.g. compounds such as *topsy-turvy* were realized as a single word.

In developing the mapping software, I found that many out-of-vocabulary (OoV) word forms were inflected variants of in-vocabulary words. For example, *George* was listed in CELEX, but *George's* was not. Thus, I added minimal inflectional processing capabilities. Specifically if the mapper found an OoV word and detected the past or plural/possessive morphemes, it attempted to recover the stem; the word was then mapped as the pronunciation of the stem plus the appropriate phonetic realization of the past/plural/possessive morpheme. This

method eliminated 95% of out-of-vocabulary items, reducing OoV tokens to less than 1% of the total corpus. The remaining OoV tokens were discarded.

Some words are in order about the CELEX phonetic representations. First, I used the built-in DISC transcription system because it enforces a 1-phone-1-grapheme transcription convention; e.g. distinguishing the diphthong [a͡u] from the vowel-vowel sequence [au], and the voiceless alveolar affricate [č] from the stop-fricative sequence [tš].

Second, the CELEX team did not represent contextual variation consistently across segments. For example, contextual variation in the phoneme /r/ is represented with 3 different allophones: word-finally as [R], deleted in non-final singleton codas, and as [r] elsewhere, consistent with the British RP (Received Pronunciation) dialect standard, and illustrated below:

(9)	[r]	wreathe	riD
		corrosion	k@r5ZH
		growths	gr5Ts
	[R]	star	st#R
	null	starchy	st#JI

Similarly, /n/ and /l/ both have distinct allophones for when they occur as syllabic nuclei. However, contextual variation between light and dark /l/ (Hayes, 2000) was not represented allophonically. Similar contextual variation in the phoneme /t/, e.g. the systematic alternation between an aspirated [t] in a singleton onset and unaspirated [t] in an [st] cluster was not

represented. In summary, allophonic variation in CELEX sometimes contains positional information, but not according to a consistent scheme, and not consistently across all segments.

Evaluation

The baseline model was trained and tested on the phonetic transcription of the BNC described above. The baseline model assigns soft parses, which were mapped to hard parses using a decision threshold θ, which was varied between 0 and 1. As discussed above, the baseline model is a probabilistic detector, and therefore has an a priori maximum likelihood decision threshold (MLDT) at θ=.5.

Results

Fig. 2.2 (below) shows the ROC for the baseline model as tested on the phonetic transcription of the BNC. The MLDT is highlighted graphically with a red circle:

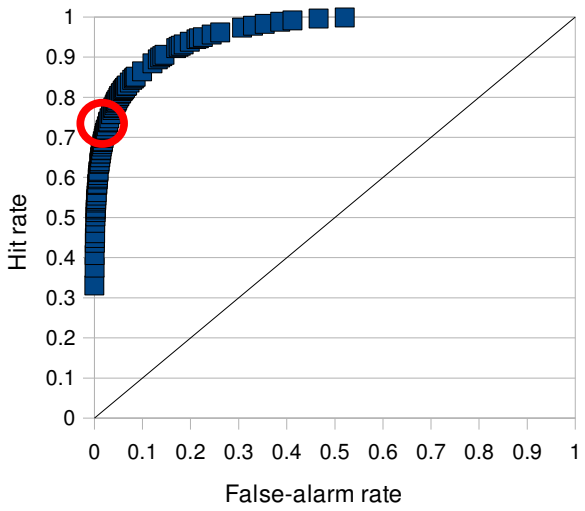


Fig. 2.2: Segmentation ROC for baseline-DiBS

The performance at the MLDT (θ = .5) is given below in the form of Table 2.1:

Baseline	**model detects**	**model not detect**	**%**
true WB	60.6 m (hits)	19.4 m (misses)	75.76% (hit rate)
not WB	8.9 m (FAs)	283.9 m (CRs)	3.05% (FA rate)
%	87.16% (precision)	6.39%	92.41% (accuracy)

Table 2.1: Performance of baseline-DiBS at Maximum Likelihood Decision Threshold

The first two rows and columns (after the header) indicate the raw number of decisions. The last row and column indicate rates, calculated by dividing the first entry of its row/column by the sum of the first and second entries (total accuracy reported in the bottom right cell).

Discussion

The ROC curve shows that the baseline model exhibits three different regimes of behavior. At lower thresholds, the model exhibits a near-floor (<5%) rate of false alarms with a hit rate well above chance[17]. At higher thresholds, the model exhibits a near-ceiling (>95%) hit

17 Chance is defined in a signal detection setting as identifying the signal with some probability p independent of any observable properties of the signal. For a given (x,y) pair on the ROC curve, the hit rate p=y that is expected by chance is binomially distributed according to the false alarm rate x: $yB \sim \text{binom}(x, B)$ where B is the total number of boundary events. In the context of this dissertation, B is so large that almost any difference between x and y will be significant, as illustrated by the following example. For 'large' B (>10) the Central Limit Theorem (Lyapunov 1900, 1901) justifies the use of a normal approximation with μ=xB, σ=$\sqrt{(x(1\text{-}x)B)}$. In the British National Corpus, B=79962011 which is greater than 10 and therefore 'large'. Thus the 95% confidence interval is $x \pm 1.96 \cdot \sqrt{x(1\text{-}x)} / \sqrt{B} = x \pm .3370/\sqrt{B} = x \pm .3370/\sqrt{B} = .0305 \pm .00003769$, so the observed hit rate of .7576 is well outside this confidence interval. Moreover, since B is in general 'large' in this way, I will assume for the remainder of this dissertation that any difference between hit rate and false alarm rate is significant.

rate with false alarms significantly below chance. There is also an intermediate range. The MLDT occurs near the top of the first regime. This is the highest hit rate that can be obtained without incurring a substantial false alarm rate. Thus, at the MLDT, overall decision accuracy is very high, about 92%.

The first point that can be observed is that diphone-based segmentation is indeed a promising approach, yielding an overall accuracy of about 92%. In terms of the two-stage framework proposed in Chapter 1, this means that the segmentation mechanism can indeed accomplish much of the segmentation work on its own – a desirable property if the proposed segmentation mechanism is to explain the developmental fact that segmentation is evident before a sizable lexicon.

A second important point, as discussed in Chapter 1, is that baseline-DiBS exhibits a pattern of *undersegmentation*: better-than-chance detection of word boundaries, without a substantial false alarm rate. To the extent that the baseline model is appropriate as an adult model, this error pattern has implications for the lexical access mechanism. Specifically, lexical access can generally rely on word boundaries discovered by the segmentation mechanism, but must discover some additional ones. Then the primary contribution of the lexical access mechanism to word segmentation is to further segment speech, presumably by matching stored lexical representations at onsets of the partially segmented signal, and positing additional boundaries for unmatched substrings. I will return to these points in Chapter 5, when I consider word learning.

The reliability of diphones in the best case is an important proof of concept for the

remainder of this dissertation. That is because the best case is an upper bound for the actual performance that listeners could exhibit with this approach. Put another way, if the upper bound is not much better than chance, then the cue is functionally useless. Of course, showing that the upper bound is good does not explain human performance, or even demonstrate that humans make use of this cue to solve the task; it simply shows that a diphone-based strategy would work well, if listeners are equipped to use it.

Corpus Experiment II: Canonical and reduced speech

The input to the baseline model in Corpus Experiment I represented a relatively phonemic transcription of English. Each word is realized invariantly in the transcript, with a single canonical form listed by CELEX. Since this input representation differs substantially from the kind of speech that listeners – in particular, infants – appear to get, it is important to determine whether these differences matter. Corpus Experiment II investigates this question by running the baseline model on a spoken corpus containing natural pronunciation variation, the Buckeye corpus (Pitt et al., 2007).

Corpus

The Buckeye corpus website (www.buckeyecorpus.osu.edu) describes the corpus as follows:

> The Buckeye Corpus of conversational speech contains high-quality recordings

> from 40 speakers in Columbus OH conversing freely with an interviewer. The speech has been orthographically transcribed and phonetically labeled.

The speakers in the corpus are age- and gender- stratified. Speech was collected in an interview format, speakers were asked their opinions about a variety of local issues such as sports and politics. The speakers were recorded and their speech was orthographically transcribed. The orthographic transcription was used to generate a phonetic transcription in two steps. First, an automatic speech recognition program was used to generate forced alignments between the orthographic transcript and a phonetic transcript. Next, the phonetic transcripts were inspected and adjusted by a research team (ongoing).

The result of this process was that the Buckeye contains two phonetic transcripts. One is the "canonical" transcript, which lists the canonical pronunciation the forced aligner used to detect wordforms. Because the aligner was designed to dynamically detect wordform variation, and the research team also modified its output, there is an additional "reduced" version of the corpus, which includes reduction processes and other pronunciation variation that is not present in the canonical transcript.

The Buckeye corpus is supplied in the form of a sequence of files which represent samples from the recorded conversations. The "canonical" and "reduced" corpora for this experiment were generated by simply concatenating these files (with necessary text preprocessing). In other words, the corpora used in this experiment collapse across speakers and conversations that are distinguished in the corpus.

It is important to note clearly how the 'reduced' Buckeye corpus differs from the CELEX transcription used in the previous experiment. One kind of difference is that a systematic effort was made to distinguish phone labels on the basis of phonetic evidence (Pitt et al, 2007). Thus, phonetically-trained transcribers went through the entire corpus and approved of corrected the machine-transcribed version. For example, a distinction was made between a full alveolar stop [t] and a flap on the basis of closure duration and presence/absence of voicing. Similarly, a distinction was made between nasalized and oral vowels on the basis of presence/absence of nasal murmur.

In addition to the abovementioned examples, an effort was made to represent conversational reduction processes such as deletion, manner assimilation, and the like. One example frequently cited in the Buckeye manual is the underlying sequence *and then*, in which the medial /d/ is elided, the interdental fricative simplifies to a stop, and the nasal gesture perseveres through the resulting stop, yielding the surface string [ənnɜn]. As a result of these efforts, the Buckeye transcription is inarguably closer to representing certain kinds of contextual variation present in natural speech.

However, the Buckeye corpus is fairly similar to the CELEX transcription in its treatment of positional variation. Like CELEX, it distinguishes syllabic allophones of /l/ and /n/, and adds a syllabic /m/ (as in *prism*). And, like CELEX, it does not distinguish aspirated and unaspirated variants of /t/ – although it does distinguish both a flap allophone and a glottalized allophone. Like CELEX, the Buckeye does not distinguish light and dark allophones of /l/. In other words, some positional variation is represented allophonically, but not entirely consistently across

segments.

A final caveat is in order: the Buckeye corpus may represent the worst-case scenario for infants. This is so because in all languages which have been tested instrumentally, caregivers use a special register known as infant-directed speech (e.g. French, English, Italian, German, and Japanese: Fernald, Taeschner, Dunn, Papousek, de Boysson-Bardies, & Fukui, 1989; English & Japanese: Werker, Pons, Dietrich, Kajikawa, Fais, & Amano, 2007; Mandarin: Papousek & Papousek, 1991). Infant-directed speech can be broadly characterized by hyperarticulation (e.g. expanded pitch range). Many of the phonetic reduction processes that occur in the 'reduced' transcription of the Buckeye are likely to be absent or less prevalent in infant-directed speech.

In summary, the Buckeye is generally more *representative* of the speech input that infants hear than the CELEX-transcribed BNC. First, the Buckeye consists of spontaneous speech, like most of the input to infants, rather than careful/read speech. Second, the Buckeye transcription attempts to faithfully represent a number of types of contextual variation, including manner/place assimilation, segment deletion, foot-medial flapping, and the like. Thus, Corpus Experiment II is intended to test to what extent this kind of variation matters for DiBS.

Method

The baseline model was run on each of the subcorpora described in the previous section. One model was trained and tested on the “canonical” version of the corpus, and the other model was trained and tested on the “reduced” version of the corpus. All other details are as described in Corpus Experiment I.

Results

Figure 2.3 shows the ROC curve for the canonical and reduced corpora, respectively. Note that these curves do not represent output from the same model that was tested on different corpora. Rather, each curve represents the output from a model that was tested on the same corpus it was tested on. As in Corpus Experiment I, the MLDT is indicated with a red circle. The capital letters represent the performance of Goldwater's (G) and Fleck's (F) models on the canonical transcript, and the corresponding lowercase letters indicate the same author's model's performance on the reduced corpus.

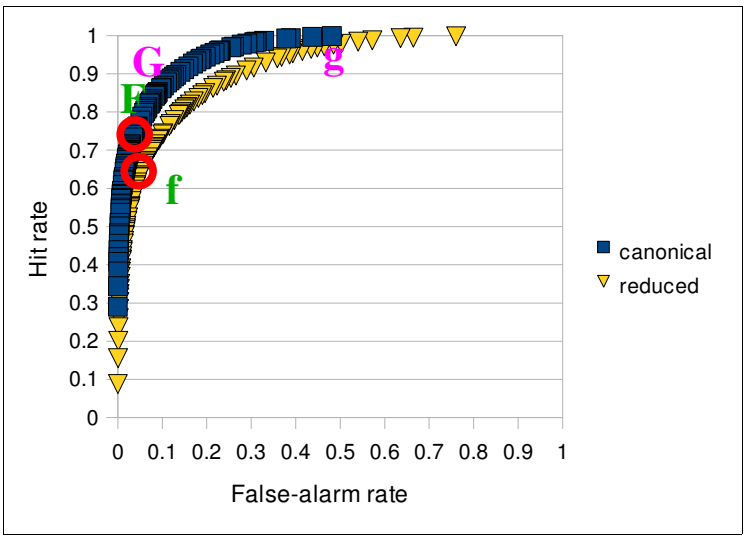


Fig 2.3: Canonical and reduced speech on the Buckeye corpus

The performance of the baseline model at MLDT is compared against the phonotactic model of Fleck (2008) and the Bayesian joint-lexical model of Goldwater (2006) on the same two

subcorpora (originally reported in Fleck, 2008) in Table 2.2:

Corpus	Canonical		Reduced	
	Precision	Recall	Precision	Recall
Goldwater (2006)	74.6	**94.8**	49.6	**95.0**
Fleck (2008)	**89.7**	82.2	71.0	64.1
Baseline-DiBS	86.8	75.9	**82.0**	66.6

Table 2.2: Comparison of segmentation models on the Buckeye corpus

Discussion

The results of Corpus Experiment II replicated and extended the findings from Corpus Experiment I to a corpus containing natural conversational pronunciation variation. Specifically, the baseline model was run on two different versions of the same corpus: in the "canonical" transcription, every word type was realized with a single canonical realization, like in the phonetic transcription of the BNC in Corpus Experiment I; whereas in the "reduced" transcription, word types were transcribed including conversational variation in their pronunciation. The first major result of Corpus Experiment II is that the baseline model exhibited a very similar level of performance on the "canonical" corpus as it did on the phonetic transcription of the BNC, namely it exhibited a pattern of *undersegmentation*. The other major result is that the baseline model, like other leading models of word segmentation, exhibited considerable degradation on the "reduced" corpus.

The finding of *undersegmentation* at the MLDT supplements the findings of Corpus Experiment I. Recall that the BNC was selected as the standard corpus owing to its large size,

and the existence of a comparable corpus in Russian. As a result, the BNC is somewhat lacking in its representativeness, since it largely consists of written language. In contrast, the Buckeye corpus consists of transcriptions of conversational speech. Thus, the results on the "canonical" version of the Buckeye show that the pattern of performance on the BNC is robust across spoken/ written modality. **This result suggests that undersegmentation is the general pattern exhibited by diphone-based segmentation for any English corpus with canonically realized wordforms.** I will discuss the cognitive implications of this finding in more detail in the general discussion; but for the present I turn to the contrasting pattern of results found for the "reduced" corpus.

As evident from Table 2.2, the basic pattern exhibited by each model remains essentially intact, but all models exhibit degraded performance on the reduced corpus. Since these two corpora were identical at the word type level, the degradation owes specifically to the conversational reduction processes described only in the reduced corpus. Thus, the assumption that words are realized with an invariant pronunciation, shared by all previous models of word segmentation and most experiments in this dissertation, has real consequences for their performance. **In particular, natural pronunciation variation has a negative impact on segmentation performance for all models tested.** A natural question is why this might occur.

One clue is given by the fact that Fleck's and Goldwater's models degrade "more" than DiBS in going from the canonical to the reduced corpus: Goldwater's model loses about 25% precision and Fleck's model loses about 18% for both precision and recall. In contrast, the diphone model loses only 5% precision and 10% recall. Presumably the greater decrement in

performance for these models owes to the assumption of a canonical, invariant wordform corresponding to a word type, which is crucially violated in the reduced version of the Buckeye corpus. However, the reason that this violation leads to the observed decrement is different for each model.

In Fleck's model, there are actually two processes at work, the core phonotactic model, and a morphological repair process. The core phonotactic model is presumably degraded by data sparsity – that is, because conversational reduction processes create a larger, sparser, and therefore noisier set of *n*-phones. In addition, the morphological repair process (which attempts to distinguish affixes from function words) is presumably impaired for the same reason – the same affix may be realized in a variety of ways, creating a greater data sparsity problem for the repair process. However, there is nothing in Fleck's model which overtly militates against a larger lexicon – the lexicon is simply the set of observed types.

In contrast, Goldwater's model is biased toward lexicons of a particular size. This stems from the Chinese Restaurant Process adaptor, whose free parameter α assigns higher probability to solutions with a particular frequency distribution. In fact, the very high recall and the very poor precision are symptoms of the model's tendency to oversegment, caused in this case by setting α too low, as I now demonstrate.

The first hint for this effect is that the boundary recall actually increases on the reduced corpus. This means that the model posits more boundaries on the reduced corpus than in the baseline corpus, which implies that the model is explaining the corpus with smaller units. The use of smaller units (e.g. morphemes) has a characteristic effect on the frequency distribution:

there are a smaller number of types, and they are used more frequently. For example, by positing the three plural allomorphs /z/, /s/, and /Iz/, the model saves itself from having to posit hundreds of other plurals; the frequencies of the three plural allomorphs go up, and the stem frequencies go up because now the singular and plural forms are counted as the same word type. This behavior is caused by the concentration parameter, which explicitly biases the model toward a particular frequency distribution. In this case, the solution the model finds is overly biased to recycle existing units. Thus, while the model overwhelmingly finds linguistically meaningful units, those units are not words, but some mixture of words and allomorphs.

More compellingly, the appropriate value for the concentration parameter can be calculated and compared to the value actually used. One way to calculate it is to equate the expected and observed probabilities of a novel word (*a la* Baayen, 2001): $p_{\text{expected}} = \alpha/(\text{N}+\alpha) = p_{\text{observed}} = n_{\text{hapax}}/\text{N}$ where n_{hapax} is the number of observed types N is the number of tokens. For the canonical transcription of the Buckeye corpus, the appropriate value is $\alpha = 3249.57$,[18] quite close to the value of $\alpha = 3000$ that was actually used. However, owing to the implicit assumption of invariant wordforms, Goldwater's model must treat pronunciation variants of the same word as distinct word types. In this case, the appropriate value is $\alpha = 18833.54$[19], much higher than the value actually used. Because the concentration parameter was not set as high as was appropriate, the model was unduly biased to recycle existing material, resulting in aggressive oversegmentation. It is this aggressive oversegmentation that explains the extreme drop in precision between the canonical and reduced transcriptions in Goldwater's model. DiBS fares

18 Canonical transcript: $n_{\text{hapax}} = 3187$, N = 166048
19 Reduced transcription: $n_{\text{hapax}} = 16915$, N = 166048

better in this context because it factors word form variation out of boundary identification.

Regardless of the precise value of the concentration parameter, the more general issue is that the appropriate value for the concentration parameter is roughly equal to the number of hapaxes in the corpus. The concentration parameter is a constant in the Chinese Restaurant Process, whereas the number of hapaxes generally increases with the size of the corpus (Baayen, 2001). In other words, the Chinese Restaurant Process is not fully appropriate as a model of word frequency distributions, because it is biased toward a fixed number of hapaxes, independent of the corpus size. In contrast, DiBS makes no assumptions with regard to word frequency distributions.

General Discussion

To summarize, Corpus Experiment I tested the baseline diphone-based segmentation model on a phonetic transcript of the British National Corpus, and Corpus Experiment II tested it on both a "canonical" version of the Buckeye conversational corpus, and a "reduced" version of the same corpus which included natural pronunciation variation owing to conversational processes such as lenition, assimilation, and elision. The baseline model exhibited a consistent *undersegmentation* pattern in all three cases, with a near-floor false-alarm rate. The overall performance of the baseline model degraded on the reduced version of the Buckeye corpus, a property exhibited to an even greater extent by other leading models of word segmentation. These results serve as an important proof of concept for diphone-based segmentation, showing that the level of performance obtainable in the best case is quite high. Moreover, these results

suggest that the diphone-based approach is fairly robust to the input representation, giving comparable results across a variety of language modalities (spoken/written) and transcription systems.

Cognitive implications

As discussed in Section 3.3.4, the predicted error pattern has significant implications for the larger processes of lexical processing. This is because in a fully-functioning adult system, any incorrect decisions made by the putative word segmentation mechanism must be caught and corrected by downstream processes such as lexical access. The present model exhibited a pattern of undersegmentation, meaning that the model filtered all or nearly all false alarms. As a result, the downstream lexical access processes can confidently rely on the word boundaries supplied by the segmentatiom mechanism, and need only "worry" about recovering additional word boundaries (rather than checking the boundaries supplied by the segmentation mechanism).

Language generality

The results of the baseline model suggests that diphone-based segmentation is of considerable utility in solving the word segmentation problem for English. This finding is encouraging from a developmental perspective, because it means that a prelexical learner could achieve near-adultlike performance on word segmentation, if they were somehow able to estimate near-optimal entries in the parse table.

The significance of such a method, however, depends on whether it could be applied to

word segmentation in a broad class of languages. If the proposed method only worked for English, it would be an interesting curiosity. However, if the proposed method gives qualitatively similar results for many other languages, this would show that it might provide the foundation for a language-general, developmental account of word segmentation.

The first step toward testing the language-generality of a diphone-based approach is to run the baseline model on one other language. In the next chapter, I take this first step by running the baseline model on Russian data, under the most comparable conditions that can be obtained.[20]

Conclusion

In this chapter, I gave a formal description of the word segmentation problem at a categorical phonetic/phonological level. Next, I described a baseline segmentation model and a framework for evaluating it. To determine the potential utility of diphone-based methods, I ran the baseline model on a phonetic transcript derived from the British National Corpus. The results of this baseline run showed that not only do diphones bear considerable information that is relevant for word segmentation, but that the baseline model almost never identifies a word boundary when there isn't one.

I then considered a number of problematic issues in the baseline experiment, such as the relatively abstract character of the input, which may fail to reflect allophonic variation that is relevant for word segmentation in human listeners. To address this issue, the baseline model was run on the more speech-like Buckeye corpus. The general pattern of results was replicated,

20 Of course, in cross-linguistic research, there are always environmental and language-specific factors that cannot be strictly equated across languages. I have made the utmost effort to make the baseline calculations as comparable as possible between the two languages, as discussed in more detail in Chapter 4.

although a detrimental effect of conversational reduction was observed for both this diphone-based model and other extant models of segmentation. These results provide a clear proof of concept for diphone-based segmentation, and suggest that it is robust to some of the variation caused by conversational reduction. However, a number of questions remain, including whether diphone-based segmentation is robust to cross-linguistic variation, or whether it is a strategy which simply happens to work for English – the question to which I turn in Chapter 3.

Appendix 2A

Suppose that the true incidence of HIV is 1/10000, and epidemiologists have created a test which gives the correct diagnosis in 98/100 cases. That is, an HIV+ person is 98% likely to be identified as such, and an HIV- person is also 98% likely to be identified as such. On the face of it, 98% sounds very good. However, let us consider the precision of the test, that is, the probability that someone has HIV, given a positive test result. Suppose that the test was administered to 1,000,000 people. Then the number of HIV+ people is 100, and out of those 98 will be correctly identified as such. Similarly, the number of HIV- people is 999,900, and out of these 19,998 will be *incorrectly* identified as HIV+. It follows that the probability that you have HIV, given that the test says you have it, is 98/(19+19998) = .5%, far less than the overall accuracy of 98%. In other words, the errors are drastically skewed: the vast majority of errors are false positives, in which the test reports HIV+ when the person was actually HIV-.

Should epidemiologists modify the test to have a lower sensitivity? If this were done, the false positive rate would decrease – a highly desirable outcome, because the emotional cost of receiving an HIV+ diagnosis is very high, and no one should have to go through that unless they really have HIV. On the other hand, the true positive rate would also be reduced, resulting in a greater number of misses. The social cost of more misses is very high, as patients may go away from the test believing they are safe, which could result in two kinds of costs. First, they may fail to take precautions to prevent further spread of the disease. Second, delaying treatment is likely to result in an overall more expensive and less effective course of treatment when the disease eventually is discovered. Thus, changing the sensitivity of the test reduces some social and

emotional costs, but increases others. The optimum value for this precision/recall tradeoff is not objectively calculable, but depends on the relative emotional and social costs of false positives versus misses.

Appendix 2B

This appendix presents back-of-the-envelope calculations which suggest that an (English-learning) infant hears on the order of 30,000 words per day during their first year of life. I have estimated this value since to my knowledge there is no more relevant published research on this question.

Some facts about speech are useful for this discussion. First, a fluent adult speaker conversing at a comfortable pace will tend to produce somewhere between 3 and 5 syllables per second (e.g., Kazlauskiene & Velickaite, 2003). Second, English word *tokens* contain an average of 1.44 syllables as calculated by the number of vowel tokens in the BNC, divided by the number of word tokens (tokens rather than types are the appropriate measure here because we are interested in estimating number of words in speech).

Thus, the number of words per second can be estimated as:

$$(4 \text{ syl/s}) / (1.44 \text{ syl/wd}) \approx 2.76 \text{ wds/s} \tag{2B.1}$$

If we assume that a typical English infant hears the equivalent of 3 hours of continuous speech in a day, the total amount of speech can be calculated as

$$(2.76 \text{ wds/s}) \cdot (60 \text{ s/min}) \cdot (60 \text{ min/hr}) \cdot (3 \text{ hr/day}) \approx 29851.2 \text{ wds/day} \tag{2B.2}$$

which can be rounded to 30,000 wds/day or about 1 million wds/year.

An alternative method for estimating this value begins with the average number of words that a speaker of English produces per day. Matthias Mehl used an inobtrusive recording device to record portions of the day-to-day interactions of several populations, finding that American speakers produce an average of about 16,000 words per day (Mehl, Vazire, Ramirez-Esparza, Slatcher, & Pennebaker, 2007).

If it is assumed that a typical infant hears all the output of a primary caregiver, plus output from a combination of other caregivers which is approximately equivalent in volume to the whole output from the primary caregiver, then we arrive at a figure of

$$(16{,}000 \text{ wds/caregiver} \cdot \text{day}) \cdot (2 \text{ caregivers}) \approx 32000 \text{ wds/day} \qquad (2B.3)$$

which is quite close to the figure above of 30,000 wds/day.

Of course, these estimates are highly approximate. The actual amount of input that an infant receives will vary from day to day and from infant to infant based on a broad variety of factors not considered here. The figure of 30000 wds/day (1 MW/year) is intended as a 'ballpark' estimate, so that for most infants it is correct to within a factor of 2 or 3.

CHAPTER 3: RUSSIAN

Abstract

This chapter describes the results from running baseline-DiBS on a phonetic transcription of the Russian National Corpus (RNC). It begins with a description of the Russian language, with particular reference to its morphology, phonology, phonetics, and orthography (the latter being relevant to generating a phonetic transcript).

When she is born, an infant does not know which language she will be exposed to. It follows that her learning mechanism must be able to handle any language she might be exposed to. Thus, from a learnability perspective, one of the best tests for the validity of a learning mechanism is its ability to handle cross-linguistic input.

In the case of word segmentation specifically, the majority of research on word segmentation has been done with English-learning infants and/or English corpora. Thus, for the models that achieve relatively good segmentation, including the baseline-DiBS model of Chapter 2, it is not usually clear whether their success owes to language-specific properties of English, or to general properties of all languages. Testing on typologically diverse languages is important, because it will determine whether a proposed segmentation mechanism is a valid language-universal strategy. This dissertation takes a concrete step in this direction by testing DiBS on Russian.

Below, I review the structure of Russian with an eye toward the factors which are likely to affect word segmentation and DiBS in particular. These include most notable Russian's complex

morphology, its paradigmatic stress system, its prosodic system, and phonetic assimilatory processes.

Русский Язык (The Russian Language)

In this section, I will discuss the structure of Russian with an eye toward its linguistic properties that are likely to substantially affect word segmentation. Broadly speaking, these properties include Russian's complex morphology and its prosodic system. Thus, I will not discuss syntactic and semantic aspects of the language except insofar as they impact word segmentation, e.g. in word formation.

I will begin with an example that illustrates the flavor of the language:

(10) вчера	я	открывал	своё	болшое	окно,	
vče`ra	ja	otkri`val	svo`jo	bol`šoje	o`kno	
yesterday	PRO.1s.nom	open.imp-past.masc	REFL-neut.s.acc	big-neut.s.acc	window-neut.s.acc	

a	сейчас	я	его	закрою.
a	sej`čas	ja	je`vo	za`kroju
and	now	PRO.1s.nom	PRO.neut.3s	close.perf-1s.nonpast

'I was opening my big window yesterday, but I shut it just now .'

Example (10) illustrates a number of important properties of the Russian language, including its extensive inflectional system, the permissiveness of its word onsets (/vč/, /kn/), and

restrictiveness of its word endings (most tokens end in a vowel or sonorant). There is finally the property that Russian is standardly written in Cyrillic, which does not impact word segmentation proper but nonetheless affects the process by which the phonetic corpus is generated. In the subsections below, I consider these subsystems of Russian in turn.

Morphology – Lexical base

Russian is an Indo-European language and as such shares a common historical origin with English (Arlotto, 1972). This common origin is evident in a number of cognate or semi-cognate forms such as *noč/nocturnal* ('night'/'night-dwelling') and *m'ed/mead* ('honey'/'honey-wine), which illustrate the most common CVC pattern for word stems. In the course of its history Russian has also experienced significant lexical borrowing from Ottoman Turkish (e.g. *karan`daš* 'pencil'), from German (e.g. *volk* 'wolf'), and most recently from English (e.g. *`komp'uter* 'computer') (Robert Bird, p.c). In terms of their historical origins and lexical bases, Russian and English are not especially dissimilar. Thus, the shared lexical base is not likely to cause significant differences in segmentation between the two languages.

Morphology – Inflection

Several aspects of the Russian morphological system are of special interest for word segmentation. One is the inflectional system, which has the following properties (Davidson, Gor, & Lekic, 1997; Martin & Zaitsev, 2001):

- nearly all content words and many function words are inflected
- nouns inflect for case (see paradigm below), number (sing/pl), and gender (masc/fem/neut)[21]
- adjective agree in case, number, and gender
- verbs inflect for tense (past/nonpast) and agree in person, number, and gender
- syncretism is pervasive

The formal effect of the inflectional system is illustrated by complete masculine and feminine possessive-adjective-noun paradigms (2) and complete past and nonpast pro-verb paradigms (3) (Davidson et al., 1997, Gor, & Lekic, 1997; Martin & Zaitsev, 2001):

(11)	***masc***	***infls***	'my nice table'	***fem***	***infls***	'my nice cat'
	Nom	-i -i --	мо**й** хорош**ий** стол	***Nom***	-a -aja -a	мо**я** хорош**ая** кошк**а**
	Acc	-i -ij --	мо**й** хорош**ий** стол	***Acc***	-a -aja -u	мо**ю** хорош**ую** кошк**у**
	Gen	-evo -evo -a	мо**его** хорош**его** стол**а**	***Gen***	-ej -ej -i	мо**ей** хорош**ей** кошк**и**
	Prep	-om -em -e	мо**ём** хорош**ем** стол**е**	***Prep***	-ej -ej -e	мо**ей** хорош**ей** кошк**е**
	Dat	-emu -emu -u	мо**ему** хорош**ему** стол**у**	***Dat***	-ej -ej -e	мо**ей** хорош**ей** кошк**е**
	Instr	-im -im -om	мо**им** хорош**им** стол**ом**	***Instr***	-ej -ej -oj	мо**ей** хорош**ей** кошк**ой**

(12)	***pers/no***	***infls***	'PRO speak-nonpast'	***gender***	***infls***	'PRO speak-past'
	1s	ja -'u	**я** говор**ю**	**m**	on -l	**он** говори**л**

21 This is a slight simplification. Nouns belong to a declension class, which typically determines gender, with some exceptions (for discussion see Corbett, 1982). Adjectives agree for gender rather than declension class.

2s	ti -š	**ты** говори**шь**	**f**	ona -la	**она** говори**ла**
3s	on -t	**он** говори**т**	**n**	ono -lo	**оно** говори**ло**
1p	mi -m	**мы** говори**м**			
2p	vi -te	**вы** говори**те**	**pl**	on'i -li	**они** говори**ли**
3p	on'i -'Vt	**они** говор**ят**			

As evident from (2) and (3) (which represent nearly the full range of regular inflectional possibilities in Russian), most word endings are drawn from a small set of sounds. In particular, most nominal/adjectival inflections end in either a vowel, or a sonorant ([m], [j]); most verbal inflections end in a vowel, a sonorant ([l], [m], [j]) or one of two consonants ([t], or [š]); and most function words also end in a vowel. In fact, the only words which systematically do not exhibit this property are masculine nouns in the nominative singular and some third-declension nouns, e.g. *volk* 'wolf'. Generally speaking it is only these cases that a stem-final consonant will appear word-finally. In other words, the inflection system imposes probabilistic but quite strong constraints on the distribution of phones word-finally.

This is a clear and significant difference from English, in which it is easy for words to end with most consonant phonemes in the language, the tense vowels, and schwa. In other words, the statistical signature of a word ending differs quite a bit between English and Russian, owing in large part to Russian extensive inflection system.

Morphology – Word formation

Like English and German, Russian is relatively permissive in terms of combining

morphemes to yield new words. For example, in both *dostaprima`čatel'ni* 'sites of interest to tourists' and *zlo`radstvo* 'pleasure at another's misfortune (lit. evil-happiness)' at least 3 pre-inflectional morphemes can be discerned. Word formation is accomplished both by prefixes and suffixes, and many prefixes derive historically from prepositions; moreover many such prefixes continue to function as prepositions, e.g. *po*, *do*, *ot*, *za*, *na*, *v*, *s*, and *k*. This fact is of special significance for word segmentation, since one and the same phone string is identified as sometimes a word, and sometimes incorporated into another (following) word.

An example where this process is specially evident is in the formation of perfectives. In Russian, aspect is realized lexically via aspectual pairs (Davidson et al., 1997; Martin & Zaitsev, 2001). Typically though not exceptionlessly, the perfective verb(s) stand in an apparently derivational relationship to the imperfective; namely a perfective form is obtained by prefixing the imperfective. For example, *p'isat'* 'write.imp' can be prefixed to yield the (attested) verbs *dop'isat'* 'finish writing (perf)' and *zap'isat'* 'write.perf (for some purpose)'. In addition, novel/unattested perfective verbs can easily be formed in this way, with the choice of prefix conveying relatively subtle meaning contrasts. As stated above, *do* and *za* are highly frequent prepositions which may occur as separate words orthographically. Thus, this process of word formation is likely to cause difficulties for a segmentation algorithm whose performance is scored according to whether it segments the prepositional forms yet does not segment the exact same phoneme string when it occurs as a prefix.

Phonology & Phonetics – Segmental inventory

Russian has a standard 5-vowel system: /a/, /e/, /i/, /o/, and /u/ (Davidson et al., 1997; Martin & Zaitsev, 2001; Hamilton, 1980). As discussed in more detail in a later subsection, there are several allophones of these underlying vowels which are triggered by stress and palatalization, for example /a/ and /o/ normally reduce to schwa in the absence of stress.

Russian consonants are distinguished by place, manner, voicing, and the secondary articulation of palatalization. The palatalization contrast is quite extensive in Russian; all consonants have a soft (palatalized) and hard (unpalatalized) variant, except for the following consonants, which are deemed inherently soft/hard owing to articulatory constraints (Hamilton, 1980):

й (soft)	[j]	front glide
ч (soft)	[č]	alveolo-palatal affricate
щ (soft)	[šč][22]	long alveolo-palatal fricative
ц (hard)	[c]	dental affricate
ж (hard)	[ʒ]	voiced retroflex fricative
ш (hard)	[š]	voiceless retroflex fricative

Phonology & Phonetics – Assimilation & Mutation

Just as Russian consonants differ in voicing, palatalization , manner, and place, so may

22 This is the sound that corresponds to the grapheme щ. For simplicity, I follow Avanesov (1967) in transcribing it as [šč].

they assimilate to one another in one or more of these properties.

Russian obstruents devoice word-finally. Moreover adjacent obstruents may not disagree in voicing, so that preceding obstruents assimilate to following obstruents in voicing (Avanesov, 1967; Hamilton, 1980; Hayes, 1984). Thus in *vstretit'c'a* the word-initial /v/ assimilates to the following voiceless /s/, yielding a voiceless [f]. The phoneme /v/ is not fully an obstruent in Russian however (for a recent discussion see Padgett, 2003), for it may disagree with preceding obstruents in voicing, e.g. *dver'* [*d'v'er'*] 'door' vs. *tver'* [*t'v'er'*] 'town'. Aside from this peculiarity, the voicing system of Russian is mercifully simple.

Palatalization assimilation in Russian is complex, variable, and under-researched. Avanesov (1967) lists some general principles but ultimately lists rules by individual segments, although contemporary phonological theory may allow a more incisive treatment by making reference to the syllable (Ito, 1986). I will simply assume that non-labial consonants assimilate in palatalization to following palatalized consonants (with the proviso that inherently soft/hard consonants never assimilate), where labial consonants do not assimilate.

Avanesov (1967) also describes retroflexion assimilation by which underlying dental fricatives (/s/, /z/) assimilate in retroflexion (becoming /š/, /ž/) to a following retroflex consonant (/š/, /ž/, /šč/). He further describes a process of manner dissimulation in which underlying /g/ dissimulates to a fricative when it is followed by an underlying /k/ (palatalized or not). For example, *`l'ogko* 'easy' is realized as [l'oxkə].

The topic of assimilation and other mutation processes in Russian is a complex one, and the processes reported above are surely incomplete. However, the principal objective of this

dissertation is a computational cross-linguistic study of the acquisition of word segmentation, rather than a laboratory-phonological description of the Russian language. Thus I will assume for the present study that the above processes describe enough of the phonology of Russian to provide some insight on word segmentation.

Prosodic system – Syllable structure

While English and Russian both permit lengthy consonant sequences, syllable structure is nonetheless quite distinct. Specifically, Russian is more permissive than English in the onset position, but more restrictive in its codas. For the onset position, Russian allows up to 4 consonants, e.g. *vstretit'c'a* 'meet up', whereas English allows only 3, e.g. *strict*. In addition, Russian onsets may contain stop-stop sequences, e.g. *kto* 'who', and even violations of the sonority sequencing principle, e.g. *lba* 'forehead-gen.s'. For the coda position, however, Russian does not generally permit lengthy consonant sequences (Kochetov, 2002). In marked contrast, English permits stop-stop codas, e.g. *act*, and regularly allows up to 3 consonants in word-final codas, e.g. *irked*, *milked*.

Prosodic system – Stress assignment and vowel reduction

Russian and English share a number of similarities in their stress systems. In particular, both languages have lexically contrastive stress, both exhibit extensive vowel reduction outside stressed syllables, and in both languages, stress is conditioned by other morphological factors, such as affixes which may induce stress on the preceding vowel (e.g. *-ic*: *`hi.sto.ry/hi.`sto.ric*,

-tel': `*dvigat'/dvi.`ga.tel'* 'move/motor').

Stress assignment in Russian depends not only the lexeme, but also varies according to the paradigm. For example, `*kniga* 'book' always has stress on the initial syllable. In contrast, `*gorod* 'city' normally has stress on the initial syllable, but in the nominative plural *goro`da*, it is the final syllable that is stressed. Several distinct stress patterns are attested. Zalizniak (1977) distinguishes the following 10 types, in roughly decreasing order of frequency:

a – stress always on the stem.

b – stress always on the ending

c – stress on the stem in sg., and on the ending in pl.

d – stress on the ending in sg. and on the stem in pl.

e – stress on the stem in sg & nom.pl., and on the ending in the other cases.

f – stress on the ending, except for n.pl.

b' – like b, but the stress is on the stem in instr. sg.

d' – like d, but the stress is on the stem in acc.sg.

f' – like f, but the stress is on the stem in acc.sg.

f'' – like f, but the stress is on the stem in instr. sg.

In addition to assignment of primary lexical stress, Russian has an extensive system of vowel reduction. The following summary refers to the literary standard (Moscow dialect) as described by Avanesov (1967) and Hamilton (1980). Three “levels” of reduction are

distinguished: tonic, pre-tonic, and unstressed. In the tonic (main stress) position, all vowel contrasts are fully realized. In the pre-tonic position (syllable immediately before the main stress), underlying back non-high vowels (/a/, /o/) are merged ([a]), and underlying front vowels (/i/, /e/) are merged ([i]). In unstressed position (elsewhere), back non-high vowels are phonetically reduced to [ə] and front vowels may be phonetically reduced to [ɪ]. For the purposes of /a/-/o/ reduction, an unstressed word-initial vowel behaves like the pretonic position, i.e. word-initial /a/ and /o/ are realized as [a].

Some secondary complication arises owing to the phonetic effects of palatalization. In particular, when the back vowel /a/ is fully stressed and occurs between two palatalized consonants, e.g. *gul`'at'* 'walk, wander', it is realized phonetically as [æ]. Similarly, when the mid-front vowel /e/ occurs before a palatalized consonant it is realized as [e], but otherwise as [ɜ]. Moreover, when the back vowel /a/ follows a palatalized consonant (i.e. when it is spelled я), it behaves like a front vowel for the purposes of vowel reduction, i.e. reducing to [i]/[ɪ] in pretonic/unstressed positions.

The full system of phonetic realizations is reported below with Ç indicating a palatalized consonant:

<table>
<tr><th>Environment: Ç_</th><th>tonic</th><th>pretonic</th><th>unstressed</th></tr>
<tr><td>/i/</td><td>[i]</td><td rowspan="3">[i]</td><td rowspan="3">[ɪ]</td></tr>
<tr><td>/e/</td><td>[ɜ], or [e]/_Ç</td></tr>
<tr><td>/a/</td><td>[a], or [æ]/_Ç</td></tr>
<tr><td>/o/</td><td>[o]</td><td colspan="2">--[23]</td></tr>
<tr><td>/u/</td><td colspan="3">[u]</td></tr>
</table>

Table 3.1: Russian vowel contrasts and reduction after palatalized consonant

23 In Modern Russian, the back mid-vowel never occurs after a palatalized consonant except under stress.

Environment: elsewhere	tonic	pretonic	unstressed
/i/	[i]	[i]	[ɪ]
/e/	[ɜ], or [e]/_Ç		
/a/	[a]	[a]	[ə]
/o/	[o]		
/u/	[u]		

Table 3.2: Russian vowel contrasts and reduction in non-post-palatal environments

In summary, the effects of stress on word segmentation should be broadly similar between English and Russian, since the stress systems themselves are so similar. In both languages, there is a single primary stress per content word and extensive vowel reduction. The location of primary stress is variable in both languages, perhaps more so in Russian[24], but not a fully reliable cue to the location of a word boundary in either language.

Orthography

A final contrast between Russian and English is the orthography. While not strictly relevant to word segmentation proper, the orthographic system of Russian was an important factor in generating the phonetic transcription used in this and future experiments. The Russian alphabet is relatively phonemic, in the sense that the phonetic form of a word can be predicted

24 A number of studies have documented stress regularities in English (Cutler & Carter, 1987; Kelly & Bock, 1988; Cassidy & Kelly, 1991). In particular, Kelly and Bock (1988) found that stress patterns were distributed asymmetrically in English according to grammatical category. In a sample of 3000 nouns and 1000 verbs, Kelly and Bock (1988) found that 94% of the nouns were trochaic (strong-weak stress pattern), while 69% of the verbs were iambic (weak-strong stress pattern). Conversely, 90% of the trochaic words were nouns, while 85% of the iambic words were verbs. The greater variety of inflectional patterns suggest that the situation is more complex in Russian, but to my knowledge, the comparable study has not been done.

from the orthography if the position of the main stress (and the general system) is known. Like English, lexical stress is not represented in the modern orthography.

One confusing aspect of Cyrillic, for eyes that are accustomed to the Roman alphabet, is that many letters are shared, but some of those shared letters have different meanings. The sound correspondences of the Russian alphabet are given below, subdivided by whether the letters have the same meaning as in English, a different meaning, or are not similar to any English letter:

Similar	Sound
а	/a/
е	/e/
к	/k/
м	/m/
о	/o/
с	/s/
т	/t/

False friends	Sound
в	/v/
н	/n/
р	/r/
у	/u/
х	/x/
ъ	hard sign
ы	hard /i/
ь	soft sign

Dissimilar	Sound
б	/b/
г	/g/
д	/d/
ж	/ʒ/
з	/z/
и	/i/
й	/j/
л	/l/
п	/p/
ф	/f/
ц	/c/
ч	/č/
ш	/š/
щ	/šč/
э	hard /e/
ю	soft /u/
я	soft /a/
ё	soft /o/

Table 3.3: Russian orthography and phonetic interpretation

As described in an earlier subsection, Russian has an extensive secondary palatalization contrast. While palatalization is without a doubt realized phonologically on consonants, it is realized orthographically on vowels (and in many cases its clearest phonetic correlates are also signaled by vowels). Specifically, the graphemes я, е, и, ё, ю, and ь indicate that the preceding consonant[25] is palatalized; whereas the graphemes а, э, ы, о, and у, indicate that the vowel does not follow a palatalized consonant (the grapheme ъ indicates the preceding consonant is unpalatalized).

There are some additional complications in the spelling system which arise from various phonetic and historical factors. First, certain consonants are inherently palatalized (й, ч, and щ) or unpalatalized (ж, ц, ш), and in these cases, the soft/hard contrast on the vowel grapheme is meaningless. Second, owing to the multiple waves of velar palatalization in the language's history, Russian enforces several phonologically unnecessary spelling rules:

5-letter spelling rule: after ш ж щ ч and ц write о if that syllable is accented and е if it is not
7-letter spelling rule: after к г х (velars) and ш ж щ ч (hushers) never write ы but always и
8-letter spelling rule: after к г х (velars), ш ж щ ч (hushers), and ц never write я/ю but always а/у

These rules are phonologically unnecessary for hushers (inherently soft coronal fricatives and affricates) and ц (hard affricate) because they are inherently soft or hard (so the palatalization contrast does not need to be signaled on the vowel). They are phonologically unnecessary for the

25 И may occur word-initially without a preceding consonant; the others indicate a preceding /j/ when no other C is present.

velars because they are obligatorily palatalized before /i/ and unpalatalized elsewhere.

Another historical/phonetic issue pertains to the sequence MvM where M is a mid-vowel. This sequence is robustly attested in high-frequency and functional items, in particular in the adjectival masculine genitive singular ending *-ovo/-evo*, in the pronominal masculine accusative/genitive *jevo* 'him', and the word for 'today' *sevod'n'a*. Historically this sequence originated from an underlying MgM sequence, and it is still spelled with the г (/g/) grapheme although it has been pronounced with a /v/ for over a century (Avanesov, 1967).

A final historical event was the Bolshevik reform of 1918 (Izvestya, 1918),[26] during which two changes were instituted, one helpful for the present purposes and one unhelpful. Unhelpfully, the grapheme ё was abolished, thereby obfuscating the underlying back-front contrast between the stressed mid-vowels /o/ and /e/ following a palatalized consonant. Helpfully, any remaining fossilized word-final soft signs (ь) were abolished, so that the word-final soft sign assumed the same meaning it had in other positions: a fully predictive signal for the presence/absence of an underlying preceding palatalized consonant. (For third-declension nouns whose stems end in an inherently soft, a word-final soft sign was retained to signal female gender; in addition, the soft sign after the second person singular verbal agreement morpheme шь was retained. Fortunately, both of these exception are phonologically vacuous since the soft sign cannot change the inherently soft/hard status of the consonant it follows).

Implications for word segmentation

To summarize, the properties of Russian that seem most relevant for prelexical word

26 The Russian National Corpus contains materials from both before and after the reform.

segmentation lie in its prosodic system and complex morphology. Prosodically, Russian allows quite complex phonotactics (e.g. 4-consonant sequences word-internally) and has a complex stress system, with lexically contrastive stress and extensive vowel reduction. However, Russian and English differ in their syllabification; Russian is generally more permissive in the syllable onset and less permissive in the coda than English. The syllabification differences are paralleled by differences in the inflectional morphologies. The net effect in Russian is that the most common word endings are drawn from a small subset of the total segmental inventory including the vowels and sonorants, but almost none of the obstruents; in other words, there are probabilistic, but quite tight constraints on the word-final distribution. In contrast, English words can and do end in nearly all of the segments in the language; the word-final distribution is much looser than in Russian.

Phonetic Transcription of the Russian National Corpus

This work is based on the University of Leeds copy of the Russian National Corpus (http://corpus.leeds.ac.uk/ruscorpora.html). Thanks to work by Serge Sharoff, it provides a richer representation than the parent copy hosted in Russia in that it is lemmatized and part-of-speech tagged. Short subsections of this corpus were downloaded, preprocessed, and phonetically transcribed serially, resulting in a phonetic transcription of the entire corpus. Preprocessing was exactly analogous to the BNC, e.g. stripping word-external punctuation and de-capitalization. The phonetic transcription process consisted of three sub-processes: recovery of phoneme string, stress assignment, and phonetic processes. These processes are described in more detail below.

Phoneme string recovery

As discussed in the orthography section above, Russian orthography is essentially phonemic, with a few exceptions. Fortunately, most of the exceptions are phonologically vacuous; one example is the spelling rules; another is the word-final *mjakij znak* (soft sign) on the second-person singular nonpast verbal conjugation (e.g. говоришь). Thus, phoneme strings were generally created with the simple expedient of a translation table.

However, there are three major exceptions to the generally phonemic nature of Russian orthography. The first exception is that orthographic MgM sequences (where M is a mid-vowel) underlyingly represent MvM sequences (e.g. the [v] in *jevo* 'his' is spelled with the grapheme that otherwise represents [g]). This exception was handled with a context-sensitive rewrite rule. The second exception is that the post-palatal back mid-vowel (traditionally written ë) is not distinguished from the front mid-vowel; in the contemporary standard, both are written e. This was handled by consulting a freely available electronic copy of Zalizniak (1977), which lists whether a word contains an underlying ë. The final exception is the palatalization system. In general, palatalization is spelled not on the consonant it occurs on, but on the following vowel. Palatalization was handled by 'moving' palatalization from the soft-series vowel or soft sign onto the preceding consonant; there are additional subtleties to this process which need not detain us here but are spelled out fully in the transcription code (which can be obtained by contacting me if it is not available from my website).

In addition, I included a process of phonological liaison for the single consonant

prepositions *v*, *s*, and *k*, which are typically syllabified with the following word.

Stress assignment

Stress is not marked in Russian orthography. (Thus, the phoneme strings in the previous subsection do not include stress). However, as discussed in a previous section, stress conditions vowel reduction, and is therefore crucial to obtain a phonetic transcription. Fortunately, some stress information is listed in the electronic Zalizniak mentioned in the previous subsection. More specifically, Zalizniak (1977) lists the position of the main stress in the *headword* (a canonical realization of the lexeme for listing, e.g. the nominative singular for nouns and the infinitive for verbs) and then indicates a letter code which corresponds to the stress pattern.

Correct recovery of the stress position requires three steps:

1. recognize the headword that corresponds to a token
2. recognize the inflectional properties of the token (e.g. for a noun, the case and number)
3. generate stress position using the listed stress code

Since the University of Leeds copy of the RNC is lemmatized, the first step is essentially included in the corpus. However, steps 2 and 3 are not included in the corpus. Collectively, steps 2 and 3 amount to a full generative model of the inflectional system of Russian, which is itself a near-dissertation sized project.

Rather than build such a generative model, I adopted a shortcut which was designed cover

the most frequent cases. The shortcut is based on the fact that the two most common stress patterns are fixed stem-stress (Zalizniak's pattern a) and fixed ending-stress (pattern b); most of the other stress patterns are a variation on stem stress. Thus, instead of the full paradigm, stress was assigned according to the following possibilities

word not listed:	stress assigned random to any vowel with equal probability
pattern a (stem stress):	stress assigned to same position as headword
pattern b (end stress):	stress assigned to final vowel
any other pattern:	default to stem stress

The resulting stress assignment does not do full justice to the richness and complexity of Russian, but nonetheless achieves coverage of the most frequent cases. This is evident from counting the number of lemmas with each stress patterns in Zalizniak (1977), as shown in Table 3.4:

Stress pattern	**Number of lemmas**	**Percentage of lemmas**
a	57449	84.1%
variant of a	5278	7.7%
b	4598	6.7%
variant of b	106	0.2%
other	911	1.3%

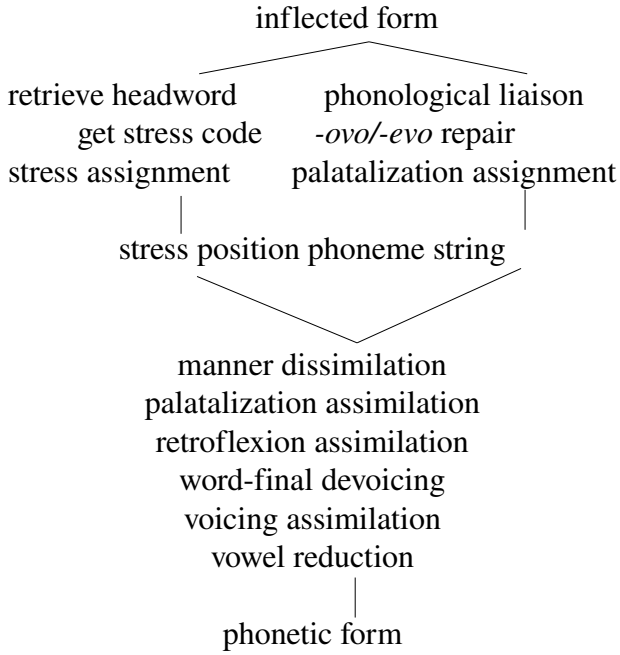
Table 3.4: Stress patterns in Zalizniak (1977) by number of lemmas[27]

27 These counts omit lemmas which are not formatted consistently, and whose stress patterns are therefore difficult to retrieve automatically from the electronic version of Zalizniak (1977).

As shown in Table 3.4, stem-stressed (pattern a) and end-stressed (pattern b) collectively make up over 90% of the lemmas in the dictionary. Together with simple variants of these (in which the stress assignment will be correct for most all but one or two inflections), these lemmas make up 98.7% of the words in Zalizniak (1977). Thus, most known word types will be assigned the correct stress pattern by this algorithm, with the caveat that since high-frequency words are more likely to be irregular, the token accuracy may be slightly lower than the type accuracy.

Phonetic processes

The phonetic form of a word was derived by applying the phonological and phonetic processes described in the summary of the Russian language above. These processes include word-final devoicing, obstruent voicing assimilation, palatalization assimilation, place assimilation, manner dissimulation, and vowel reduction, summarized below:



In summary, the phonetic form of an individual word is derived by creating a phoneme string from the orthography, assigning stress from the headword and stress code in Zalizniak (1977), and then applying hand-crafted rules to derive the phonetic representation.

Corpus Experiment III: Baseline-DiBS on the RNC

The preceding sections gave an overview of the Russian language and a phonetic transcription was generated for the Russian National Corpus. In Corpus Experiment III, I applied the baseline model developed in Chapter II to the Russian data. In other words, Corpus Experiment III is a replication of Corpus Experiment I, but with Russian language data. Except for the input corpus, every other detail of the model is the same.

Corpus

The Russian National Corpus is a large corpus of modern Russian,[28] which was explicitly modeled after the BNC (http://www.ruscorpora.ru/en/corpora-intro.html). As such, it includes a variety of material such as fiction, newspaper articles, and transcribed speech. In the phonetic transcript created as described above, there are 33,876,860 word tokens, somewhat smaller than the BNC but still quite large.

Method

The method was identical to Experiment I of the previous chapter, except the corpus was

28 It includes some materials from the previous century, such as short stories by Pushkin.

different.

Results

The performance of the baseline model on the phonetic transcript generated from the Russian National Corpus is shown below in the form of an ROC curve. The maximum likelihood decision threshold is highlighted with a red circle, as in the previous chapter.

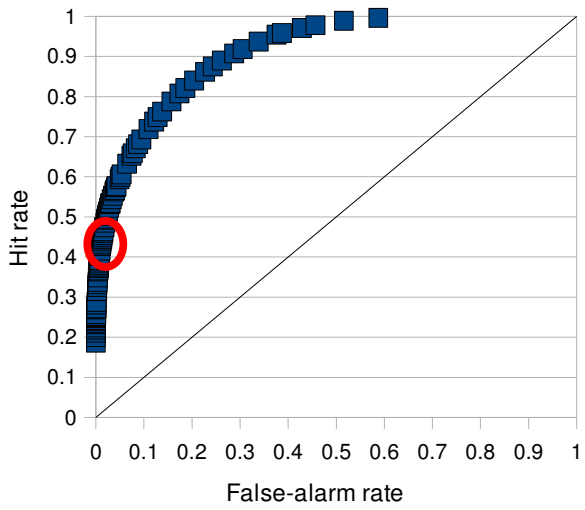


Fig 3.1: Segmentation of baseline-DiBS on RNC

Discussion

The baseline model exhibits a pattern of *undersegmentation* at the maximum likelihood decision threshold (MLDT), with a hit rate of 46% and a false-alarm rate of 1%. The overall accuracy at the MLDT is 92%. For comparison, when the baseline model was run on the BNC, the model (at the MLDT) yielded a hit rate of 75% with a false-alarm rate of 5%, and an overall

accuracy of 92%.

One natural question is why the hit rate is so much lower for Russian than for English, especially given that the overall accuracy was the same across both languages. One reason is that word boundaries are generally harder to find in Russian because there are less of them, and there are less of them because Russian words are on average longer than English words. In other words, the sensitivity of the word boundary detector is lower for Russian than for English, because there is in fact a lower percentage of true boundaries in the signal. Another reason may be the frequent occurrence of prepositions which may also occur as prefixes; since there is a low false alarm rate these items evidently do not cause DiBS to posit word boundaries when they occur as prefixes, which means it is likely they do not cause DiBS to posit word boundaries when they occur as prepositions either.

These results show that DiBS exhibits broadly similar performance on both the Russian and English data. Namely, baseline-DiBS exhibits an overall pattern of *undersegmentation* (high precision with better-than-chance recall) on both languages, with an overall accuracy of about 92%.

More broadly, these results provide support for the hypothesis that DiBS is a language-general word segmentation strategy. This follows from the fact that Russian and English are typologically distinct along two different dimensions. First, Russian is richly inflected whereas English is not; second, English allows highly complex codas whereas Russian does not. As I discussed in greater detail earlier in this chapter, these phonotactic and morphophonological

differences are the ones most likely to matter for word segmentation. The fact that the algorithm exhibits broadly similar performance – in particular, undersegmentation – despite these differences is highly suggestive evidence that the algorithm would perform broadly similarly for any language, meaning that it could be a valid acquisition strategy.

At the same time, it is important to acknowledge the limitations of this experiment. First, Russian, like English, possesses lexical stress; second, and relatedly, both languages have complex phonotactics at syllable and word onsets. It is therefore possible that the results of this experiment crucially depend on the complex onset phonotactics of Russian and English. One way to test this alternative interpretation would be to run the diphone model on a phonetically-transcribed Japanese corpus, since Japanese has a relatively simple phonotactic structure.

An additional limitation of this study is the phonetic transcription process of the Russian corpus itself. While I am proud to have accomplished the task of generating a phonetic transcription of such a large corpus at all, there are several simplifying assumptions which detrimentally affected the quality of the transcript. To select but two examples, the stress assignment algorithm only allowed two options, fixed stress or end stress, whereas the actual paradigmatic possibilities of Russian stress are much richer. Moreover, the palatalization assimilation process is overly simplified, assuming that palatalization assimilation occurs for all and only non-labials. Even with these simplifying assumptions, it was an enormous amount of work to generate this corpus. Thus, it is highly satisfying to find that in broad strokes it replicates the findings of Chapter 2.

CHAPTER 4: LEARNABILITY

Abstract

This chapter develops a Bayesian learning framework for estimating the parameters p(# | xy) for DiBS using a generative model p(xy | #). With the assumption of phonological independence across word boundaries, the generative model can be estimated by factoring it into the word-edge token distributions p(x ← #) and p(# → y). Then, two specific learning models are developed. The lexical learner estimates these distributions from a (possibly small) lexicon, whereas the phrasal learner estimates them from phrase-edge distributions, i.e. without knowing any words at all. It is shown that these learning models achieve performance near the level of the upper bound of baseline-DiBS. For comparison, a range of coherence-based learning models are implemented; it is shown that they fail to achieve good segmentation at any decision threshold.

Diphone models are *prima facie* appealing from a learnability perspective. One reason is that they do not require infants to remember very much of the current input phrase in order to make a segmentation decision; rather, the infant need only remember back one or two segments. Another appeal is that because the diphone domain is small, there aren't very many of them and they occur relatively frequently, which implies that comparatively little training data is needed to estimate parameters. Thus, DiBS is a *prima facie* appealing theory of segmentation, and the previous chapters provide empirical support by demonstrating that the statistically optimal baseline model achieves high accuracy and a similar performance profile in both English and Russian.

However, baseline-DiBS is not a suitable theory of infant segmentation: to calculate the optimal diphone statistics it is supplied with the location of word boundaries in the training corpus, whereas finding these boundaries is precisely the segmentation problem the infant faces. In other words, baseline-DiBS is supervised because it utilizes information that is not observable to infants. This does not mean that DiBS itself is inherently supervised; rather, the challenge is to estimate the relevant diphone statistics using only information that is available to infants. This chapter takes up challenge. As a comparison, it also implements the coherence-based approaches discussed in Chapter 1.

Estimating DiBS from observables

This section addresses the question of how DiBS' model parameters can be estimated from information that is observable to infants. To reiterate briefly, the core parameters of DiBS are statistics of the form

$$p(\# \mid xy) \tag{4.1}$$

which indicate the probability that a word boundary (#) falls between the phones *x* and *y*, given that they have occurred in succession. Thus, this section establishes a framework to estimate p(# | xy) from infant-observable information. To put this discussion on a solid footing, however, it is first necessary to discuss what information is observable to infants.

What is observable?

I assume that infants can observe the following kinds of information:

context-free distribution of diphones

distribution of phones at phrase-edges

frequency of words in their lexicon

context-free probability of a word boundary

Each of these is discussed in turn below.

Before this, however, a word of clarification may be in order. As stated in Chapter 1, I assume that infants perceive speech categorically or have access to a categorical level of representation in which speech is represented as a sequence of 'phones'. By *phone* I mean a sound category which can be reliably distinguished by adults on the basis of acoustic/phonetic and distributional evidence. For example, I would distinguish voiceless, unaspirated [t] from aspirated [t^h] as two distinct phones. In this particular case, there is an alternation between the two phones that is conditioned by the prosodic context: when the phoneme /t/ occurs syllable-initially before a vowel, it is inevitably realized as [t^h] but when it occurs in a syllable-initial *st* cluster, the same phoneme is inevitably realized as [t] (for discussion see Pierrehumbert, 2002). I assume that infants perceive the difference between these two allophones of /t/. However, I do not assume that they have analyzed them as two distinct allophones of the same underlying phoneme. Thus, [t] and [t^h] are reliably distinguished by English-speaking adults in production on the basis of

distributional criteria (each is produced in their appropriate context).

I assume that infants track the context-free distribution of diphones in their input. This assumption, which is shared in some form by all existing models of phonotactic word segmentation, is motivated by evidence that infants attend to local statistical relationships in their input (Saffran et al, , 1996; Mattys & Jusczyk, 2001).

By 'distribution of phones at phrase-edges', I mean in particular the probability distribution over phones in the phrase-initial or phrase-final position. Of course, this assumption presupposes that infants can distinguish phrase boundaries, which indeed appears to be the case, as reviewed in Chapter 1 (Christophe, Gout, Peperkamp, & Morgan, 2003; Soderstrom, Kemler-Nelson, & Jusczyk, 2005). As a consequence, for example, English-learning infants might learn that phrases begin with [h] relatively frequently, but never with [ŋ]; and conversely, that [ŋ] is a reasonably frequent phrase-finally whereas [h] is impossible in that position. This assumption is motivated by pervasive effects in the memory literature of *primacy* and *recency*, i.e. showing that listeners are better able to recall items which they heard first (primacy) or last (recency). That is, given that infants track any distributions at all (a prerequisite for any phonotactic theory of word segmentation), the most conservative assumption is that they track distributions over the positions which are easiest for them to remember and encode, namely the first and last positions in a phrase.

I further assume that infants track the relative frequency of words in their lexicon. There are several reasons to believe this assumption is correct. First, there is a massive body of evidence documenting demonstrating that adults attend to the frequency of words and other linguistic

events in their input (for a review see Jurafsky, 2003). In the absence of compelling evidence to the contrary, the simplest theory is that the mechanisms which cause frequency sensitivity are present from birth (continuity theory of development). Second, although I am unaware of any studies which specifically demonstrate word frequency effects in infants, there are a number of studies that demonstrate frequency effects for other, closely related linguistic units. One such study pertains to phonotactics: by 9 months of age infants prefer high-frequency phonotactic sequences over low-frequency sequences (Jusczyk, Luce, & Charles-Luce, 1994). Another pertains to phonetic categories: infants begin to exhibit adult-like[29], language-specific discrimination earlier for higher-frequency coronal stops than for lower frequency velar stops (Anderson, Morgan, & White, 2003). The final studies pertain to grammatical categories: by the second year of life infants use high-frequency function words to infer the grammatical category of novel words (Mintz, 2003; Peterson-Hicks, 2006). Taken together, these studies suggest that infants attend to frequencies of the many of the same linguistic events as adults do; in particular, these studies suggest that infants know the relative frequencies of words in their lexicon.

Finally, I assume that infants can infer/estimate the context-free probability of a word boundary $\dot{p}(\#)$. To avoid terminological confusion, I will use the notation p(#) to refer to the infant or model's estimate of the probability of a word boundary in the input, and $\dot{p}(\#)$ to refer to the true probability of a word boundary in the input. Note that $\dot{p}(\#)$ is well-defined and can be

29 As a broad generalization, 7-month-olds exhibit a discrimination pattern that is essentially independent of their language background (Kuhl, Stevens, Hayashi, Deguchi, Kiritani, & Iverson, 2006; Tsao, Liu, & Kuhl, 2006). By the time they are 11 months old, infants' discrimination of sounds that are not contrastive in their native language typically declines (Werker & Tees, 1984) whereas their discrimination of difficult native contrasts improves (for exceptions and discussion see Best, McRoberts, & Sithole, 1988; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992 ; Polka, Colantonio, & Sundara, 2001)

calculated a variety of ways; for example, it is the inverse of the average word length. I will argue that it is not difficult for infants to obtain a reasonably accurate estimate for $\dot{p}(\#)$ using observable information and some assumptions about language.

For example, suppose that the infant is exposed to a language with one primary accent per phonological word. In this case, the infant might observe that primary accents tend to be separated from each other by some average number of phones. For concreteness, suppose that the inter-accent interval is on average 4 phones. If the infant is willing to assume that there is one primary accent per phonological word,[30] they are licensed to conclude that the average length of a phonological word is 4 phones, and the probability of a word boundary is inversely related to this length, i.e. p(#) = ¼.

Another way that the infant might estimate $\dot{p}(\#)$ is with a prior distribution over the number of words in a phrase. For example, spoken English phrases typically contain at least 1 and not more than 4 content words. For concreteness, suppose an average of 2.5 words per phrase and an average of 7.5 phones per phrase; then the infant is licensed to conclude an average word length of 3 phones, so that p(#) = 1/3. Alternatively, the infant may have a prior distribution over word lengths, e.g. an innate preference for bimoraic forms[31]. Depending on how frequently the language allows heavy syllables (so that a single syllable is bimoraic), the average word length should be somewhere between 3 and 4 phones, yielding a p(#) somewhere between 1/3 and ¼.

---

30 Infants appear to acquire the rhythmic organization of their language as early as 6 months (Nazzi, Bertoncini & Mehler, 1998; Nazzi, Jusczyk, & Johnson, 2000); certainly English-learning 7.5 month-olds have acquired the generalization that stressed syllables usually signal word onsets.

31 Such an innate preference is consistent with the observations that a number of phonological processes target bimoraic units, including reduplication (McCarthy & Prince, 1986/1996) and nicknaming in Japanese (Mester, 1990; Poser, 1990; Rose, 2005).

It is no coincidence that the estimates of these different methods do not differ very substantially. Rather, there appear to be fairly tight constraints on $\dot{p}(\#)$. For example, Russian and English exhibit average word lengths of 5.90 and 3.82, phones yielding word-boundary probabilities of .17 and .26, respectively. This constrained range of variation can be thought of as a consequence of the Zipfian distribution of language,[32] whereby shorter words are hugely more frequent. That is, even though English and Russian allow very long words such as *antidisestablishmentarianism* and *dostaprimacatel'ni*, words of this length are so rare that they do not really make any difference to the average word length. The average word length, and therefore $\dot{p}(\#)$, is highly constrained. Moreover, an approximate estimate may be quite sufficient. In the present case, the most important factor is whether the estimate is under or over MLDT. Given that p(# | xy) is bimodally distributed with its modes at 0 and 1 (Hockema, 2006), small variations in p(#) are unlikely to cause many word-spanning diphones to be mis-classified as word-internal or *vice versa*. In summary, the assumption that infants can estimate $\dot{p}(\#)$ is motivated by the fact that there are reasonable ways infants could estimate this value; I will therefore assume that infants can estimate this value correctly without explicitly modeling the cognitive processes by which they obtain their estimate.

32 Technically, the mean of a distribution is only well-defined if the distribution is *stationary*, i.e. if different samples are always drawn from the same distribution. This is general not true of Zipfian distributions, and in particular not true for corpus samples. For example, the relative frequency of *Ronald Reagan* in newspaper corpora of 1984 is much higher than on newspaper corpora of 2004, owing to the fact that Ronald Reagan was a more salient public figure in 1984 than in 2004; the opposite applies to *Britney Spears*. In fact, even samples from within the same corpus are not stationary. For example, the BNC contains articles from multiple genres, including newspaper articles, medical reports, and patents. The relative frequency of different words will naturally vary between these different genres; in fact there are likely to be differences in average word length across these genres, e.g. owing to the high percentage of latinate technical terms in medicine. Fortunately, owing to the central limit theorem (Lyapunov 1900; Lyapunov, 1901), the sample mean is bound to not fluctuate too heavily, so it can be estimated.

Bayes' rule

Bayes' Rule allows a conditional probability distribution p(X | Y) to be re-written in terms of the 'opposite' conditional probability p(Y | X). Formally speaking, Bayes' Rule falls out straightforwardly from the definition of conditional probability (Manning & Schutze, 1999):

$$p(X \mid Y) = p(X \wedge Y) / p(Y)$$
$$= p(X) \cdot p(Y \mid X) / p(Y) \quad (4.2)$$

Thus, on the surface Bayes' Rule is simply a consequence of the concept of conditional probability.

The immense utility of Bayes' Rule becomes clear when Y is interpreted as some set of data to be explained, and X is interpreted as a hypothesis space. To make this point explicit, Bayes' Rule is re-written below, with *Hyp* standing for a hypothesis space and *Data* standing for a data set:

$$p(Hyp \mid Data) = p(Hyp) \cdot p(Data \mid Hyp) / p(Data) \quad (4.3)$$

Re-interpreted this way, *Bayes' Rule provides a way to assign probabilities to hypotheses* (Manning & Schutze, 1999). This is desirable from a theoretical standpoint, since the scientist is obligated by the quest for truth to seek hypotheses which are more likely. And to the extent that

learning is like doing science, Bayes' Rule is similarly of utility to the learner, by allowing them to determine which explanations of their environment are good ones.

The crucial ingredients of a Bayesian model are given in Equation 4.3. The term *p(Data | Hyp)* is called the *data model* or sometimes simply a *generative model* (Manning & Schutze, 1999): it assigns probability mass to the data set given some hypothesis. The term *p(Hyp)* is called the *prior* distribution (Manning & Schutze, 1999): it assigns probability mass to different hypotheses based on some prior criterion such as simplicity. The final term *p(Data)* is typically dispensed with, since in practice it functions as a normalization constant whose purpose is to ensure that the *posterior distribution* p(Hyp | Data) is a true probability distribution (Manning & Schutze, 1999).

The practical functioning of these components can be illustrated with the classic example of an unfair coin. Suppose that the learner has observed 8 heads and 2 tails, and is attempting to infer the underlying distribution of heads and tails. (The linguistically minded reader can easily turn this into a language learning problem by interpreting heads as, for example, observations of a surface Verb-Object constituent order, and tails as a surface Object-Verb order.) Further suppose that the learner's model is that coin tosses are independently and identically distributed according to a Bernoulli process with parameter p_{H}, representing the underlying probability of a 'head' outcome, so that the appropriate data model is the binomial distribution. Finally, suppose that the learner considers the hypothesis space of all multiples of .1 for p_{H}, i.e. $Hyp = \{p_{\mathrm{H}} = .1 \cdot n \mid 0 \leq n \leq 10\}$.

The data model can now be used to assign probabilities to the observed data, given some

hypothesis. For example, consider the hypothesis $p_H = .5$. The binomial distribution tells us that the probability of observing 8 heads and 2 tails, given that the probability of a heads is .5, is

$$\begin{aligned} p(8\ H, 2\ T \mid p_H = .5) &= \text{binom}(8,2; .5) \\ &= {}_{10}C_8\ (.5)^8(1-.5)^2 \\ &\approx .0439 \end{aligned} \tag{4.4}$$

where ${}_{10}C_8$ ('10-choose-8') is the combinatorial function. In contrast, the hypothesis $p_H = .8$ assigns higher likelihood to the data:

$$\begin{aligned} p(8\ H, 2\ T \mid p_H = .8) &= \text{binom}(8,2; .8) \\ &= {}_{10}C_8\ (.8)^8(1-.8)^2 \\ &\approx .302 \end{aligned} \tag{4.5}$$

The hypothesis $p_H = .8$ has a special status with respect to these data – it is the observed relative frequency of heads, i.e. the number of heads divided by the total number of observations: 8/(8+2) = 8/10 = .8. It is therefore the hypothesis which assigns maximal probability to the data set[33], and for this reason the relative frequency is called the *Maximum Likelihood Estimator* (MLE) for p_H (Manning & Schutze, 1999).

33 This can be seen from the fact that the derivative of the log-likelihood function at .8 is 0:
$d/dp\ \ln \text{binom}(8,2; p)_{|p=.8} = d/dp\ \ln ({}_{10}C_8\ (p)^8(1-p)^2)_{|p=.8} = d/dp\ \ln({}_{10}C_8) + 8 \ln p + 2 \ln (1-p)_{|p=.8} = 8/p - 2/(1-p)_{|p=.8} = 10 - 10 = 0$.
The logarithm is an increasing function, so a maximum at the log-likelihood must also be a likelihood maximum.

Up until now I have omitted mentioning the *prior*. The prior distribution represents the learner's biases in terms of hypotheses. For example, a reasonable bias in the context of coin-tossing would be to strongly prefer the hypothesis that the coin is underlying fair. This might be encoded by assigning a prior probability of .99 to the hypothesis that the coin is fair, and a prior probability of .001 to every other hypothesis in the hypothesis space. The product of the prior and the data model constitutes the joint distribution over hypotheses and data. For the two hypotheses under consideration, the joint probabilities are shown below:

$$\begin{aligned} \text{p}(8\text{ H}, 2\text{ T} \wedge p_{\text{H}} = .8) &= \text{p}(p_{\text{H}} = .8) \cdot \text{p}(8\text{ H}, 2\text{ T} \mid p_{\text{H}} = .8) \\ &\approx .001 \cdot .302 \\ &\approx .00302 \end{aligned} \tag{4.6}$$

$$\begin{aligned} \text{p}(8\text{ H}, 2\text{ T} \wedge p_{\text{H}} = .5) &= \text{p}(p_{\text{H}} = .5) \cdot \text{p}(8\text{ H}, 2\text{ T} \mid p_{\text{H}} = .8) \\ &\approx .99 \cdot .0439 \\ &\approx .0435 \end{aligned} \tag{4.7}$$

Now, the axioms of probability require that *p(Hyp | Data)* sum to 1. This can only be true if $p(8\text{ H}, 2\text{ T}) = \sum_{p_{\text{H}} \in Hyp} p(p_{\text{H}}) \cdot p(8\text{ H}, 2\text{ T} \mid p_{\text{H}})$, which is a constant. It follows that the relative likelihood of these two hypotheses does not depend on *p(8 H, 2 T)* since it is a constant.

Accordingly, if the learner were forced to select a single hypothesis, the more likely

hypothesis is that the coin is fair. This conclusion is licensed by the fact that although the unfair hypothesis assigns a higher likelihood to the data, the hypothesis that the coin is unfair is a priori extremely unlikely. In this way, the combination of the prior and data model end up selecting the 'best' hypothesis through a combination of factors, including its ability to explain the data (data model) as well as *a priori* theoretical grounds of simplicity (prior).

A final property of Bayesian models can be illustrated by considering the related case in which the learner has now observed 80 heads and 20 tails. Keeping the prior and hypothesis space the same, the new probabilities are:

$$\begin{aligned} \text{p(80 H, 20 T } \wedge\ p_H = .8) &= \text{p}(p_H = .8) \cdot \text{p(80 H, 20 T} \mid p_H = .8) \\ &\approx 9.93\text{e-}5 \end{aligned} \tag{4.8}$$

$$\begin{aligned} \text{p(80 H, 20 T } \wedge\ p_H = .5) &= \text{p}(p_H = .5) \cdot \text{p(80 H, 20 T} \mid p_H = .8) \\ &\approx 4.19\text{e-}10 \end{aligned} \tag{4.9}$$

While the relative frequency of heads and tails has stayed the same, the data are now overwhelmingly more consistent with the MLE hypothesis rather than the fair coin hypothesis. In other words, when there is not a lot of data, the prior exerts an overwhelming effect on the interpretation. When there is a lot of data, it will overwhelm even the strongest prior, provided it allows (assigns nonzero probability mass to) the hypothesis at all.

In the models discussed below, the prior probability will be the context-free probability of

observing a word boundary, and the data model will model the probability of a diphone, given the presence of a word boundary. In other words, applying Bayes' Rule to the basic diphone equation yields the following equation:

$$p(\# \mid xy) = p(\#) \cdot p(xy \mid \#) \,/\, p(xy) \tag{4.10}$$

However, unlike the Bayesian model scenario discussed above, in which the data was fixed and the goal was to select the optimum hypothesis from among a large hypothesis space, it is the data (diphones) which vary here, and the hypothesis space is simply the binary choice between the presence and absence of a word boundary.

Phonological independence

Bayes' Rule provides the first step by which a learner might estimate $p(\# \mid xy)$, because it allows the learner to rewrite this unobservable probability in terms of $p(xy \mid \#)$. The next move is to define a generative model for $p(xy \mid \#)$ whose parameters can be estimated from observables. This can be done, I argue, with the assumption of conditional independence given a word boundary:

$$\text{phonological independence:}\quad p(xy \mid \#) \approx p(x \leftarrow \#) \cdot p(\# \rightarrow y) \tag{4.11}$$

Here I use the notation $p(x \leftarrow \#)$ and $p(\# \rightarrow y)$ to refer to the distribution of phones at word token

edges:

$p(x \leftarrow \#) = p(x\# \mid \#)$	probability of phone x, given word-final position
$p(\# \rightarrow y) = p(\#y \mid \#)$	probability of phone y, given word-initial position

Note that this is not an assumption of phonological independence *within* words. The assumption of phonological independence within words is strongly false. It would, for example, predict that the sequence /kæts/ 'cats' is equiprobable with other licit sequences /kæst/ 'cast', /stæk/ 'stack', /tæks/ 'tacks', /sækt/ 'sacked', /skæt/ 'scat', /æskt/ 'asked', as well as with the illicit sequences /stkæ/, /sktæ/, and so on.

To summarize, the assumption of phonological independence allows the data model $p(xy \mid \#)$ to be factored into two components $p(x \leftarrow \#)$, $p(\# \rightarrow y)$, which correspond to the distribution of phones at word token boundaries. This assumption, though not strictly true, is reasonable for infants before they have had the opportunity to observe any data to the contrary. As discussed in the previous section, it also seems reasonable to suppose that infants can estimate the average length of a word in their language, thereby obtaining the context-free *prior* probability of a word boundary p(#). Thus, the problem of estimating DiBS diphone statistics has been reduced to the subproblems of estimating the distribution of phones at word token edges.

For readers who are acquainted with Bayesian networks, the factored data model can be visualized as a *dynamic Bayesian network* (Ghahramani, 1998), which is a *graphical model* for sequential data. In graphical models, directed arrows represent probabilistic dependencies and

the absence of an arrow crucially represents the absence of a direct dependency (conditional independence). The generative model described in Equation 4.11 can be depicted with Fig. 4.1, where '%' indicates a phrase onset/offset, 'ϕ' indicates phones in the phrase, and '#?' is the random variable indicating the presence/absence of a word boundary:

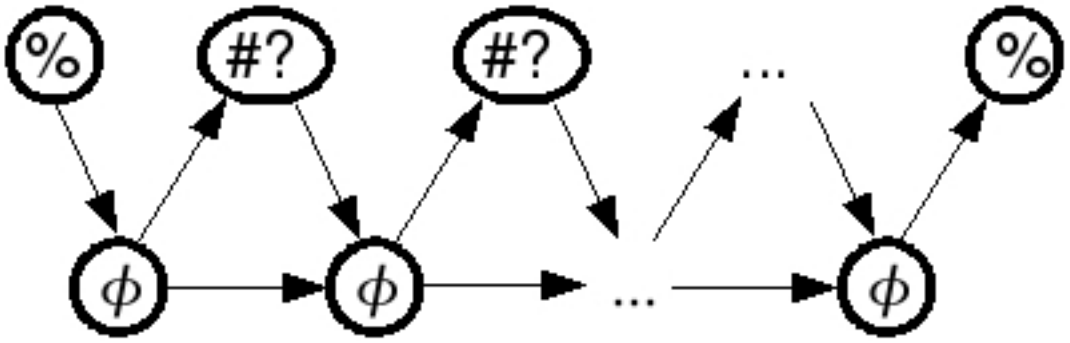


Fig 4.1: Graphical model for DiBS with phonological independence

The repeated configuration in which a phone ϕ_i points to #?, and both ϕ_i and #? point to the next phone ϕ_{i+1} indicate that the next phone is generated from the previous phone, contingent on the presence or absence of a word boundary. The absence of other arrows indicates there are no other dependencies in the generative model; in particular the presence/absence of a word boundary is conditionally independent of the presence/absence of preceding word boundaries, given the value of the intervening phone.

Remaining terms

Although the generative model p(xy |#) is the core conceptual element of this Bayesian

model, the remaining terms p(#) and p(xy) are just as important mathematically. For the present purposes I have assumed these values are available to the learner, as motivated above.

Summary

In summary, the Bayesian formulation provides a principled means to estimate the fundamental DiBS statistic p(# | xy) from simpler distributions, specifically generative models for diphone occurrences conditioned on the presence of a word boundary, p(xy | #), the prior probability of a word boundary p(#), and the context-free diphone probability p(xy). The first term can be factored by the assumption of phonological independence into two models which represent the distribution of phones at word token edges, $p(xy \mid \#) = p(x \leftarrow \#) \cdot p(\# \rightarrow y)$. The factored model p(xy | #?) has a convenient graphical formulation as a dynamic Bayesian network. Alternatively, the learning model can be interpreted in terms of boundary-spanning versus word-internal counts from a 'virtual corpus', which makes the analogy to baseline-DiBS formally rigorous. Under either formulation, the learner need only specify counts or probabilities corresponding to the distributions $p(x \leftarrow \#)$, $p(\# \rightarrow y)$.

Lexical-DiBS

The infant must somehow estimate the distribution of phones at word token edges. By assumption, the infant does not have access to the phrase-medial distribution at word edges. However, if the infant has already learned some words, then she clearly has access to at least some word edges, namely the beginnings and endings of the words in her lexicon. There is one

obstacle to using these words: the words in an infant's lexicon are word types, whereas the needed distribution refers to word tokens. (The token distribution is the one that is encountered in running speech, and is therefore the appropriate domain for segmentation statistics.) This section demonstrates how the token edge distributions can be estimated from types, making crucial use of the assumption that infants know the relative frequency of words in their lexicon.

The idea is to calculate the relative frequency with which a phone begins/ends a word by estimating token frequencies from the lexicon. Some formal definitions may serve to make this notion precise.

Def'n: A wordform ω consists of a string of phones $(\phi_1\phi_2...\phi_n)$.

Def'n: A lexicon Ω consists of a collection of wordforms ω with associated frequencies $f(\omega)$.

Def'n: The notation ω_0 == y is an indicator variable, whose value is 1 if ω's initial phone is [y].

Def'n: The notation ω_{-1} == x is an indicator variable, whose value is 1 if ω's final phone is [x].

Then the edge distributions are given by:

$$p(x \leftarrow \#) = \sum_{\omega\in\Omega} f(\omega) \cdot (\omega_{-1} == x) \,/\, \sum_{\omega\in\Omega} f(\omega)$$

$$p(\# \rightarrow y) = \sum_{\omega\in\Omega} f(\omega) \cdot (\omega_{0} == y) \,/\, \sum_{\omega\in\Omega} f(\omega) \qquad (4.12)$$

The logic of these formulae can be seen by imagining a 'virtual corpus' in which every known lexical type ω occurs with its attested frequency $f(\omega)$. The number of times that [x] occurs word-

finally is the sum over types of how many times it occurs for each word type. For a given type ω, this is either $f(\omega)$ (if the word ends with [x]) or 0 (if the word ends with anything else). The total frequency of word endings in the 'virtual corpus' is simply the total frequency of words.

I refer to this learning model as the *lexical* learner, or lexical-DiBS, because the diphone statistics are estimated from a lexicon. This is relevant to the infant's situation because the infant can exploit this learning algorithm as soon as they have learned a few words. Hence, the strategy is valid even for learners in the early stages of lexical acquisition. Note that unlike baseline-DiBS, which is trained on the occurrence of word boundaries in the corpus, these statistics are estimated from the learner's mental lexicon, even in the early stages of lexical acquisition, when the learner has not acquired very many words.

Phrasal-DiBS

The *lexical* learner described above estimated DiBS diphone statistics from a lexicon, crucially assuming that the learner has access to a lexicon. However, as argued in Chapter 1, infants appear to be able to segment speech before they have acquired much of a lexicon at all. Therefore, a more satisfactory learning account would provide a way to estimate the DiBS statistic p(# | xy) without reference to a lexicon at all. This section addresses that challenge by proposing a *phrasal* learner.

The core idea is the insight of Aslin et al (1996) that utterance boundaries contain information that is useful for word boundaries. This insight can be formalized in DiBS using the notion of edge distributions developed above. It is specifically motivated by the observation that

phrases always begin with a word and always end with a word[34]. Thus, the distribution of phones at utterance edges should be a reasonable proxy for the distribution of phones at word edges:

utterance-edge approximation:

$$p(x \leftarrow \#) \approx p(x \leftarrow \%)$$

$$p(\# \rightarrow y) \approx p(\% \rightarrow y) \quad (4.13)$$

In these formulae, the symbol % refers to a phrase boundary, and the notation $p(x \leftarrow \%)$, $p(\% \rightarrow y)$ refers to the probability of [x] in the phrase-final position, and [y] in the phrase-initial position, respectively.

Corpus Experiment IV: Lexical- and phrasal-DiBS

Corpus Experiment IV is designed to test the *phrasal* and *lexical* models described in the previous sections. Ultimately, what is of interest is how these models perform on a relatively small subset of data, since that is the situation the infant is faced with. However, for maximal comparability to the baseline results in previous chapters, this experiment will train and test the models on the whole corpora. The early lexical model will be tested in a later experiment for its ability to perform based on a small lexicon.

Corpora

The phonetic transcriptions of the BNC and RNC were used, as described in previous

34 Excepting word-medial disfluencies. I assume disfluent phrases can be neglected in modeling.

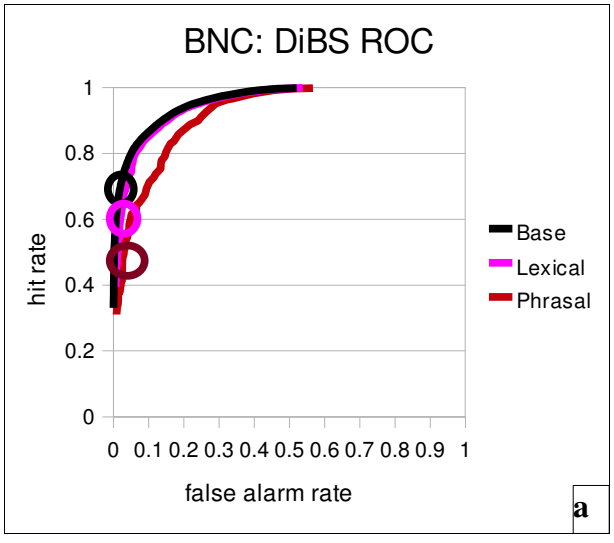
chapters.

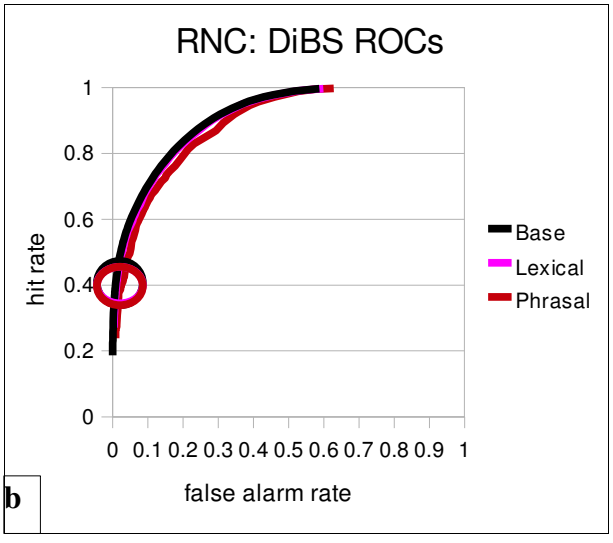
Method

The method is identical to Corpus Experiments I, II, and III, except that p(# | xy) was calculated according to the equations described above for both the *phrasal* and *lexical* learners.

Results

The results are plotted below in an ROC curve for each language, with the baseline shown for comparison. In addition, the F-score is shown as a function of the decision threshold. (The F-measure F = 2PR/(P+R) is a composite measure of precision and recall frequently used in the machine learning literature. It is analagous to accuracy, but adjusted for response bias. In particular, when the signal is rare, it is possible to get good accuracy by never detecting the signal, but this will yield a low F score.)





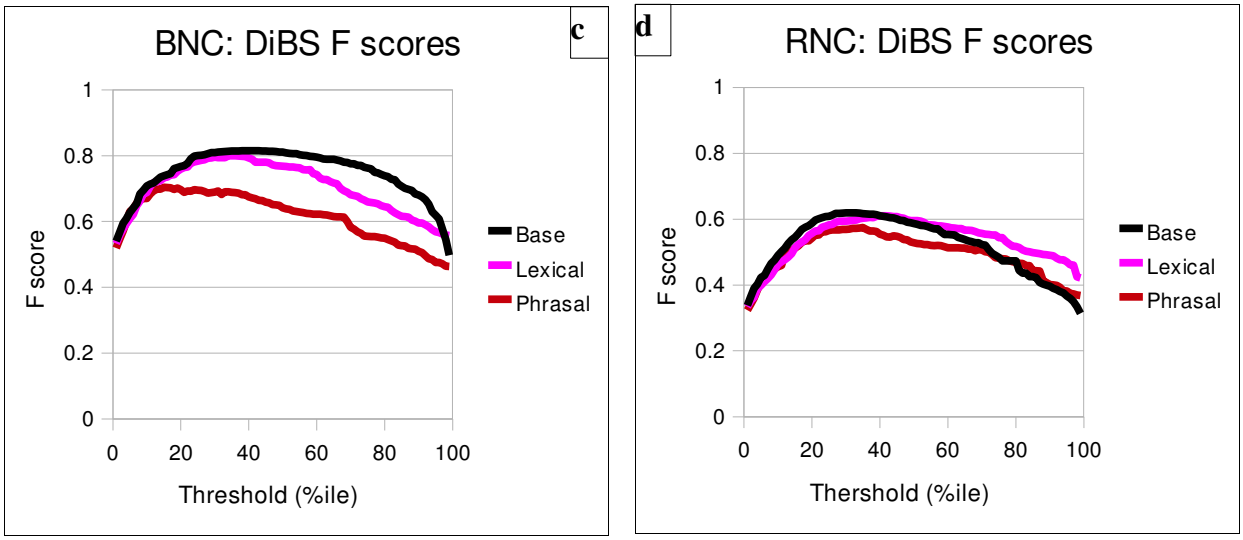


Fig. 4.2: Segmentation performance of learning-DiBS models, including ROC curves for (a) BNC and (b) RNC with MLDT indicated with colored circle, and F score as a function of threshold for (c) BNC and (d) RNC.

Discussion

Experiment IV tests the segmentation performance of the two learning models, phrasal-DiBS and lexical-DiBS, and compares them against baseline-DiBS for both of the large language corpora used in earlier chapters. Comparison of the ROC curves (Fig 4.2ab) illustrates that the learning-DiBS models exhibit performance generally 'close' to that of baseline-DiBS. Lexical-DiBS in particular exhibits almost the same segmentation as baseline-DiBS. Thus, while a reduced level of segmentation performance is evident for either learning-DiBS model (as expected, given that baseline-DiBS is the statistically optimal prelexical parser), the extent of reduction is quite small; both learning-DiBS models achieve a level of segmentation that is near the statistical optimum.

A second important fact is evident from consulting Fig 4.2ab: both of the learning models, like baseline-DiBS, exhibit *undersegmentation* at MLDT (i.e. false alarm rate less than 5%). In other words, a learner using these methods to estimate DiBS statistics is predicted to undersegment. (The F score is reported across all decision thresholds for comparability with other models of word segmentation.)

Corpus Experiment V: Coherence-based models

Experiment IV showed that DiBS can be estimated from phrase-edge distributions and or a budding lexicon, and that once these models are fully trained they achieve favorable performance relative to the baseline model. But, it may be asked, is this a genuine step forward? As reviewed in Chapter 1, a number of other prelexical phonotactic learning models have been proposed centering around various measures of phonological coherence. As further discussed in Chapter 1, these proposals have not been computationally implemented in a rigorous, systematic, and comparable manner. The next section systematically implements a variety of coherence-based approaches, to enable a fair comparison against DiBS.

Corpora

The phonetic transcriptions of the BNC and RNC were used, as described in previous chapters.

Method

The method is identical to Corpus Experiments I-IV, except that word boundaries were identifed using a decision threshold over the following coherence-based statistics (Saffran et al, 1996; Cairns et al, 1997; Hay, 2003; Swingley, 2005):

forward transitional probability	$FTP(xy) = p(xy)/p(x)$
pointwise mutual information	$PMI(xy) = \log_2 p(xy)/(p(x)*p(y))$
raw diphone probability	$RDP(xy) = p(xy)$

The coherence-based measures yield a statistic for every diphone, e.g. FTP(xy) yields the forward transitional probability for the diphone [xy]. In the terminology of Chapter 2, these statistics were mapped to hard decisions using a detection threshold. That is, for some threshold θ, [xy] is treated as always signalling a word boundary if $FTP(xy) > \theta$ to a word boundary, and as always signaling the absence of a word boundary otherwise.

Results

The results are plotted below in an ROC curve for each language, with the baseline shown for comparison. In addition, the F-score is shown as a function of the decision threshold.

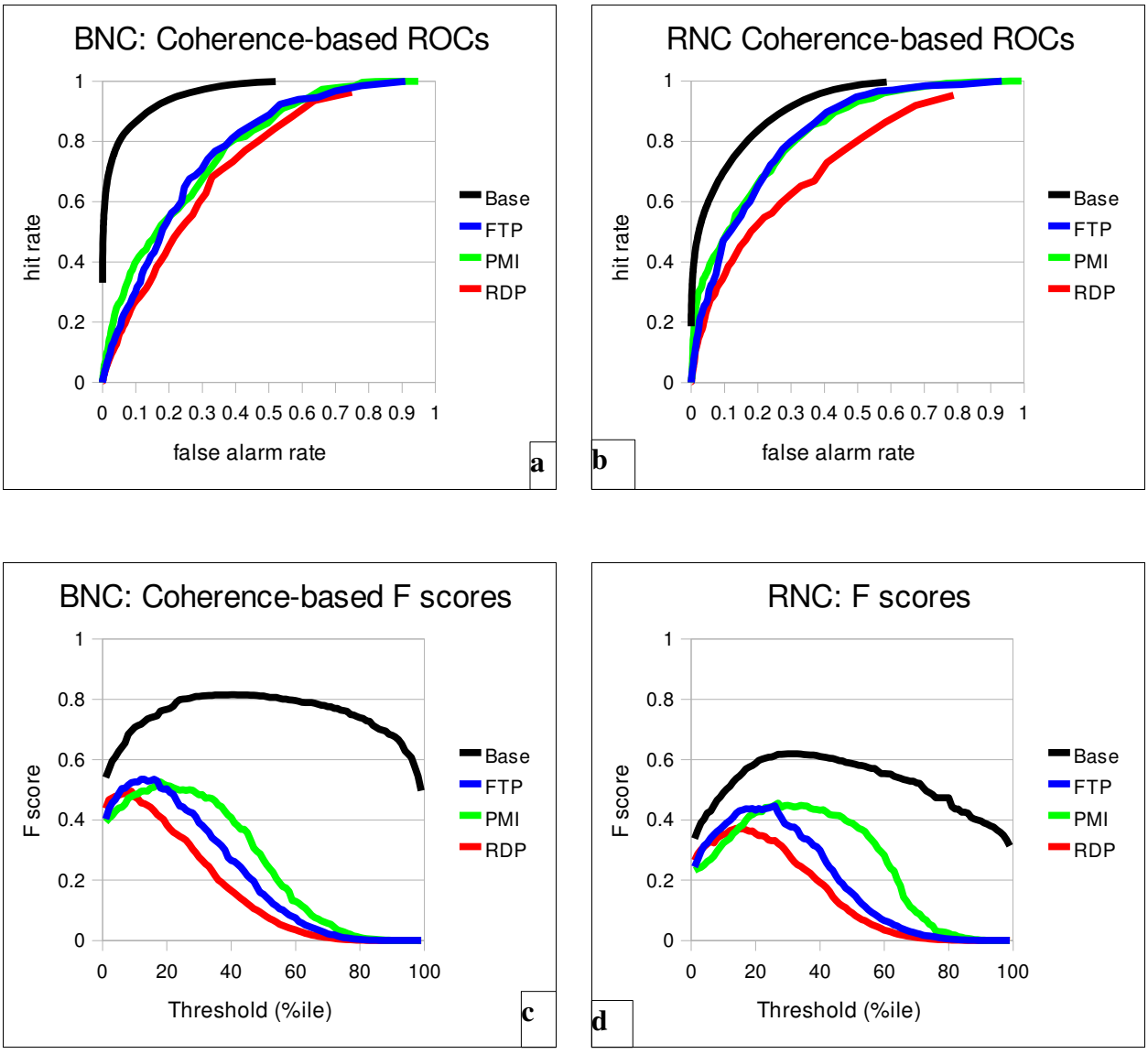


Fig 4.3: Segmentation performance of coherence-based models, including ROC curves for (a) BNC and (b) RNC and F score as a function of threshold for (c) BNC and (d) RNC.

Discussion

The naïve prelexical statistics all yield generally comparable patterns of performance. The pointwise mutual information measure (PMI) appears to be generally more robust (e.g. near

maximum F score for a broadest range of thresholds), and the raw diphone measure consistently yields poorer segmentation, but overall, the three coherence-based measures behave similarly. Moreover, inspection of the ROC curves shows that there is no regime which exhibits undersegmentation or oversegmentation while also exhibiting much-better-than-chance performance. In other words, all of these naïve prelexical statistics exhibit poor discrimination of word boundaries from word-internal diphones, and all exhibit *over+undersegmentation* when they do better than chance.

A natural question is why the coherence-based measures do so much worse than DiBS when superficially they are quite similar. After all, both models are defined with reference to unigram and bigram statistics only. The central difference is in the amount of context that is modeled. The coherence-based models are based on the premise that word boundaries will cause lower coherence, so the presence of a word boundary can be inferred by a lower degree of coherence. Thus, coherence-based models attempt to find word boundaries indirectly, according to a statistic that should be associated with them. In contrast, DiBS models word boundaries directly. Another way to think about this is that DiBS explicitly models word-positional context, whereas the coherence-based models don't. That is, DiBS tracks the relative frequency with which a sound *x* occurs word-initially, word-medially, and word-finally. In contrast, coherence-based models do not. It is the better modeling of positional context that allows DiBS to do so much better than coherence-based models.

Corpus Experiment VI: Bootstrapping lexical-DiBS

Experiment IV showed that both learning models exhibited performance that was generally comparable to baseline-DiBS. However, these models were trained on the entire data set, whereas what is ultimately of interest is the models' segmentation when trained on a limited subset of data – the situation an infant faces. Experiment VI addresses this issue by evaluating the early lexical model's segmentation as a function of lexicon size.

For the greatest verisimilitude, the early lexical learner should be supplied with actual infant lexicons, e.g. from the MacArthur CDI vocabulary assessment forms that parents typically fill out when their children participate in child language research studies (Dale & Fenson, 1996). Unfortunately for the present purposes, individual vocabulary assessments are not a matter of public record in English or Russian.[35] Therefore, infant lexicons were generated for this experiment under the hypothesis that infants learn words according to their frequency. For each such generated lexicon, the early learning model was then applied to calculate the DiBS statistic p(# | xy). Segmentation was assessed at the MLDT, as in previous chapters.

Although this method sacrifices something in the way of realism, it yields a high degree of control. In particular, it is possible to generate a large sample of lexicons which are all matched in overall size. Thus, it can give some idea of the stability of the early lexical model with respect to vocabulary size. In particular, if the algorithm fails to stabilize within some reasonable vocabulary size, this would constitute strong evidence that the algorithm is not an adequate model for infant segmentation. This follows from the fact that infants do vary in their lexicons, but appear to achieve consistent and good segmentation relatively early in development.

35 The MacArthur CDI website (http://www.sci.sdsu.edu/lexical/) reports averages across infants for a particular age. I was unable to find equivalent norms for Russian-learning infants.

The experiment is described in more detail in the following subsections.

Corpora

The corpora are the phonetic transcriptions of the BNC and RNC developed in previous chapters.

Method

To investigate the predicted developmental trajectory, a spectrum of target lexicon sizes was considered. Specifically, the following sizes were preselected:

lexicon size 20, 50, 75, 100, 125, 150, 200, 250, 300, 400, 500, 1000

For each lexicon size L, a sample of L wordforms was drawn. This sample was drawn from the frequency distribution of the corpus. In other words, it was sampled without replacement from the set of all wordforms that occur in the corpus, weighted by the word frequency[36]. Wordforms in the sample were assigned the same frequency with which they occurred in the corpus, preserving their relative frequency distribution.

36 Several caveats are in order. First, word-learning in infants is driven by a variety of factors, of which frequency is only one (Hall & Waxman, 2004). In particular, phonological factors such as phonotactics and lexical neighborhood density affect word-learning (Storkel et al., 2006). All other things being equal, it seems reasonable to suppose that infants are predisposed to learn words which exemplify the most typical patterns of the language, cf. the trochaic bias in English and Dutch (Swingley, 2005). Thus, this frequency-weighted sampling method is likely to overestimate the phonological complexity of the infant's lexicon for large sampling sizes. This effect is somewhat counter-balanced by the Zipfian fact that ultra-high-frequency function words such as *he/on* 'he' and *and/i* 'and' are disproportionately phonologically simple. In small samples these ultra-high-frequency items are likely to be over-represented. The interested reader is encouraged to contact me for further details.

For each lexicon size, 100 lexicons were generated as described above. For each such lexicon, segmentation was assessed on the entire corpus at the MLDT, yielding recall and false alarm rates.

Results

The results are shown below in the form of an ROC curve. It should be noted that unlike in the previous experiments, the ROC curve need not be monotonic, since the underlying parser is changing.

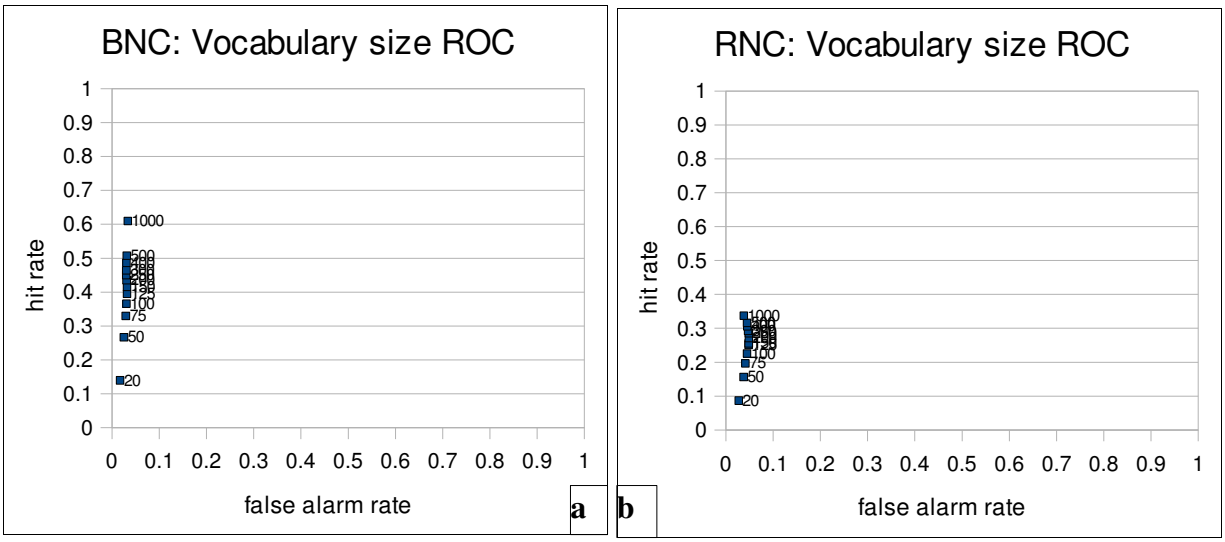


Fig 4.4: Segmentation of lexical-DiBS as a function of vocabulary size in (a) BNC and (b) RNC

Discussion

Consistent with the previous results, the early lexical learner exhibits undersegmentation

for every vocabulary size tested. Moreover, a general pattern is evident whereby the lexical learner's segmentation initially begins with both hit and false alarm rates near 0; as vocabulary size increases, the false alarm rates stays basically constant, but the hit rate increases. By 1000 words, both models have come reasonably close to the adult-level performance, which is a hit rate of 75% for English and of 45% for Russian.

A word is in order about the Russian results. Owing to the extensive inflectional system of Russian, most lexemes have multiple wordforms; for example typical masculine nouns such as *stol* 'table' have about 10 phonologically distinct forms in their paradigms. Furthermore, relative frequency within paradigms is relatively stable across lexemes (Daland, Sims, & Pierrehumbert, 2007), which means that if a wordform is frequent, other wordforms sharing the same lexeme are also likely to be frequent. As a result, the '1000 words' in the Russian data consist of fewer lexemes than the the 1000 words of the English data. Thus, it is not really clear whether 'number of words' is strictly comparable across these two languages. The answer to this question depends on whether infants perceive the relationship between different forms of a paradigm; a question for which research is in its infancy (Kajikawa, Fais, Mugitani, Werker, & Amano, 2006; Fais, Kajikawa, Amano, & Werker, in press). Thus, for the present I will simply note this factor, and pass on the general discussion.

General discussion

This chapter has accomplished one of the main goals laid out in Chapter 1, of providing a full learnability account for prelexical word segmentation. Specifically, the fundamental DiBS

statistic p(# | xy) can be rewritten using Bayes' Rule. With the assumption of phonological independence across word boundaries, p(# | xy) can be estimated using the simpler the word-edge distributions p(x ← #) and p(y → #) (the context-free probabilities of diphones p(xy) and a word boundary p(#) are assumed to be observable). Two learning algorithms to estimate these distributions were then proposed: the *lexical* learner bootstraps from the learner's lexicon, and the *phrasal* learner bootstraps from phrase-edge distributions, without any lexicon at all.

The segmentation performance of these learning theories was compared against both baseline-DiBS and against several of coherence-based threshold models discussed in Chapter 1. The results showed a generally similar undersegmentation pattern as with baseline-DiBS, with some degradation owing to the faulty independence assumptions. However, both bootstrapping models significantly outperformed the coherence-based models in two different ways: first, they exhibited far greater accuracy at all thresholds, and second, at MLDT the DiBS models exhibited undersegmentation, whereas the coherence-based models all exhibited over+undersegmentation.

These results raise a number of issues. One concerns Goldwater's (2006) result that models which assume lexical independence are bound to undersegment, as a result of the large number of collocations in natural language. Given this result, it is natural to wonder whether the undersegmentation pattern exhibited by DiBS is simply a consequence of the same collocational facts. A related issue concerns the relationship between collocations and morphologically complex words; both of which have multiple sub-parts, but whose organization is categorically distinguished in DiBS. More broadly, the clear prediction of these models is ofprelexical undersegmentation throughout the lifespan. This has implications for the architecture of the

lexical access system, in addition to the implications for the wordforms that children will initially acquire. Each of these issues is addressed in separate subsections below.

Collocations: lexical vs. phonological independence

Goldwater (2006) argued forcefully that models of word segmentation which assume lexical independence between successive words are bound to undersegment. This assumption is clearly related to the assumption of phonological independence across word boundaries (called *p-independence* in this section) which is crucial to the bootstrapping models presented in this chapter. Thus, it is natural to wonder whether the observed pattern of undersegmentation for the DiBS bootstrapping models is simply a consequence of the same mechanism that Goldwater uncovered. In this section, I will argue that p-independence is crucially different than the assumption of lexical independence; in other words, undersegmentation in DiBS is not simply caused by a faulty independence assumption.

A brief review of Goldwater's (2006) result is in order. Goldwater first created a baseline version of her model, which made the lexical independence (unigram) assumption; in this model she found that many collocates were posited as single words, yielding undersegmentation. Next, she relaxed the independence assumption by tracking adjacent (bigram) dependencies, and found undersegmentation to be drastically reduced. Then, to show that the effect was due specifically to the independence assumption rather than to the superior statistical modeling properties of the richer model, she randomly permuted the order of words in the original corpus, in effect forcing the lexical independence assumption to be true. She ran the baseline (unigram) model on this

modified corpus, and found that the undersegmentation effect again disappeared. Finally, and perhaps most convincingly, Goldwater initialized the baseline model with the correct lexicon, and found that the undersegmentation effect returned, indicating that the model posited the collocations as new words *even though it already knew the sub-parts were words*. In terms of Goldwater's model, the probability mass lost by positing the novel word was more than compensated for by the probability mass saved in treating the independence-violating collocate as a single word. In effect, the collocation "looked more like a word" (its distribution was more consistent with the model's expectations for a word) than its component words did (since they strongly violated the independence assumption precisely by co-occurring with each more frequently than expected).

Thus, the question is whether p-independence is what causes undersegmentation in the DiBS bootstrapping model. I will argue not. The argument hinges on the fact that p-independence refers to a different level of representation than lexical independence. Thus, even though the two assumptions are related, violations of lexical independence do not necessarily imply violations of p-independence. There are three arguments to this effect. First, collocations would have to target or avoid specific clusters in order to generate a significant violation of p-independence; such an effect would constitute a strong violation of the Saussurian principle of the arbitrary relationship between word meaning and word form (Saussure, 1983). Second, no such violation is apparent; rather, both the English and Russian data show that p-independence is approximately true. Finally, baseline-DiBS does not assume p-independence, but nonetheless exhibits undersegmentation. These points are addressed in turn below.

To see the first point, suppose that English generally obeys p-independence, but has a single frequent collocation, for concreteness suppose it is *million dollars*. There is no effect of this collocation on any word-internal diphones, as they are driven by word frequencies alone. Thus, the only explicit/positive effect is to strongly inflate the boundary-spanning counts for the diphone [nd] over what would be expected under p-independence. There is also a corresponding implicit/negative effect of weakly deflating all the other boundary-spanning counts *under* what would be expected under p-independence (since probabilities must sum to 1). Now let us consider the effect of a related but slightly different collocation, for concreteness suppose it is *million people*. The effects of this collocation are the same: strongly inflating the boundary-spanning counts for [np] over what is expected, and weakly deflating the boundary-spanning counts under what is expected for all other diphones. Crucially, the strong inflation of [nd] is partially countered by the weak deflation caused by [np], and *vice versa*. The inflationary effects caused by one collocation will tend to be countered by the deflationary effects caused by all other collocations with a different boundary-spanning diphone; in other words, the violations of independence caused by one collocation will tend to cancel out the violations caused by most others. Thus, to generate a systematic violation of p-independence, collocations would have to specifically target or avoid particular boundary-spanning diphones.

There is little evidence that this occurs. In fact, p-independence is approximately true for both English and Russian, as shown by Fig. 4.5ab:

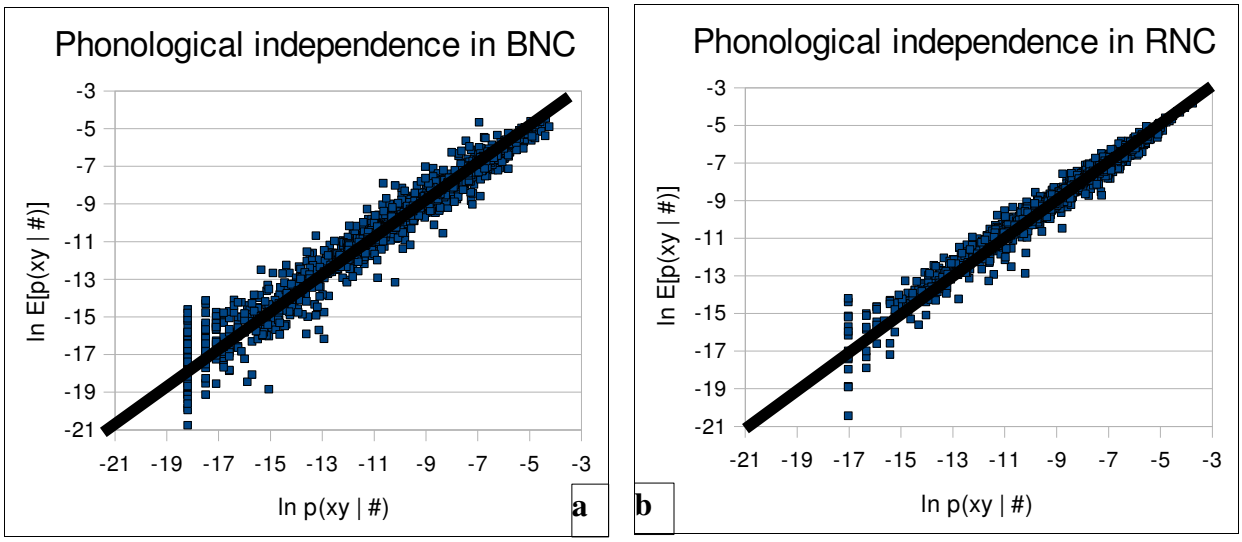


Fig 4.5: Phonological independence in (a) BNC and (b) RNC

The x-axes of Fig. 4.5 represent the natural log of the actual probability p(xy | #), as estimated by normalizing the observed boundary-spanning counts, and the y-axes represent the natural log of $p(x \leftarrow \#) \cdot p(\# \rightarrow y)$, which is the expected probability under p-independence. Each point represents a boundary-spanning diphone, and the identity line represents the shape that is expected if p-independence is strictly true. As shown by Fig 4.2, while there is some deviation from the identity line, the approximation is in general quite good. More specifically, all deviations between log observed probability and log expected are between -2.84 and 3.4, with a standard deviation of .45 in the RNC (between -3.6 and 3.77, standard deviation of .71 in BNC), meaning that the estimated and observed probabilities are always within a factor of 50 (less than two orders of magnitude of error), and the majority of estimated probabilities are correct to within a factor of 2.

The regression slopes and intercepts are given below with their standard errors (95%

confidence interval):

$$\begin{aligned} &\text{BNC:} \quad \text{slope} = .843 \pm .009, \ \text{intercept} = .000 \pm .000 \\ &\text{RNC:} \quad \text{slope} = .968 \pm .005, \ \text{intercept} = .000 \pm .000 \end{aligned} \tag{4.14}$$

Interestingly the slope is slightly less than 1 in both cases, suggesting a small but systematic deviation from phonological independence. In fact, this is a sampling artifact, caused by the fact that ultra-low probability diphones are under-sampled. Since probability distributions must sum to 1, the consequence is that the relative frequency estimator (the 'actual' probabilities on the x-axes in Fig 4.2) will overestimate the true probability of diphones which have been observed (Baayen, 2001). This same effect does not occur with the expected probabilities, which are generated from the much smaller unigram distributions where undersampling is not an issue. In other words, the x-values are inflated over what they should be by this sampling artifact, resulting in a slightly lower slope. Thus, although the slope of the regression line deviates slightly from the expected value of 1, these data are nonetheless strongly consistent with p-independence, demonstrating it is a very reasonable assumption for learners to make in the absence of better data.

There is a final argument which shows that p-independence does not cause undersegmentation in DiBS: baseline-DiBS exhibits undersegmentation, but it does not assume p-independence. This suggests there is some other factor in DiBS besides p-independence which causes undersegmentation, undermining the claim that undersegmentation pattern must be

attributed to the assumption of p-independence.

In summary, while collocations introduce a substantial violation of lexical independence, this does not automatically imply substantial violations of phonological independence (across word boundaries). In fact, the data suggest that p-independence is approximately true in both English and Russian. Thus, undersegmentation in DiBS models is not a straightforward consequence of the collocation mechanism outlined in Goldwater (2006); even if the assumption of p-independence contributes to undersegmentation in the bootstrapping models, it cannot be the full or even the main cause, as baseline-DiBS also undersegments and it does not make the p-independence assumption. The assumption of phonological independence across word boundaries does not cause DiBS to undersegment.

Segmenting collocations vs. morphologically complex words

As reviewed in Chapter 1, there is a conceptual cline between single and multiple words, with complex forms (*trans-Siberian*), compound words (*hotdog, penknife*), and collocations (*apple pie*) exhibiting a mixture of properties that are typical of simplex forms or multiple word sequences. In this section, I consider the evolutionary implications of DiBS for this cline; that is, how the iterative effects of DiBS parsing might accumulate and drive changes in a language's morphophonological structure. However, before this, it is important to distinguish the *conceptual* cline between single and multiple words (discussed above) from the *operational* distinction in the corpus data that DiBS is tested on. Operationally, multi-word sequences are categorically

distinguished from single word sequences, because if the underlying orthographic sequence contains an internal space or other sentential punctuation, the sequence consists of multiple words, and otherwise it is a single word.

The fact that baseline-DiBS does not exhibit a substantial degree of oversegmentation minimally implies that it does not generally parse simplex and complex words into multiple components. Moreover, the generally high accuracy implies that baseline-DiBS generally agrees with the orthography (of English and Russian) in what counts as a word boundary. This fact is hardly surprising, given that baseline-DiBS is trained with corpus data whose word boundaries derive from the orthography. What is surprising, or at least not obviously predicted, is that the two bootstrapping models described in this chapter also generally agree with the orthography. In fact, lexical-DiBS exhibits almost identical performance to baseline-DiBS when supplied with the full lexicon of the language. One way of interpreting these results is that not only does DiBS draw the line between single and multiple words according to phonotactics (by construction), but that the natural languages tested here draw the line in generally similar way.

Additionally, these results replicate and extend the results of Hay and colleagues on phonotactic juncturehood. Hay and colleagues showed that phonotactic juncturehood is highly correlated with decomposability in complex words specifically (e.g. Hay & Baayen, 2002). The present results shows the effects of phonotactic juncturehood prove useful for distinguishing multi-word sequences as well. Moreover, they suggest that natural languages appear to be structured so as to yield near-complementary distributions even for short sequences: *sequences which are licit and typical across word boundaries are systematically absent word-internally,*

and sequences which are licit and typical word-internally tend to not occur across word boundaries.

This interpretation receives some additional support from Pierrehumbert's (1994) study of triconsonantal clusters. This study showed that the set of *attested* word-internal clusters was much smaller than the set of *expected* clusters under the hypothesis that they are generated from the cross product of word-final and word-initial clusters. More specifically, Pierrehumbert conservatively estimated that over 8,700 word-internal consonant clusters were predicted, whereas only about 50 were actually observed.[37] In other words, monomorphemic words appear to systematically avoid word-internal phonotactics that are typical of word boundaries.[38]

In fact, this finding is not just consistent with DiBS, it is straightforwardly predicted by the parsing mechanism of DiBS. To see this, suppose that a complex wordform with strong junctural phonotactics is introduced into the language. For example the [tn] cluster in *post-9/11* is quite rare word-internally. DiBS would accordingly predict that listeners would tend to segment off *post* from the rest of the word, treating it as a free morpheme. Indeed, a recent post on Language Log (http://languagelog.ldc.upenn.edu/nll/?p=1260) provides evidence that *post* has become a preposition in English. For example, in "*Post the wash-out from the credit crunch, most assets globally were overpriced*", *post* both stands alone and takes a multi-word noun phrase complement, clear diagnostics of being a separate word. In sum, as soon as a word with strong

37 Pierrehumbert (1994) argued that many of these 'absent' clusters were not problematic because their expected frequency was less than 1. Even taking this into account, the expected number of clusters was at least 200, so 150 clusters were 'missing'.

38 Pierrehumbert (p.c.) indicated that in compounds which were listed in the dictionary, the boundary-spanning junctures were systematically under-represented. Supposing that dictionary listing is a reasonable proxy for non-compositional semantics, this finding is also explained by DiBS, with the further assumption that words are more likely to undergo semantic drift and acquire noncompositional meaning if they are parsed as single words rather than as multi-word units.

junctural phonotactics is introduced, DiBS will segment the word at the juncture by the very nature of its parsing mechanism.

In fact, this same mechanism may explain the errors the DiBS makes. Recall that compounds occupy the most unclear position on the cline between single and multiple words; indeed, some compounds are variously written as both single, multiple, and hyphenated words (*flowerpot, flower pot, flower-pot*). Even so, compounds are generally written without a space in English, but exhibit most of the other properties of multiple word sequences; in particular, they should exhibit strong junctural phonotactics since they are composed of two free morphemes. The prediction is that DiBS should decompose compounds on the basis of their phonotactics; but since most compounds are written without a space, this would be counted as an error. In other words, the relatively limited degree of oversegmentation that DiBS exhibits may be due to compounds, which to some extent behave phonologically like multiple words. Unfortunately, it is not an easy matter to test this interpretation, as compounds are not marked as such in the corpora used here.

The opposite kind of evolutionary effect is predicted for collocations with weak junctural phonotactics. As discussed briefly in Chapter 2, many of the collocations that are frequently undersegmented are function word + function word sequences such as *he is* and *and in*. Thus, DiBS predicts that these frequent function-function collocations may be parsed as a morphologically complex unit, as appears to have already happened for *he is ~ he's*.

To summarize, the generally good agreement on word boundaries between DiBS, which is entirely phonotactically based, and the language corpora used here, which are orthographically based, suggest that junctural phonotactics are a very powerful determinant of word boundaries in natural languages. This finding can be explained as an evolutionary consequence of DiBS, which will tend to parse items apart if they have strong junctural phonotactics, and together if they lack strong junctural phonotactics. Natural languages are structured so that junctural phonotactics are inherently gradient, but nonetheless exhibit a bimodal distribution, with most sequences being a reliable cue to the presence or absence of a word boundary.

Implications for lexical access

Both baseline-DiBS and the two bootstrapping-DiBS models presented in this chapter exhibit undersegmentation, so it is a clear prediction of DiBS that *prelexical parsing will undersegment the input throughout the lifespan*. Thus, the lexical access system will encounter the characteristic error pattern of undersegmentation throughout the lifespan. As a consequence, the lexical access system might operate more efficiently than would be otherwise possible. This follows from the principle that it is more efficient to do one thing at a time than different things at the same time.

In particular, if the lexical access system is conceived of, at least in part, as a *filter* for parsing errors, there are 2 kinds of errors that in principle it could catch. One error is an error in which the prelexical parser failed to identify a word boundary. In this case the lexical access mechanism must identify the missed word boundary, either by recognizing the end of the

preceding word, or by recognizing the beginning of the following word. The other kind of error is an error in which the prelexical parser has falsely identified a word boundary. In this case, the lexical access mechanism might identify the false boundary by failures of lexical access for the material on both sides of it, together with some kind of repair mechanism. It seems intuitively plausible that the lexical access mechanism would function far more efficiently if it did not have to guard against both of these kinds of errors. This is exactly the prediction of DiBS: the lexical access does not have to guard against both kinds of errors, because DiBS only very rarely identifies a false word boundary.

It would make sense for the lexical access system to take advantage of this fact. One way it could do so is in interpreting lexical access failures as a reliable signal for the presence of a novel word. To see this, let us begin by observing that a genuinely novel word should always trigger a failure of lexical access, since a listener cannot recognize a word they do not know. Now let us revisit the case in which the lexical mechanism has been passed a false word boundary. The most plausible way to catch this error would be by a failure of lexical access, because the false word boundary breaks up the word improperly into nonword subparts. Under this scenario, the failure of lexical access is ambiguous: it might signal the advent of a novel word, or it might signal a false word boundary. If, instead, the lexical access mechanism can rely on the prelexical parser to not pass false word boundaries, lexical access failures will not be triggered in this way. Thus, lexical access failures will *only* occur when a novel word has been encountered. In other words, a failure of lexical access will be an unambiguous signal for a novel word; the lexical access mechanism need not waste precious cognitive resources attempting to discover the false

word boundary and re-parse. In this way, constraints on prelexical errors translate directly into more efficient processing in the downstream lexical access system.

Implications for word learning

Recapitulating this chapter's results, both the baseline model and the two bootstrapping models exhibited undersegmentation. Thus it is a clear prediction of DiBS that the prelexical parser will undersegment throughout the lifespan. In particular, it will undersegment during the initial stages of lexical acquisition. This is a crucial point, because as discussed in Chapter 1, the prelexical segmentation mechanism is what segments candidate word-forms, thereby making them available to be learned. It follows that the set of forms that infants are exposed to as possible words will consist of both properly segmented and undersegmented forms. In other words, the implication of the undersegmentation pattern exhibited by DiBS is that infants will learn both properly segmented and undersegmented wordforms, but very rarely oversegmented forms. This prediction is explored in more detail in the next chapter.

CHAPTER 5: TOWARD WORD-LEARNING

Abstract

This chapter begins by arguing that wordform learning crucially implicates lexical access. Thus, a theory of lexical access is proposed in which phrases (the output of the parser) are further decomposed. The adequacy of the proposed lexical access mechanism is tested by applying it to the output of the parsers developed in preceding chapters. Then, the theory is extended to account for wordform learning: 'candidate' wordforms are added to the lexicon when they have been accessed sufficiently frequently. These components are integrated with the learning-DiBS models from the previous chapter to yield a full boostrapping model.

As set out in the two-stage framework of Chapter 1, this dissertation has proceeded by treating word segmentation as its own problem, related to but logically distinct from word recognition and word learning. The preceding chapters develop models of how prelexical word segmentation might change as a function of increasing lexical knowledge, yielding a full developmental trajectory for this problem. But, because word segmentation is related to word learning, no account can be deemed satisfactory without addressing the nature of that relationship. The goal of this chapter is to do exactly that: develop a model of how the lexicon might change as a function of parsing the input; in other words, how to learn words from the output of the parser.

In the context of the two-stage framework set out in Chapter 1, previous chapters have

concentrated on the representations and processes associated with the prelexical parser specifically. This chapter focuses on the representations and processes associated with the lexicon.

What's in a lexicon?

Minimally, an infant lexicon must consist of some word forms that the infant recognizes as words. Note that this does not imply that the infant must know the meaning of all the words, only that for each form in her lexicon, the infant recognizes the form is a word. (For evidence that word forms can be learned prior to meaning, see Graf-Estes et al, 2007). Thus, as a simplification, I omit meaning entirely here. As an additional simplification, I treat lexical status as binary – specific wordforms are either in the lexicon, or not. Thus, 'wordform learning' is the process of adding a wordform to the lexicon.

While 'being in the lexicon' is binary, I do assume that infants track the frequency of wordforms in their lexicon, as discussed in Chapter 4. Finally, I assume that the lexicon includes 'lexical candidates', sound sequences for which the listener is trying to decide whether they are words or not. This assumption is essentially motivated by the need to account for word-learning, so I will defer further discussion for now to consider the question of what triggers word learning.

Locus of word learning

I claim that word learning crucially implicates the lexical access mechanism (rather than the prelexical parser). This argument is predicated on two core assumptions. First, as laid out in

Chapter 1, I assume there is both a prelexical parser and a lexicon. Second, as laid out in the section above, I assume that wordforms either are in the lexicon, or they are not. There are then two arguments that word learning implicates lexical access.

The first argument for this claim is theoretical simplicity – the normal purpose of the lexical access mechanism is to access the lexicon. Adding a new word to the lexicon certainly requires accessing the lexicon. Thus, it would not be unnatural for the lexical access mechanism to mediate word-learning. On the other hand, the normal purpose of the prelexical parser is to assign phonological parses to incoming material, *before* any knowledge of the incoming material's lexical status becomes available. In other words, the parser is supposed to operate without needing to access the lexicon. If the prelexical parser is what triggers word-learning, it would need to access the lexicon especially for this purpose. Since the lexical access mechanism normally accesses the lexicon, and the parser normally doesn't, it is theoretically simpler and more natural to posit that word-learning is mediated by lexical access.

The second argument for this claim depends on the fact that word-learning processes should only occur for new words. This follows straightforwardly from the assumption that a word is either in the lexicon or not: it is only logically coherent to add a wordform to the lexicon when it is not already there. Determining that a word is novel (not already in the lexicon) requires lexical access; more specifically, it requires a lexical access failure. Thus, there is a simple and natural connection between lexical access and word-learning: word-learning should only be triggered when lexical access fails. This can be straightforwardly accounted for if lexical access is the locus of word learning. If, on the other hand, the prelexical parser is what mediates word

learning, it requires feedback from the lexical access mechanism to determine when a form is new. In other words, the lexical access mechanism has to be involved in word-learning since a lexical access attempt is necessary to determine that a word is new; it is therefore simpler to posit that the lexical access mechanism mediates word learning directly, rather than feeding back to the parser for this purpose only.

For these reasons, I will assume that word learning implicates lexical access failure. Thus, to account for word learning, I must also account for lexical access. The following section sketches a theory of lexical access which is similar in spirit to MATCHECK (Baayen, Schreuder, & Sproat, 2000). The theory proposed below differs from MATCHECK in that it is fully probabilistic. That is, rather than assigning 'activation values' to lexical entries as a function of 'input time', it assigns a probability distribution over possible parses; updating phrase-by-phrase rather than millisecond by millisecond.

Lexical access

Lexical access refers to the process of identifying the lexical elements (words) that correspond to a spoken input representation. In a full theory of speech processing, lexical access should assign a meaning representation to a sequence of forms, but since the focus of the present dissertation is more narrowly on word segmentation, I will simplify matters by treating lexical access as a *matching* process, i.e. simply recognizing word forms as known or not, without regard to their meaning.

The parsing results in preceding chapters demonstrated a consistent pattern of

undersegmentation, meaning that in general, at least for adults, the lexical access mechanism can rely on the prelexical parser's decision when it has positively identified a word boundary, but must consider the hypothesis that word boundaries have been missed in its input. An example is given below to illustrate this point:

(15) orthographic transcription: 'from an infected mother to her baby'

correct phonetic trans'n: frQm {n InfEktId mVD@R tu h3r b1bI

prelexical parser output: frQm{nInfEktId mVD@R tuh3r b1bI

The 'correct transcription' has spaces to indicate the true word boundaries; the 'prelexical output' indicates the word boundaries recovered by the prelexical parser (hard parsing at MLDT). Thus, for this phrase, the lexical access mechanism will receive four distinct inputs. Of these, 'from an infected' and 'to her' will have been undersegmented by the parser, so the lexical access mechanism must further decompose these forms by matching the corresponding known words. The remaining words, 'mother' and 'baby' will have been correctly parsed by the prelexical parser, so the lexical mechanism need only match these full forms (without incorrectly further decomposing these sequences). In other words, the problem faced by the lexical access mechanism is to further decompose its input by matching against known forms.

A related question is the desired *output* of the lexical access mechanism. Since the lexical access mechanism cannot know in advance whether its input will contain word boundaries or not, there are in principle multiple possible decompositions. For example, the word 'captain' could be

parsed as 'cap ten' (assuming an unreduced pronunciation for the second syllable) or 'captain'. Similarly, the phrase 'and it' could be decomposed as 'and it' or as a novel word 'andit'. From a theoretical perspective, the most natural way to address this issue is to consider all possible decompositions, as this is the natural generalization of the 'dual-route' model of inflectional morphology (Marcus, Brinkmann, Clahsen, Wiese, & Pinker, 1995; Baayen, Dijkstra, & Schreuder, 1997; Baayen & Schreuder, 1999; Baayen, Schreuder, & Sproat, 2000; Hay, 2003; Hay & Baayen, 2002). From an evaluation perspective, however, it is much simpler to consider the single best decomposition. As with the parser, I will take a probabilistic approach to this problem: the theory will identify and assign a probability distribution over possible decompositions. However, I will only evaluate the maximum likelihood decomposition, so for the present purposes this will be the only output of the lexical access mechanism. The following sub-sections tackle each of these problems in turn.

Identifying possible decompositions

I define a decomposition as an assignment from the input sequence of phones to a sequence of words. I assume that

(1) a well-formed decomposition partitions the input

(2) decomposition occurs chronologically

Decompositions may be ill-formed in two ways. First, some element(s) of the input may not be

assigned to any words at all. Second, some element(s) may be assigned to multiple words. In other words, a decomposition is well-formed if and only if every phone in the input sequence is explained by exactly one word, and all phones between the initial and final phone of a word belong to that word. An example of a well-formed decomposition is illustrated below:

(16)

input	k{p	tEn	
partition	k{p	tEn	
word seq	'cap'	'ten'	

The implication of these assumptions is that in the process of constructing a single decomposition, the lexical access mechanism need only initiate lexical searches at known word edges. And, since decomposition occurs in the order that words become available), access begins from the 'left'most unmatched edge. Thus, the access mechanism goes through the entire lexicon, identifying any lexical items which exactly match a substring of the input from the leftmost unmatched edge.

For example, for the input phrase 'captain', the current leftmost unmatched edge is the very beginning of the phrase, so the lexical access mechanism would look for words which begin with /k/. It would find the word 'cap', which exactly matches 'captain' on the leftmost 3 segments. At this intermediate stage of decomposition, the lexical access mechanism has matched 'cap' against the longer string 'captain', leaving a *residue*, the as-yet-unmatched right substring '-tain'. The left edge of this residue becomes the new leftmost unmatched edge, so now the lexical access mechanism will try to match '-tain'. The lexical access mechanism finds 'ten', which for

the sake of illustration we will assume has the same segments. Then the lexical access mechanism has exhaustively matched the residue, so this particular decomposition is done.

However, if the lexical access mechanism stopped here, it could only find one decomposition, in this case the wrong one. It must also consider other matches for the input. For example, the lexical search might continue and find the word 'captain'. This item would result in an exact match, fully explaining the input as well. Thus, the lexical access mechanism should find two parses for this sequence: 'cap ten' and 'captain'.

It may happen that the lexical access mechanism encounters some input it is simply unable to decompose in this manner. For example, suppose that the listener encountered a novel name, such as my last name 'Daland'. After searching through the whole lexicon, it cannot find a single word that exactly matches 'Daland' from the left edge. At this point, the lexical access mechanism *fails*, concluding that 'Daland' must be a novel form. In this case, it outputs a single decomposition: 'Daland'.

Owing to the recursive nature of decomposition, lexical access will parse embedded words as long as it can recognize the words that precede them. For example, if the input is 'andit', the lexical access mechanism will first match 'and', and then match 'it', yielding the correct decomposition 'and it'. And, owing again to the recursive nature of decomposition, the proper result will occur even if the embedded word is novel. For example, if the input is 'aDaland', lexical access will first match 'a', then posit 'Daland'. This is because, after the lexical access mechanism has matched 'a', it is in the same position as in the previous paragraph: the leftmost unmatched edge begins at 'Daland'.

A final consequence of this recursive search is that if the lexical access mechanism is unable to find a single form that matches a whole residue, it will posit the whole residue as a new form, even if it is able to decompose the residue. Thus, if lexical encounters the name 'Adele', it will posit the decomposition *a+dell*, but it will also posit the whole form *Adele*. Similarly, if the input 'andit' is encountered, not only will lexical access find the (correct) parse 'and it', it will also posit a new form 'andit'. In other words, multi-word sequences can be identified as constituting a single unit if their phonotactics and relative frequency support such a decomposition.

Note that this property is not specially coded into the model, but emerges from the recursive structure of lexical access and the need to process unfamiliar words. Nonetheless it may form part of the explanation of how distinct words become assimilated into a single word over time, for example *don't know* → *dunno* (Scheibman, 2000). Of course this process necessarily involves a production component with phonological/prosodic reduction (Bien, Levelt, & Baayen, 2005). However, in order to produce these two words as a single one, it is first necessary to perceive them as a functional unit, i.e. to parse them together in perception. Supposing that the prelexical parser consistently segments out this sequence, the lexical access mechanism will posit the whole form *don'tknow* even if it is able to decompose this sequence into its two components. As the phrase *don't know* continues to be experienced at higher-than-expected rates, the holistic parse *don'tknow* may become more and more probable, so that eventually it competes with the compositional parse in perception.

Parse probability and decomposability

In addition to identifying parses, the lexical access mechanism assigns probabilities to distinct parses. This is done both for theoretical and for practical reasons. Theoretically speaking, a distribution over parses is desirable because it provides a way to account for gradience in decomposability, as discussed in more detail below. Practically speaking, it is easier to evaluate a single parse, and assigning a probability distribution provides a principled way to select the best one. Thus, some confusion may arise as to the cognitive stance I am taking. To be clear, I believe that the human lexical access mechanism does assign a probability distribution over possible decompositions of its input, and that the full distribution plays an important role in lexical processing. However, for evaluative purposes, I only use the probability distribution to select the single most likely decomposition. Thus, I will begin by explaining why it is theoretically desirable to compute a distribution over parses, and then I will lay out a theory by which the learner may do this.

It is theoretically desirable to compute a probability distribution over parses because, among other things, such a distribution would provide a neat account of gradience in decomposability judgements. For example, Hay (2003) found that in pairs like *unleash/unscrew*, the relative frequency of the base and derived forms was a strong predictor for relative decomposability judgments; specifically, complex forms which were more frequent than their base (*unleash*) were rated as less complex than complex forms which were less frequent than their base (*unscrew*). A similar effect obtains even when junctural phonotactics are matched, e.g. *swiftly/softly*. This gradience in decomposability judgements is a mystery if it is assumed that both words are fully decomposed or fully undecomposed in listeners' mental lexicons. However,

it can be accounted for straightforwardly if we assume that listeners assign probabilities to both the decomposed and undecomposed forms. A more decomposed judgement could be represented by assigning higher probability to the decomposed parse, and a less decomposed judgement could be represented by assigning lower probability to the decomposed parse.

The same general argument applies for other factors which are correlated with decomposability. For example, Hay & Baayen (2002) illustrate a 'productivity cline' for 80 affixes in English. This gradience in productivity can also be accounted for by assigning a probability distribution over both wholistic and decomposed forms.

A final argument for representing a probability distribution over parses comes from analogy with syntax. Bod (1998) argues at length for representing a distribution over multiple syntactic parses of the same sentence, and implementations of this theory have had remarkable success in explaining acceptability judgements (Bod, 2001), ambiguity resolution (Bod, Scha, & Sima'an, 2003) and construction learning (Borensztajn, Zuidema & Bod, 2008) . To the extent that morphological and syntactic structure are similar (or identical, as assumed in Halle & Marantz's (1993) Distributed Morphology), the same idea should hold for morphological structure as for syntax.

For simplicity, I use a unigram model to assign the probability of a decomposition. That is,

$$p(\omega_1 + \omega_2 + ... + \omega_n) = p(\omega_1) \cdot p(\omega_2) \cdot ... \cdot p(\omega_n) \quad (5.1)$$

For example, the probabilities of the two parses of 'captain' are shown below:

$$p\left(\begin{array}{c} k\{ptEn \\ / \quad \backslash \\ k\{p \quad tEn \\ | \quad | \\ 'cap' \quad 'ten' \end{array}\right) = p(k\{p) \cdot p(tEn) \qquad p\left(\begin{array}{c} k\{ptEn \\ | \\ k\{ptEn \\ | \\ 'captain' \end{array}\right) = p(k\{ptEn) \tag{5.2}$$

In general the probability of a word is proportional to its frequency:

$$p(\omega) = f(\omega)/F \tag{5.3}$$

where F is a normalizing constant chosen to make $p(\omega)$ a true probability distribution. In the simplest case, *F* is simply be the sum of the frequencies over all words.

It may prove instructive to apply this theory to the *unleash/unscrew* example from Hay (2003) discussed above. She lists the following frequencies:

$$\begin{array}{llll} f(\text{leash}) = & 16 & f(\text{screw}) = & 187 \\ f(\text{unleash}) = & 65 & f(\text{unscrew}) = & 44 \end{array} \tag{5.4}$$

Since *un-* is the same morpheme in both contexts, let us assume it has some constant probability p_{un} and not worry about the precise value. Then the probabilities of the decomposed vs. whole

word parses for the complex forms are:

$$\begin{aligned} &p(\text{unscrew}) = f(\text{unscrew}) / F \qquad = 44/F \\ &p(\text{un+screw}) = p(\text{un}) \cdot p(\text{screw}) = p_{un} \cdot 187/F \\ \\ &p(\text{unleash}) = f(\text{unleash}) / F \qquad = 65/F \\ &p(\text{un + leash}) = p(\text{un}) \cdot p(\text{leash}) \; = p_{un} \cdot 16/F \end{aligned} \tag{5.5}$$

If only a single decomposition is selected as the winner, it must clearly be the whole word parse in *unleash*, since $16/F < 65/F$ and $p_{un} < 1$ implies that $p_{un} \cdot 16/F < 16/F$. The same does not hold true in the case of *unscrew*, or rather it depends on the precise value of p_{un}. Thus, in a minimal sense, this theory captures the greater decomposability of *unscrew*.

It is possible to do better than this with more specific assumptions about decomposability. For concreteness, let us assume that 'decomposability' is the log odds of the decomposed parse to the whole parse. Then

$$\begin{aligned} d(\text{unscrew}) &= \log (p(\text{un + screw})/p(\text{unscrew})) \\ &= \log ((p_{un} \cdot 187/F)/(44/F)) \\ &= \log (p_{un} \cdot 187/44) \\ &= \log p_{un} + \log (187/44) \end{aligned} \qquad \begin{aligned} d(\text{unleash}) &= \log (p(\text{un + leash})/p(\text{unleash})) \\ &= \log ((p_{un} \cdot 16/F)/(65/F)) \\ &= \log (p_{un} \cdot 16/65) \\ &= \log p_{un} + \log (16/65) \end{aligned}$$

$$\begin{aligned} d(\text{unscrew}) - d(\text{unleash}) &= \log p_{un} + \log (187/44) - (\log p_{un} + \log (16/65)) \\ &= \log (187/44) - \log (16/65) \\ &> 0 \qquad \text{(from properties of logs)} \end{aligned} \tag{5.6}$$

According to these assumptions, *unscrew* is predicted to be more decomposable than *unleash*, regardless of the precise value of p_{un} of F; in fact, these terms simply cancel out because the prefix and total frequency mass are the same in both forms.

Note that the present theory does not actually decompose complex words unless their constituent morphemes are listed in the lexicon. For example, if *in* (the preposition) is listed but *un* (the prefix) is not, then the lexical access mechanism will posit *in+fected* as one parse for *infected*, but not *un+screw* as a parse for *unscrew*. I will defer the question of whether complex words should be decomposed for the time being, returning to it when I discuss word learning proper.

Reserving frequency mass for novel words

As stated above, the simplest way to calculate the probability of a word is to divide its frequency by the total frequency of all words F

$$\begin{aligned} p(\omega) &= f(\omega)/F \\ F &= \sum_{\omega \in \Omega} f(\omega) \end{aligned} \tag{5.7}$$

However, this would assign zero probability to any word that had not been encountered yet; whereas we do encounter new words throughout life. As discussed at length in Baayen (2001), most individual linguistic events that could happen are quite rare, but taken together they have a significant impact on the statistical behavior of the lexicon, and how speakers process language. Baayen (2001) shows that the probability of encountering a new word p_ν is well-predicted by the relative frequency with which a novel word of frequency 1 (the *hapax* types) has been observed before:

$$\mathrm{p}_\nu = n_{\mathrm{hapax}} / \mathrm{F}$$

$$n_{\mathrm{hapax}} = |\{\omega \in \Omega \mid \mathrm{f}(\omega) = 1\}| \tag{5.8}$$

It follows that the appropriate normalization constant F is the sum of the observed frequency mass (sum over words) and the number of hapaxes (estimate of unobserved mass):

$$\mathrm{F} = n_{\mathrm{hapax}} + \sum_{\omega \in \Omega} \mathrm{f}(\omega) \tag{5.9}$$

Lexical phonotactic model

Note that the probability mass that is reserved for new word forms must be distributed somehow over all possible forms, i.e. a *lexical phonotactic* model. Ultimately this should be done with a probabilistic phonotactic learner, such as the one presented in Hayes & Wilson (2008). For simplicity, I adopted the same solution adopted by Goldwater (2006): a generative model which

first generates the length of a novel word according to a geometric distribution, together with a uniform unigram model that assigns equal probability to every sequence of phones:

$$p(\omega_\nu = \phi_1\phi_2...\phi_n) = (1-p_\#)^{n-1}p_\# \cdot (1/|\Phi|)^n \quad (5.10)$$

where $p_\#$ is the context-free probability of a word boundary,[39] $|\Phi|$ is the number of phones in the language and $1/|\Phi|$ is the corresponding uniform probability of observing a phone. Note that this geometric distribution conservatively assigns lower probabilities to longer words.

Although crude, this model meets the minimal criteria for assigning probabilities to parses: it assigns higher probabilities to parses made up of higher-frequency sub-parts, it assigns an empirically reasonable probability of encountering new words, and it assigns a well-defined probability distribution over new wordforms. Note that assigning a probability that an input sequence contains a new word is not the same as actually learning the novel sequence as a word.

Summary

The preceding subsections described a theory of lexical access, in which input from the prelexical parser is further decomposed into known words or identified as unknown; more specifically, probabilities are assigned to all possible decompositions on the basis of the relative frequencies of their sub-parts.

The ultimate goal is to integrate this theory of lexical access with both the learning theory

39 As discussed in Chapter 4, I assume this value is available to the learner through prosodic cues and/or prior expectations about word length.

of prelexical parsing from the previous chapter, and a theory of word learning. The ideal bootstrapping model would consist of a DiBS parser and a lexicon, both of which develop in response to language input. The DiBS parser might be modeled as a 'mixture' of the phrasal and lexical models, and would correspondingly develop in two ways. First, the phrasal component of the parser should improve its diphone statistics incrementally as more input phrases are encountered. Second, as words are learned, the lexicon component of the parser should improve its diphone statistics, and be correspondingly weighted more strongly. Finally, as more words are learned, the lexical access mechanism should be able to make an increasing contribution in decomposing the input.

However, performing this integration in one fell swoop is vulnerable to problems arising from 'too many moving parts', i.e. the behavior of the model as a whole cannot be understood without understanding the behavior of each of the parts and their interaction. Thus, the next section describes an experiment whose goal is to verify the adequacy of the lexical access theory when it is integrated with a DiBS parser, without the additional complication of word learning, which can be sidestepped by equipping the learner with a full lexicon to begin with. (A full bootstrapping model will be presented later in this chapter.)

Corpus Experiment VII: Verifying the theory of lexical access

Before integrating this theory of lexical access into a full bootstrapping model, it is important to verify that it works in a simpler setting. Thus, this experiment is designed to assess the effect of the lexical access mechanism on word segmentation. The canonical scenario for

models of lexical access is one in which the listener already knows all the words they might encounter (e.g. McClelland & Elman, 1986; Norris & McQueen, 2008). This scenario can be realized here by chaining baseline-DiBS to the lexical access mechanism, i.e. by supplying the partially parsed output of DiBS as input to be further decomposed in lexical access. The results of this, called the *base* condition, will illustrate whether the lexical access mechanism is truly adequate for decomposing speech into words in the best-case, supervised scenario.

It is possible to come a step closer to full boostrapping using phrasal-DiBS instead of baseline-DiBS. In this case, even though the lexicon is fixed, the parser can adapt and change incrementally as it is exposed to more and more language input. The results of this, called the *phrasal* condition, will illustrate whether the success of lexical access depends on the high quality parsing afforded by baseline-DiBS. In other words, if much worse performance is obtained in the phrasal condition than in the base condition, it would be clear that the reason was the poorer quality of prelexical segmentation. Conversely, if comparable decomposition were obtained in the phrasal and base conditions, it would license the conclusion that the lexical access mechanism is not too adversely affected by getting input from phrasal-DiBS instead of baseline-DiBS. (Note that this experiment does not correspond to any natural learning situation, since the parser begins with little language experience, like an infant, whereas the lexicon is adult-like. The ability to conduct such 'thought experiments' is one of the great virtues of model-based research.)

This information will prove invaluable in interpreting the results of the upcoming bootstrapping model. To appreciate this, suppose that the bootstrapping model fails, i.e. exhibits

very poor decomposition and/or by and large fails to learn words to which it was exposed. To what component or interaction should this failure be attributed; and more importantly, what could such a failure tell us about the acquisition of word segmentation? If we can be confident in the lexical access mechanism when it has a sizable lexicon, then such a failure would strongly suggest a failure in the word learning mechanism specifically. If the lexical access mechanism is not tested in this way, then failure of the bootstrapping model would be much harder to interpret. It could arise from a deficiency in the word learning mechanism, or it could result from a deficiency of the lexical access mechanism, or from some unforseen interaction. Such a failure would not be very informative.

Corpora

The phonetic transcription of the BNC was used, as described in previous chapters. The equivalent experiment with the RNC was omitted owing to the computationally-intensive nature of this experiment.[40]

Method

Sample set. Each corpus was divided into samples consisting of an equal number of phrases. The number of phrases per sample was set to 4,000. This number was chosen as a coarse approximation to the amount of speech input that a typical infant might receive in a single day.

40 Specifically, the lexical access algorithm requires a recursive lexical search, with the result that the search time increases super-linearly with the length of the sequence-to-be-decomposed, and exponentially with the number of distinct items in the lexicon. Russian is therefore more computationally-demanding on all fronts. First, Russian words are longer in general. Second, DiBS parses Russian phrases into sequences consisting of more underlying words, because it undersegments Russian more than English. Finally, owing to the rich morphology of Russian, there are many many more distinct types in the RNC (824132) than in the BNC (67034).

The motivation for this sample size is that a typically-developing English infant might be exposed to about 30,000 wds/day (see Chapter 2, Appendix B for the rationale behind this estimate) and phrases (in the BNC) consists of an average of 7.5 words.

Just to be clear, 4,000 phrases/day was selected because it is a *standardized amount of input* in rough correspondence with the amount of input that English infants might hear in a day. This is not a strong claim that infants hear exactly this much input every day, with the implication that 'number of samples' is fully equivalent to number of days of input exposure. In fact, it stands to reason that the exact amount of input that a given infant receives will vary widely according to a host of factors, and will differ across infants and days. In fact, this kind of variation even occurs in the BNC, as the number of words in a phrase may vary throughout the corpus (e.g. as a function of genre), and so the input will exhibit some natural variation in numbers and types of words per sample.

In each condition, the model was first exposed to 180 samples, corresponding to roughly 6 months of language experience. The model was then further trained on an additional 180 samples, corresponding roughly to the second 6 months of language exposure. During this 'second six months', the model was evaluated every 30th sample, i.e. representing one-month samples. Thus, language exposure is indicated in 'months', with the understanding that there is only a rough correspondence between these months and infants' language exposure in the first year.

Parser. The baseline-DiBS parser of Chapter 2 and the phrasal-DiBS parser of Chapter 4

were used.

Lexicon. The lexical access mechanism was equipped with the full lexicon that is observed in the input corpus, including the frequency of each word.

Processing. The phrases in a sample were processed iteratively. Each phrase was stripped of its word boundaries and then passed to the parser, which posited hard word boundaries at the MLDT (.5). This parsed the sequence into several putative words. Each such 'word' was then passed to the lexical access mechanism, which attempted to decompose it into a sequence of items from the lexicon as described in the preceding section of this chapter. Thus, there are three sequences of interest: the original sequence, the prelexically parsed sequence, and the decomposed sequence after lexical access. An example is shown below for illustration, with the corresponding orthographic representation listed on the right:

(14)	original:	h6 d5z It @fEkt ju	'How does it affect you'
	parsed:	h6d5z It@fEktju	'Howdoes itaffectyou'
	decomposed:	h6 d5z It @fEkt ju	'How does it affect you'

Results

The prelexical parser's performance on a sample was evaluated by calculating signal detection statistics on the *parsed* sequences (with reference to the original). The whole system's

performance on a sample was evaluated by calculating the total number of hits, etc.. on the *decomposed* sequences. Thus, for each sample, there are two sets of numbers: the parsing performance when just the prelexical parser has applied, and the further decomposed sequence after lexical access has applied.

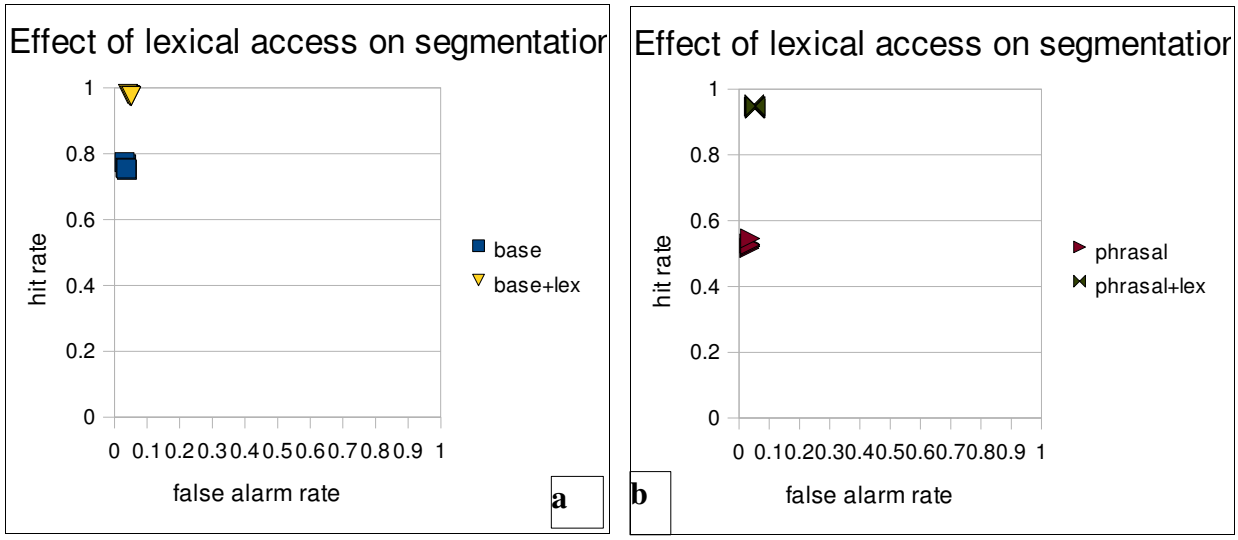


Fig. 5.1: Effect of lexical access on segmentation with (a) baseline-DiBS and (b) phrasal-DiBS

Discussion

The results in the base condition illustrate several facts. Perhaps most importantly, they illustrate that the combined operation of DiBS and the lexical access mechanism result in near-ideal decomposition. That is, not only does it typically correctly decompose sequences such as *howdoes* into their consituent words (raising the hit rate to near-ceiling), it typically correctly doesn't decompose sequences such as *infected* which contain an embedded word (failing to raise false alarm rate). A second point concerns variability among samples. Specifically, the observed

variability is relatively small, suggesting that DiBS is robust to genre variation.

The results in the *phrasal* condition are similarly informative. First, the tight cluster of points corresponding to phrasal-DiBS without lexical access is highly consistent with the segmentation performance yielded by phrasal-DiBS in Chapter 4. This is a highly promising finding, since the parser in Chapter 4's experiment was trained on the entire corpus, whereas the parser here is trained incrementally. This indicates that phrasal-DiBS can achieve its near-ceiling performance with a relatively small amount of training data, as intended; more specifically, it achieves near-ceiling segmentation by the time it has received roughly 6 months' worth of language input. Second, and equally importantly, the effect of lexical access is almost the same in the phrasal condition: near-ideal decomposition is achieved.

Together, these results suggest that the lexical access mechanism is robust to variation in the prelexical parser, as long as the error pattern it exhibits is undersegmentation. This follows from the fact that the degree of undersegmentation is significantly worse in phrasal-DiBS than baseline-DiBS, but this difference had almost no effect on the kind of decomposition that was ultimately achieved. As a result, it is safe to conclude that the lexical access mechanism achieves superior decomposition when equipped with a full lexicon.

Corpus Experiment VII has moved an important step closer to a full bootstrapping model by verifying the efficacy of the lexical access mechanism proposed in this chapter. The results of this experiment show that the lexical access mechanism exhibits superior decomposition in a typical adult lexical access scenario, in which the listener is familiar with all the words they hear. The results further show that the lexical access mechanism is robust to variation in the quality of

its input: as long as the prelexical parser exhibits undersegmentation, nearly-ideal decomposition will be obtained.[41] It only remains to develop a word-learning theory, to which I now turn.

Toward word-learning

For the purposes of this dissertation, word-learning consists of entering a wordform into the mental lexicon, i.e. recognizing that a given form corresponds to a true word. Although there is a vast body of research on word learning (for a recent overview see Hall & Waxman, 2004), much of it focuses on other questions, such as what kinds of meanings infants attribute to novel words (e.g. Booth & Waxman, 2003), what kinds of meanings are easy to discover from contextual cues (e.g. Gillette, Gleitman, Gleitman, Lederer, 1999), and what kind of social support infants receive in learning word meanings (e.g. Baldwin, Markman, Bill, Desjardins, Irwin, & Tidball, 1996). Comparatively little research has focused on word-learning in the sense defined above, which for clarity I will call *wordform learning*. Before describing a theory of wordform learning, I will summarize the relevant research.

Previous research on wordform learning

Much of the available research on wordform learning, as I have defined it, is focused on properties of word learners rather than on formal properties of words themselves. For example, a number of studies have shown that short-term memory and expressive vocabulary contribute independently to predicting word-learning (e.g., Gathercole, Hitch, Service, & Martin, 1997;

41 I assume that the prelexical parser continues to operate in adulthood. In part this is motivated by the more general assumption of a continuity theory of development. More practically, even adults may benefit from prelexical segmentation, for example when they are in a discourse setting such as a classroom in which they encounter many novel words.

Henry & MacClean, 2003; Masoura & Gathercole, 2005; Majerus, Poncelet, van der Linden, & Weekes, 2008).

It is possible or likely that multiple mechanisms mediate these effects. For example, Storkel et al (2006) shows dissociable effects of sublexical phonotactic probability and lexical neighborhood density on new word-learning. This finding is predicted under the two-stage framework assumed here (see Chapter 1 for further explication) with the additional assumption that distinct memory processes are at work in prelexical versus lexical speech processing (e.g. buffering a sublexical phonological representation versus establishing a long-term lexical phonological representation).

A final point of special relevance for infant learning concerns pronunciation variability. Swingley and Aslin (2000) used an eye-tracking paradigm to assess 18-month-olds' interpretation of variants of familiar words which could only be phonemically distinct lexical items for adults, e.g. *baby/vaby*. The results showed both that infants treated the variant as a label for the familiar word (e.g. interpreted *vaby* as referring to the baby), and that they were much slower to do so than with the canonical/correct pronunciation. These results are especially interesting in light of the massive body of research documenting a general shift in phonological development in the second six months of life, whereby infants begin to exhibit language-specific perception of segments (e.g. Werker & Tees, 1984), robustly integrate a variety of prelexical cues for word segmentation (Jusczyk et al, 1999b), and continue to acquire more and more words (Dale & Fenson, 1996). These perception results suggest that infants know all or most of the phonetic contrast system of their language, yet exhibit an incomplete mastery of licit allophonic variation,

e.g. are still willing to treat [v] as an allophone of /b/ word-initially. Similarly, Werker, Fennell, Corcoran, & Stager (2002) found that 20-month-olds but not 14-month-olds succeeded in learning a minimal pair (*bin/din*) whereas even at the younger age infants can learn phonologically dissimilar pairs (*lif/neem* – Stager & Werker, 1997). Together, these results suggest that phonetic and/or phonological similarity presents a special problem for word learning in early word learners. This is an active area of research on which we can expect to learn more in the next few years. Thus, for the time being I turn to my own modeling proposal.

Proposal

As discussed in the introduction to this chapter, word-learning should be triggered by a failure of lexical access. That is, in the course of processing speech input, the learner will first parse her input prelexically; she will then initiate lexical access attempts at the onsets indicated by the prelexical parser. Inevitably, some of these lexical access attempts will fail; specifically, they will fail whenever the learner encounters a word she doesn't know (and may also fail if the prelexical parser has over-segmented, which does occur though quite rarely). These forms, for which the lexical access mechanism fails to match a known word item, are *candidates* for entry into the lexicon. The essence of my word-learning proposal is that learners should add a candidate to their lexicon when they are sufficiently confident that it is a legitimate word. Thus, a word-learning theory should specify how a learner becomes confident that a candidate is truly a word.

As a first, crude pass at this problem, I propose that a candidate wordform is added to the

lexicon if it occurs 10 times or more. To be more precise, the proposal is that every time the lexical access mechanism selects the winning (maximum likelihood) decomposition, it increments the frequency count of every word that is successfully matched, and also increments counts for any novel/non-word candidates in the winning decomposition. When a candidates count reaches 10, it is added to the lexicon. That is, instead of being treated as a nonword for the purposes of lexical access (with a probability determined by the product of the prior probability of a novel word and the lexical phonotactic probability of this new form), the candidate is now treated as a word. Moreover, if the learner's parser depends on the lexicon, the new word now figures into the calculation of DiBS statistics.

This proposal is inadequate in a number of respects. Perhaps most obviously, single-exposure learning is attested in adults (e.g. Storkel et al, 2006). Second, formal properties of the word itself do not directly make any difference to whether the word is learned; the only influence of formal properties is indirect, insofar as they determine how it is segmented. In particular, no memory effects (Masoura & Gathercole, 2005) or effects of confusability with existing words is modeled, whereas English-learning 14-month-olds are known to interpret phonemically distinct forms as realizing a familiar word (e.g. interpret *vaby* as referring to *baby*, Swingley & Aslin, 2000). In addition, phonotactic likelihood and lexical neighborhood density have been shown to be important predictors of word learning (Storkel et al, 2006), which is unpredicted by this proposal.

Faced with a proposal that is inadequate in these ways, it is tempting as a modeler to try to find an alternative that is better somehow. While there are any number of alternative proposals,

it is likely that any simple proposal would suffer from many of the same issues. For example, word learning could be stochastic, i.e. when a learner encounters an unfamiliar wordform they would learn it with some constant probability. This proposal would account for learning with a single exposure, but, like the frequency-threshold proposal, it would not account for how formal properties of a wordform affect its learnability. While it is tempting to try to model these other formal and social factors, this would require a more complex model of word learning, which would entail additional complicating and perhaps unmotivated design choices. The benefit of choosing such a simple deterministic model as this one is that it is very easy to understand exactly how it works in the initial modeling runs. If necessary, it can always be modified later. This is the principal motivation for using such a simple model.

Moreover, despite its shortcoming, the frequency-threshold proposal does capture a number of important properties of word-learning. One such property is that the probability of knowing a word increases with its frequency; this is modeled straightforwardly by learning all words whose exposure frequency exceeds some minimal threshold. Another property is that it is easier to learn words that occur with strong junctural cues (Mattys & Jusczyk, 2001); this property falls out from the parsing mechanism of DiBS, which makes candidates available to be learned precisely because of their junctures. A final property, which could be regarded as the limiting case of the juncture cues case, is that words are easier to learn if they occur at a phrase edge (Dahan & Brent, 1999); this also follows straightforwardly from DiBS since only one boundary needs to be estimated instead of both.

Mixture-DiBS

In the full bootstrapping scenario described above, the infant begins with little language experience and no lexicon. The only DiBS model which is well-defined in this case is phrasal-DiBS. However, after the infant has begun to parse her input prelexically, she should be able to learn some words, which can be modeled with the word-learning proposal described above. At this point, the learner will begin to have access to lexical-DiBS. Note that while lexical-DiBS yields better segmentation than phrasal-DiBS in the best case (Experiment V), it requires a sizable lexicon to achieve its near-ceiling level of parsing (Experiment VI), whereas phrasal-DiBS is at its ceiling before word-learning commences in earnest (Experiment VII). Then the ideal developmental trajectory would be for the learner to rely on phrasal-DiBS initially, and then to gradually shift to relying on lexical-DiBS as more and more words are learned. What is called for is some way of combining the parsing statistics of these two DiBS models to achieve this.

The simplest way to do this is with a linear mixture, that is, a weighted average of the phrasal-DiBS and lexical-DiBS. The resulting parser will be referred to as mixture-DiBS, and has the form:

$$p_{mixture}(\# \mid xy) = (\omega_{phrasa} \cdot p_{phrasal}(\# \mid xy) + \omega_{lexical} \cdot p_{lexical}(\# \mid xy))/(\omega_{phrasal} + \omega_{lexical}) \qquad (5.11)$$

where ω_{phrasa} and $\omega_{lexical}$ are the mixture weights for phrasal-DiBS and lexical-DiBS, respectively. This leaves the question of how to determine the mixture weights.

One appealing option is to weight each parser according to the amount of data that

underlies the statistics in each parser. In the case of phrasal-DiBS, this is proportional to the total number of phrases the infant has experienced (since each phrase contributes once to the left-edge and once to the right-edge distribution). In the case of lexical-DiBS, it is proportional to the total frequency mass of the learner's lexicon (since each token contributes once to the left-edge and once to the right-edge distribution). Thus:

$$\omega_{\text{phrasa}} = \text{number of input phrases}$$

$$\omega_{\text{lexical}} = \sum_{\omega \in \Omega} f(\omega) \tag{5.12}$$

Thus, mixture-DiBS is an incremental learner which is well-defined for any nonzero amount of language input, and any size lexicon (including no lexicon). Moreover, mixture-DiBS initially relies completely on its phrasal component when it has no lexicon, but gradually shifts to its lexical component as more and more words are learned, a property which emerges naturally from weighting by evidence. With all the components in place, it is possible to integrate them into a full bootstrapping model.

Corpus Experiment VIII: Full bootstrapping

Corpus

The phonetic transcription of the BNC was used, as described in previous chapters.

Method

Sample set. The corpus was divided into samples in the same way as in Corpus Experiment VII.

Parser. The parser was mixture-DiBS, as described above. During the just-learning phase, the parser's statistics were updated on the basis of language exposure. During the next phase, two steps occurred after presentation of each learning sample. First, any word candidates which met the word-learning criterion were added to the lexicon. Second, the parser's statistics were updated according to the additional input (the sample) and any additional words that were learned.

Results

As in the previous experiment, the prelexical parsing and decomposition after lexical access are plotted on an ROC curve.

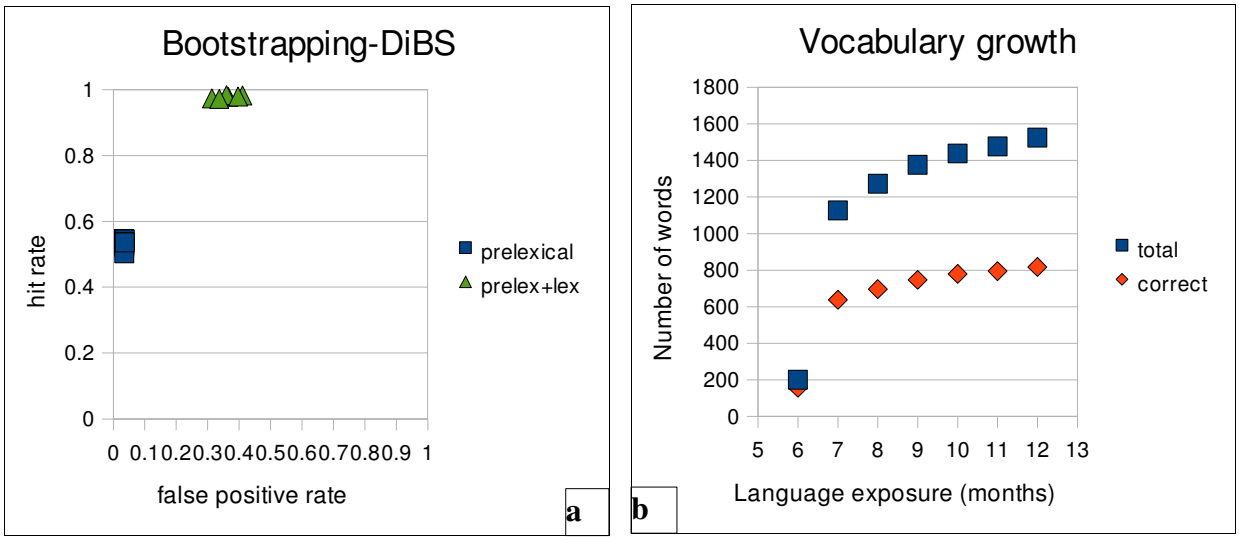


Fig 5.2: (a) Segmentation ROC of bootstrapping model and (b) its vocabulary growth.

Discussion

Two facts are evident from Fig 5.3. First, the system as a whole exhibits rather aggressive oversegmentation, as evident from the false alarm rate of about 40%. Second, the prelexical segmentation mechanism alone exhibits undersegmentation.

By itself, the first fact is not necessarily troubling. It is possible that the lexical access mechanism is discovering meaningful sublexical units, i.e. affixes and stems. Since these items have their own meaning (even if non-compositional), it is a reasonable outcome if the system as a whole ends up positing them as separate lexical units. However, it is also possible that the system is oversegmenting in a different way; i.e. non-meaningful units. Note that the CELEX lexicon on which the phonetic transcript of the BNC was built does not contain a morphemic decomposition of its words. Thus, it is not possible to use this resource to determine what percentage of 'words' found by the bootstrapping model are meaningful units. However, some insight on this question can be gleaned by inspecting the parsing output of the system, and the high-frequency words it learns.

Parsing output. The end-behavior of the system is illustrated below with four phrases taken from the final testing sample. The top line gives the underlying orthographic sequence and the next line (original) gives the phonetic transcription with the correct word boundaries indicated as spaces. The next line (parsed) indicates the word boundaries identified by the prelexical parser. The line below that (decomposed) indicates the word sequence identified by the

lexical access mechanism. The final (ortho) line indicates an orthographic transcription of the decomposed line:

(17) ortho: *A commitment to an economic transformation with several components*
original: 1 k@mItm@nt tu {n ik@nQmIk tr{nsf@m1SH wID sEvr@l k@mp5n@nts
parsed: 1k@mIt m@nt tu {nik@nQmIktr{ns f@m1SH wI DsEvr@l k@mp5n@nts
decomposed: 1 k@mIt m@nt tu {n ik@nQmIk tr{ns f@m1SH wI DsEvr@l k@m p5 n@nt s
ortho: *a commit ment to an economic trans formation wi thseveral com po nent s*

(18) ortho: *government was best*
original: gVvHm@nt wQz bEst
parsed: gVvH m@nt wQz bEst
decomposed: gVvH m@nt wQz bEst
ortho: *govern ment was best*

(19) ortho: *two years earlier*
original: tu j7z 3l7R
parsed: tuj7z3l7R
decomposed: tu j7z 3l7R
ortho: *two years earlier*

(20) ortho: *on a more limited franchise*
original: Qn 1 m$R lImItId fr{nJ2z
parsed: Qn1m$R lImItId fr{nJ2z
decomposed: Qn 1 m$R lI mItId fr{nJ2z

ortho: *on a more li mited franchise*

As evident from these phrases, many of the sublexical units that the system discovers are indeed meaningful. For example, the system has correctly identified the words *a*, *to*, *an*, *economic*, *was*, *best*, *two*, *years*, *earlier*, *on*, *more*, and *franchise*. In addition it has identified *-ment*, which is a meaningful affix, and has segmented it off from *commit* and *govern*, both of which can occur as separate words without *-ment*; similarly it has segmented *trans-* off from *formation*, which can occur as a separate word. So the system has considerable success in identifying the meaningful units in this sample.

At the same time, the system exhibits some notable failures. In particular, it decomposes *components* into *com+po+nent+s*. The affixes *com-* and *-s* are meaningful affixes in English, but neither *po* nor *nent* are. Similarly, the system incorrectly decomposes *limited* into *li* and *mited*. The former item is recognizable as the adverbial ending *-ly* (which is realized with the lax vowel [ɪ] in British RP, the phonetic standard for CELEX), but *mited* is certainly not a meaningful word in English. Tentatively, then, it can be concluded that the model is too aggressive in segmentation, which results in learning not only linguistically meaningful sublexical units (*-ment*, *-ly*, *trans-*, *-s*), but also linguistically unprincipled sublexical items (*-po-*, *-mited*).

This conclusion is all the more interesting given the evidence that the prelexical segmentation mechanism undersegments. One potential concern with the bootstrapping model is an 'error snowball' in which some oversegmentation errors cause sublexical units to be learned as words, which alter the statistical signature of word boundaries in the lexicon, thereby causing the prelexical parsing mechanism to oversegment more. This appears not to happen *even though* the

system learns a number of sublexical units as words. In other words, the results of Experiment VIII suggest that the prelexical parsing mechanism is robust against this kind of error snowball. Rather, oversegmentation appears to be driven entirely by the lexical access mechanism – more specifically, it is overly aggressive in learning sublexical items as words. One simple way to see which sublexical items are learned is to investigate the highest-frequency 'words' the system learns.

Lexicon. The 100 most frequent lexical items acquired by this system (at the end of the final sample) are given below:

word	freq	word	freq	word	freq	word	freq	word	freq
s	489924	S	136341	{n	57652	ri	30045	h{d	23691
l	451967	IN	131304	D{t	57415	#R	29954	n5	23559
k	402528	v	123315	g	56896	hIz	29220	r@	23410
Di	377412	Iz	112208	bi	56644	s5	29109	wIJ	22447
d	371569	i	102174	@nt	55826	Ent	28746	wi	22207
t	369396	rI	94048	f$R	54756	@d	27948	@rI	21920
p	324217	_	92671	J	53427	D1	27931	6t	21429
1	314078	$	90936	wQz	50003	h{v	27833	7	21352
m	300372	It	90056	$I	48948	@s	27674	EI	20845
I	241273	H	88031	{t	46817	frQm	27247	w3R	20550
In	231210	@R	87345	@t	46584	nQt	27200	w2	20077
n	213113	ju	81725	@m	44767	Es	26787	I2	19905
z	201908	II	80682	{z	41194	bI	26550	Its	19441
Qv	185551	D	79882	$R	40487	@ns	26400	Id	18945
{nd	172134	5	78667	@z	39304	wVn	26043	@II	18871
f	168578	P	73101	b2	39015	u	25985	h3R	18785
tu	165186	b	71658	If	35156	hi	25391	r1	18599
2	160120	wI	65126	En	34779	bVt	25309	Em	18376
@	153178	Qn	64199	T	34360	Vn	25015	r2	17541
st	142035	@n	60477	#	31425	DIs	24615	r5	17329

Table 5.1: The 100 most frequent lexical items learned by the bootstrapping model

The first several 'words' include [s], [t], [d], [Di], and [In], which are recognizable as allomorphs of the plural/possessive marker *-s*, the past tense marker *-ed*, the definite article *the*, and the preposition *in*. In addition, the most frequent 'words' include a number of functional items including *a*, *to*, *of*, *and*, and *-ing*. Similarly, the items [l], [v]. and [m] are recognizable as licit contractions following *I* and some other pronouns (*I'll*, *I've*, *I'm*). On the basis of these items, the system must be counted as a partial success for identifying meaningful elements.

However, there are a number of items which clearly do not correspond to meaningful units. For example, [k], [f], and [n] in the first column are not words or other meaningful units in English. Similarly, in other columns [g], the voiced palato-alveolar affricate (transcribed [_]), [b], and the voiced inter-dental fricative [D] are not words or other meaningful units of English. It is evident from these 'words' why the system oversegments – there are too many single-consonant 'words'.

The results of Experiment VII are useful for interpreting this result. Recall that in Experiment VII, in which the learner was supplied with the correct lexicon to begin with, nearly perfect segmentation was achieved, regardless of whether the prelexical parser was phrasal-DiBS or baseline-DiBS. This suggests that *when* the lexical access mechanism is equipped with the proper lexicon, it functions properly (as long as the prelexical parser undersegments). The implication for the present case is that the lexical access mechanism is oversegmenting because too many improper/sublexical units have been admitted into the lexicon. The solution, then, is to somehow block sublexical units from being admitted to the lexicon.

As a first, crude pass at this problem, I implemented a single word-learning constraint: a lexical candidate must contain a vowel to be entered into the lexicon. The results from running this adjusted bootstrapping model are described below in Corpus Experiment IX.

Corpus Experiment IX: Full bootstrapping with word-learning constraint

Corpus

The phonetic transcription of the BNC was used, as described in previous chapters.

Methods

The method is identical to Corpus Experiment VII, except that a constraint was added to the word-learning mechanism: words could only be learned if they contained a vowel.

Results

As in the previous experiment, the prelexical parsing and decomposition after lexical access are plotted on an ROC curve. Beside this, the total vocabulary size is plotted along with the number of items which correspond to wordforms in the original corpus.

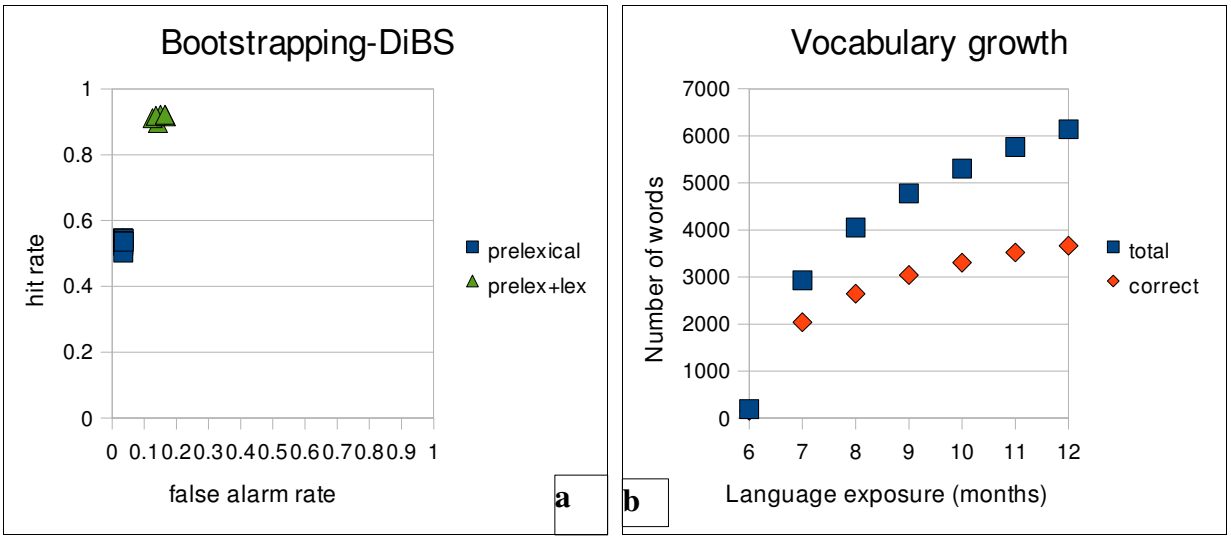


Fig. 5.3: (a) Segmentation ROC of bootstrapping model w/ vowel constraint and (b) its vocabulary growth

Lexicon. As in the previous experiment, the 100 most frequent lexical items learned by the system are given below.

word	freq	word	freq	word	freq	word	freq	word	freq
Di	337536	{n	40635	sI	25205	@z	16477	m$R	13126
I	173645	bi	40471	p@	24787	hu	16373	kVm	13045
In	168724	{t	39100	DIs	24167	2m	16200	l2	12375
1	162812	b2	38883	k@n	24146	tId	16023	k@m	12331
{nd	162196	{z	38368	$R	23779	wUd	15766	nju	12157
tu	132041	$I	36060	h{d	23663	bin	15706	Si	12079
Qv	129610	$	35415	n5	23652	mi	15662	nI	11954
@	107960	Qn	30816	5	22894	h{z	15620	@d	11562
Iz	85042	@R	29596	wIJ	22306	wII	15171	d1	11487
IN	78097	s5	28289	wi	22159	sVm	15123	h6	11241
II	72947	hIz	28057	si	21059	2d	14842	D5	11106
2	72603	D1	28046	w3R	20510	If	14805	wQt	11093
wI	62126	#R	27839	m@nt	18955	Dis	14760	hIm	11035
rI	60679	h{v	27700	{k	18458	k{n	14563	Vp	10974
It	59952	nQt	27110	dIs	18240	Its	14463	kQn	10915
D{t	55159	bVt	27089	h3R	17920	w2	14444	l2k	10907
f$R	53106	frQm	26326	tI	17805	m2	14440	Vn	10888
wQz	49778	bI	26132	m1	17037	6t	13725	dId	10775
ju	45297	hi	26094	D8R	16602	w1	13391	mEn	10726
dI	42473	wVn	26002	wEn	16491	D8	13329	D{n	10587

Table 5.2: The 100 most frequent lexical items learned by the bootstrapping model with vowel constraint

In comparison with the previous experiment, the most frequent 'words' learned by the bootstrapping model with vowel constraint are much more satisfactory. Of the ultra-high frequency items in the first columns, all are clearly recognizable as meaningful units of English except [@] and [I], and an argument could be made for the latter as an adjectival/diminutive marker (*happy*, *pretty*, *ready*, *Bobby*, *Johnny*, *Kathy*, *Christy*). For clarity, the units of the first column are given in conventional orthography:

(21) [Di] *the* [In] *in* [1] *a* [{nd] *and*

[tu]	*to*	[Qv]	*of*	[Iz]	*is*	[IN]	*-ing*
[lI]	*-ly*	[2]	*I*	[wI]	*we*	[rI]	*re-*
[It]	*it*	[D{t]	*that*	[f$R]	*for*	[wQz]	*was*
[ju]	*you*	[di]	*de-*				

Interestingly, but perhaps not surprisingly, all of these ultra-frequent items are function words (prepositions, determiners, copula, pronouns) or function morphemes (adverbial marker, progressive marker) except *re-* and *de-*. This is perhaps unsurprising because these are some of the most high-frequency items of the language; if the system is learning words at all, these are some of the ones it will get the most exposure to. It is nonetheless an interesting finding that the system identifies so many functional items, since many these items are systematically absent in infant productions before the age of 2 (Guasti, 2004), which has led some researchers to posit that infants lack any formal representation of functional items (e.g. Wexler, 1994), although a growing body of evidence suggests that infants are sensitive to functional items in *perception* well before they begin to produce them (Shady & Gerken, 1999; Mintz, 2003; Peterson-Hicks, 2006; Shi, Cutler, Werker, & Cruickshank, 2006; Shi, Werker, & Cutler, 2006; Soderstrom, White, Conwell, & Morgan, 2007). The fact that this system identifies so many functional units highlights the importance of further research on infants' knowledge of functional morphemes.

Some further insight on the system's behavior can be gleaned by inspecting words from the *middle* of the frequency spectrum. The words that the system learns with recognized frequencies of 98 or 99 are given below in (22) (a * is used to indicate a non-word, and in cases

where the non-word is a recognizable sub-part of a real word, the 'missing' letters are supplied in parentheses):

(22)

[strVglIN]	*struggling*	[rEk@n]	*reckon*	[kiz]	*keys*
[k3rIN]	*(han)kering	[h#vIst]	*harvest*	[frVstr1SH]	*frustration*
[_3n2lIst]	*journalist*	[DI]	*thi(s)	[wEdIN]	*wedding*
[spQndtu]	(re)*spond to	[sp@d]	*sped	[sk$t]	(e)*scort
[s3kPz]	*circles*	[p{sIv]	*passive	[p2ld]	*piled*
[ndIt]	(a)*nd it	[kl{m]	*clam	[gQTIk]	*gothic*
[dt@w$dz]	*d towards	[bl2_d]	(o)*bliged	[JI]	*chi
[pQpjUl{r@tI]	*popularity*	[s3v@tIvp#tI]	(con)*servative party		
[plEks@tI]	(com)*plexity				

A reasonable proportion of these items are words, and many of the errors bear morphologically transparent relationships to words. There are nonetheless phonotactically implausible items such as the sequence [dt@w$dz] which contains an illegal onset cluster.

Discussion

Comparison of Fig 5.3a with Fig 5.4a shows a significant reduction in oversegmentation. Whereas the system in Experiment VIII exhibited a false alarm rate of around 40%, the false alarm rate in this experiment is a more modest 15%. Thus, the vowel constraint is substantially

though not completely successful in blocking the aggressive oversegmentation apparent in Experiment VIII. This fact supports the interpretation that oversegmentation was caused by the entrenchment of single-consonant 'words' in the previous experiment.

Comparison of Fig 5.3b with Fig 5.4b shows another, perhaps more interesting outcome: both the total number of words learned and the number of correct words are greater than in Experiment VIII. This is a highly counterintuitive finding because relative to the previous experiment, *there are stronger constraints on learning a word, and yet more words are learned.* I postpone further consideration of this for the general discussion.

Size of lexical inventory. One concern arising from these results is the relatively large number of words that the system learns, as compared with what infants are purported to know. As reviewed in Chapter 1, mother's reports suggest that English-learning infants' receptive vocabulary contains on average about 40 words at 8 months, about 80 words at 12 months, and about 180 words at 18 months. In contrast, the system here is learning on the order of 6000 words. With respect to this disparity, two remarks can be made.

First, the total vocabulary size that the system acquires can be adjusted by tweaking the frequency threshold. For Experiments VIII and IX an arbitrary threshold of 10 was imposed, based loosely on the fact that word-learning is evident in laboratory studies with about 10 exposures (e.g. 14-month-olds in Booth & Waxman, 2003). It is possible – indeed, it is likely – that infants' word-learning ability increases developmentally, so that 6-month-olds require more exposures to learn words than 14-month-olds. Of course, word-learning is not based only on

number of exposures, so the theory is making a relatively crude simplification to begin with.

Second, one of the reasons the vocabulary size is so high in these results is because the infants' lexicon includes a large number of function words and morphemes. These items will not show up in estimates of (English-learning) infants' vocabularies, for the simple reason that many function words (e.g. determiners, adverbs, *he*, *we*) and all bound affixes (*-ing*, *-ed*, *re-*, *-en*, *-s*) are not listed on the MacArthur CDI forms (Dale & Fenson, 1996) which are standardly used to assess infant vocabularies. Moreover, caregivers who fill out these forms are instructed to report if their child understands the *meaning* of the word; as remarked repeatedly throughout this dissertation, it is possible for an infant to know that a form is a word without understanding its meaning. In other words, the results of this experiment could be interpreted as suggesting that infants 'know' significantly more wordforms than contemporary vocabulary measures would suggest.

General discussion

Summary

This dissertation began by introducing three related acquisition problems, of *word segmentation*, *word recognition*, and *word learning*. Previous chapters have focused on the word segmentation problem specifically, developing and testing a learning theory for prelexical word segmentation. The goal of the present chapter was to broaden this research to address the other two problems of word recognition and word learning.

This chapter began by arguing that under the assumptions adopted here, word learning

must be mediated by the lexical access mechanism. Then, a theory of lexical access was developed in which a probability distribution is assigned over exhaustive lexical decompositions of an input sequence; for evaluation purposes, only the maximum likelihood decomposition is used. This theory was tested by chaining DiBS to the lexical access mechanism, i.e. feeding the output of the prelexical DiBS parser into the lexical access mechanism for further decomposition. In Experiment VII, the lexicon was initialized to the full/correct lexicon; the results showed near-ceiling decomposition. This means that the lexical access mechanism correctly identifies all and only word boundaries that were missed by the prelexical parser; in particular, it does not systematically decompose sequences like *infallible* into *in* + *fallible*, in which each of the constituents can itself occur as a word. These results show that the theory of lexical access developed here is adequate as long as the lexicon itself is appropriate.

The remaining sections were devoted to the word-learning problem and a full bootstrapping model. As a first pass, Experiment VIII implemented a crude word learning mechanism in which a 'candidate' wordform was added to the lexicon once the lexical access mechanism had attempted to access it 10 times. Experiment IX, the final experiment in this dissertation, was identical to Experiment VIII, except with the added word-learning constraint that a wordform could not be added to the lexicon unless it contained a vowel. While these experiments are not strictly cognitively plausible, they are of considerable utility in illustrating where the system succeeds, where it fails, and what must ultimately be done to achieve a better model.

In the remaining subsections, I consider these issues in greater depth. First, I explore a

highly counterintuive difference between the results of Experiment VIII and Experiment IX. Specifically, even though there were objectively stronger constraints on word-learning Experiment IX, the result was that a greater number of words were learned overall. After considering this results in some detail, I outline some options for how to modify the lexical access theory to address the shortcomings discovered here; for example, how to handle affixes. Finally, I argue that this research highlights the need for a richer understanding of how listeners learn word*forms* specifically. Each of these are considered in turn below.

Effect of vowel constraint on word-learning

One highly counterintuitive finding of this chapter is more words were learned in Experiment IX than in Experiment VIII, although there were stronger constraints on word-learning in Experiment IX. The goal of this subsection is to explain this finding.

Perhaps the most important part of the explanation is the absence of morphological structure in the lexical access mechanism. More specifically, the unigram model adopted here means that the lexical access mechanism lacks the means to represent any relationships between lexical items, such as morphosyntactic dependencies between 'heads' (stems, roots) and affixes. In fact, the model does not even distinguish these as distinct categories. Items are simply in the lexicon (with some frequency), or not. As illustrated by Experiment VII, this is not a problem if the lexicon contains the 'right' words to begin with. But as illustrated in Table 5.1, the unconstrained system of Experiment VIII isolates single-segment sublexical units like [s] as possible words. In one way, this is a positive outcome, since [s] does indeed correspond to a

meaningful unit (the plural/possessive marker) in English. The problem with the current system is that when [s] is entered into the lexicon, there is nothing that marks it as appropriate for word-final position only. In other words, it becomes possible to (incorrectly) decompose an [s] in other word positions as well.

To see why this is a problem in Experiment VIII, suppose that the system has identified [s] as a word and now it encounters the novel word *struck* [strVk], and let us further suppose the learner has not yet learned the word *truck*. According to the recursive lexical decomposition process proposed earlier in this chapter, the lexical access mechanism should consider two parses for this sequence. One parse consists of the whole form (residue) since the lexicon does not have a lexical entry for it already. The other parse, beginning from the onset of the word, identifies the unit [s] as one part of the decomposition and, failing to identify any subsequences matching the residue *-truck*, posits the remaining residue *-truck* as a lexical item:

(22) decomposition 1: w = [strVk]

decomposition 2: w = [s] + [trVk]

The first decomposition will be assigned a probability with terms for the probability of a new word, the and the lexical phonotactic probability of this new word *struck*. The second decomposition will be assigned a probability with terms for the familiar unit [s], as well as the probability of a new word and the lexical phonotactic probability of the new word *truck*:

$$p([strVk]) = p_{\nu} \cdot p(\omega_{\nu} = [strVk])$$

$$= p_{\nu} \cdot (1-p_{\#})^{4} p_{\#} \cdot (1/|\Phi|)^{5}$$

$$p([s] + [trVk]) = p([s]) \cdot p_{\nu} \cdot p(\omega_{\nu} = [trVk])$$

$$= f([s])/F \cdot p_{\nu} \cdot (1-p_{\#})^{3} p_{\#} \cdot (1/|\Phi|)^{4} \quad (5.13)$$

where F is the frequency mass of the lexicon and $|\Phi|$ is the total number of phones in the language. Which of these is the maximum likelihood parse can be determined by taking the likelihood ratio:

$$p([s] + [trVk])/p([strVk]) = (f([s])/F \cdot p_{\nu} \cdot (1-p_{\#})^{3} p_{\#} \cdot (1/|\Phi|)^{4}) / (p_{\nu} \cdot (1-p_{\#})^{4} p_{\#} \cdot (1/|\Phi|)^{5})$$

$$= f([s])/F \cdot |\Phi|/(1-p_{\#})$$

$$= 802394/24344258 \cdot 51/(1-.2618)$$

$$= 2.277 \quad (5.14)$$

where the specific values for f([s]), F, $|\Phi|$, and $(1-p_{\#})$ are taken from the end-state of Experiment VIII. This ratio indicates that the decompositional parse p([s] + [trVk]) is almost 3 times as likely as the wholistic parse p([strVk]), so the more decomposed parse wins, re-inforcing the likelihood of [s] to be decomposed in similar situations.

If the system already knows the word *truck*, the situation is similar, except that the more decomposed parse is assigned even higher relative probability. This follows from the fact that the

probability of a word $p(\omega) = f(\omega)/F$ is almost inevitably higher than the probability of encountering this same wordform as a new word $p_\nu \cdot (1-p_\#)^{n-1} p_\# \cdot (1/|\Phi|)^n$, that is $p(\textit{truck}) > p_\nu \cdot p(\omega_\nu = \textit{truck})$.

In summary, it is not necessarily a problem by itself that the system identifies [s] as a meaningful unit. However, given that it identifies [s] as a meaningful unit, the problem is that there is no morphological constraint which ensures that it is only identified in lexically appropriate positions. The result in Experiment VIII is that a number of single-segment 'words' are learned; owing to their great frequency they become entrenched, so that any *novel* words that begin with these segments are decomposed into them and a sublexical residue. This explains the 'plateau' in word-learning in Experiment VIII – many or most of the word tokens are decomposed into small/phonological units (e.g. single segments), of which there are such a limited number that the peak is eventually reached.

In contrast, the vowel constraint of Experiment IX prevents these single-consonant units from being identified as words. As a result, the single-consonant words cannot become entrenched. When novel words are encountered, there is less opportunity to overdecompose them, and the system learns units that more closely resemble words according to the underlying corpus. The vowel constraint actually promotes learning by forcing the system to not overdecompose, i.e. forcing it to learn larger-sized units, which of course there are more of. This is why the system learns more words in Experiment IX even though there are stronger constraints against word-learning.

Toward a better theory

Although the bootstrapping experiments in this chapter exhibited a number of notable successes, they also exhibited some failures. In particular, the results of Experiment VII suggest that when the learner is equipped with the 'correct' lexicon to begin with, the theory of lexical access developed here achieves superior decomposition. In contrast, the comparison between Experiments VIII and IX suggests that the lexical access mechanism is underconstrained. As I suggested in the previous subsection, one underlying problem is an impoverished morphological representation. The lexicon/lexical access mechanism adopted here does not distinguish affixes from heads, and therefore is incapable of representing relationships between them. In other words, the solution is to adopt a richer view of the lexicon.

Perhaps one of the best ways to model this richer structure, I will suggest, is to incorporate a richer model of prosodic structure. Through this dissertation, I have attempted to develop a probabilistic formalism which can readily be extended to richer structures. In particular, the core domain considered in this dissertation is sequences of two phones (diphones), but most of the key aspects of the model can work on arbitrary domains; I focused on the phone domain because the evidence that infants attend to it is particularly strong (Mattys & Jusczyk, 2001).

In fact, I find it remarkable that the model is able to succeed as well as it does without explicitly modeling lexical stress, syllabification, or prominence. The acquisition literature suggests that at least for English, stress cues are used for segmentation even before other phonotactic cues (Jusczyk, Houston, & Newsome, 1999). Moreover, native-language

syllabification effects are evident even in highly proficient non-native speakers (Dupoux et al, 1999), which suggests or at least is consistent with the hypothesis that syllabification is perceptually entrenched extremely early in development, perhaps before or during the rapid phonological development of the second six months of life (see Chapter 1 for a review).

It is likely that the prelexical parser's performance could be substantially improved by modeling the full prosodic hierarchy (Selkirk, 1984) rather than simply word boundaries. In turn, the richer structure at the prelexical level can only improve parsing at the lexical level, owing to the tight confluences between phonological and morphological junctures (Hay, 2003; Hay & Baayen, 2002; Pierrehumbert, 2003). Similarly, a richer model of prosodic structure should result in the opportunity for a more appropriately constrained theory of word-learning. For example, a sufficiently rich prelexical representation of the prosodic word (Selkirk, 1984) may facilitate distinguishing content words from function words in the model's input.

Conclusion

This chapter has outlined a theory of lexical access and a theory of word-learning, and it has integrated them with prelexical parsing models developed in earlier chapters to make a first pass at a full bootstrapping model. The results of the bootstrapping model reveal a pattern of successes and failures that is highly informative for further research. Specifically, the results suggest that prelexical parsing *a la* DiBS is remarkably effective, but that lexical access and word-learning require a richer theory of morphological structure. Put more generally, it could be said that this dissertation has pushed a relatively naïve statistical approach as far as it will go: its

more or less sufficient for prelexical parsing, but more sophisticated morphology is required for lexical access and/or word-learning.

CHAPTER 6. CONCLUSIONS

Abstract

This chapter begins by recapitulating the problem of the infant word segmentation acquisition, and DiBS as a solution in a two-stage speech processing framework. Next, it reviews the major accomplishments of each chapter in fleshing out and testing the proposal. Finally, it reviews some of the open questions and issues of this research, and future directions.

Words are fundamental elements of language. In order to make use of words for language comprehension and acquisition, infants must first be able to pick out one word from another as they occur in fluent speech – the problem of word segmentation. In adults, this process can plausibly be explained largely as an epiphenomenon of word recognition (McClelland & Elman, 1986; Norris & McQueen, 2008). However, infants between 6 and 10.5 months of age do not know many of the words that they hear (Dale & Fenson, 1996; van de Weijer, 1998; Brent & Siskind, 2001), whereas by this age they already demonstrate impressive segmentation abilities in laboratory studies (Saffran et al, 1996; Jusczyk et al, 1999a; Jusczyk et al, 1999b; Mattys et al, 1999; Mattys & Jusczyk, 2001; Johnson & Jusczyk, 2001; Bortfeld et al, 2005). Thus, infants must have access to some other speech segmentation mechanism besides word recognition, based on prelexical cues such as phonotactics (Mattys & Jusczyk, 2001).

This fact, coupled with evidence of dissociations between sublexical and lexical factors in speech processing (Vitevitch et al, 1997; Luce & Pisoni, 1998; Vitevitch & Luce, 1998; Vitevitch & Luce, 2004), supports a *two-stage* view of speech processing in which speech input is assigned

a phonological parse by a *prelexical parser* (or "Fast Phonological Preprocessor") (Pierrehumbert, 2001; Hay, 2003) and then further decomposed during *lexical access*:

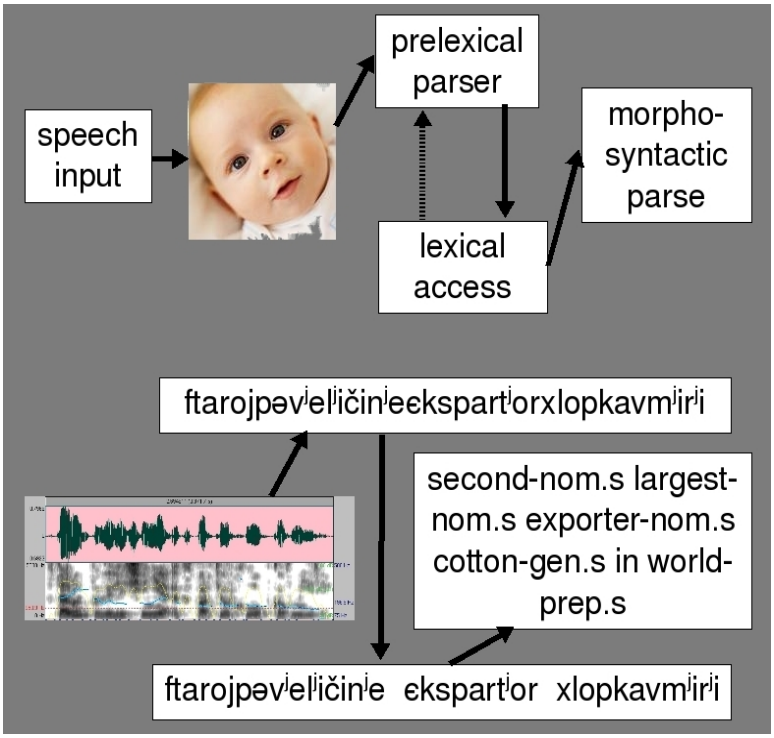


Fig 6.1: Two-stage speech processing framework

The bidirectional links between the prelexical parser and the lexicon represent the interdependency between these components of speech processing. In particular, the prelexical parser should affect the lexicon by isolating novel wordforms for lexical acquisition (Davis, 2004), and the lexicon should project lexical phonotactics onto the prelexical parser for more accurate phonological parsing, including word segmentation (Pierrehumbert, 2001; Pierehumbert, 2003).

This dissertation fleshes out this general framework by making and testing a proposal as to the specific formal representations, knowledge content, and processing mechanisms of the prelexical parser. DiBS – **Di**phone **B**ased **S**egmentation – is predicated on the idea that the presence/absence of a word boundary can be predicted by the immediate phonological context, and more specifically by the identity of the phones immediately preceding and following the possible boundary. Thus, DiBS assumes that the listener has access to a categorical representation, e.g. phones, of the speech sounds in their speech input. It proposes that listeners estimate the probability of a word boundary between every pair of phones that occur in the language, i.e. for every sequence [xy] the listener calculates the probability of a word boundary given [xy] p(# | xy). Finally, this knowledge is put to use in online speech processing by initiating lexical access attempts at all locations in the speech stream with a high word boundary probability.

The main body of this dissertation is devoted to working out the concrete details of this proposal more specifically and testing it. To this end, a sequence of models is developed, including baseline-DiBS, lexical-DiBS, and phrasal-DiBS. The first model, baseline-DiBS, is a supervised learning model and therefore not appropriate for modeling (unsupervised) infant acquisition; rather, it is an important proof-of-concept which was used to illustrate the comparatively high level of segmentation that can be achieved by DiBS in the best case. The next two models, lexical-DiBS and phrasal-DiBS, address the learnability issue by showing how to estimate the DiBS statistics p(# | xy) using Bayes' Theorem from information that is or can reasonably be assumed to be available to infants, including any words in the infant's lexicon and

the distribution of speech sounds at phrase-edges. In Chapters 2-4, these models are tested on phonetic transcripts derived from large corpora in English (the British National Corpus) and Russian (the Russian National Corpus). Thus, the greatest focus of this dissertation is on the prelexical parser specifically.

Ultimately, it is not just the acquisition of segmentation which must be explained, but the interlocking relationship between word segmentation, word recognition, and word learning. Thus in Chapter 5, models of lexical access and word learning are developed. These models are integrated with the final DiBS model, mixture-DiBS, to form a full bootstrapping model. Mixture-DiBS combines phrasal- and lexical-DiBS to model the changing effect of the lexicon on the prelexical parser as more words are learned.

Summary of acquisition problem

Before considering this work in more detail, it is worth reviewing the motivation for it, i.e. the infant's problem. Speech processing in infants is characterized by the following properties:

1. Phrase-internal word boundaries are not consistently marked acoustically (Lehiste, 1960)
2. On average mothers report their infants know 40 words at 8 months of age and 80 words at 12 months of age (Dale & Fenson, 1996)
3. During this period infants can segment unfamiliar words from fluent speech (Saffran et al, 1996; Jusczyk et al, 1999a; Jusczyk et al, 1999b; Mattys et al., 1999; Mattys &

Jusczyk, 2001; Johnson & Jusczyk, 2001; Bortfeld et al, 2005)

4. Only 10% of word types that infants hear are presented in isolation (Aslin et al, 1996; van de Weijer, 1998; Brent & Siskind, 2001)

The first property is why there is a word segmentation problem at all; adults face this problem just as infants do. However, unlike infants, adults know most of the words they encounter. Thus, word segmentation in adults can plausibly be explained by word recognition, as in the TRACE model of speech perception (McClelland & Elman, 1986) and related models (Norris & McQueen, 2008). Owing to the second property, this kind of account is inadequate to explain word segmentation, as infants appear not to know most of the words they hear. It has been argued (Dahan & Brent, 1999; Brent & Siskind, 2001) that infants must hear a word in isolation or adjacent to a familiar word to learn it, a position that is hard to reconcile with the significant body of literature demonstrating segmentation of novel words from novel contexts (e.g. Saffran et al, 1996; Mattys & Jusczyk, 2001). The fact that infants are only exposed to 10% of novel word types in isolation, and yet end up learning most of the words they encounter, can only be explained if infants learn words by segmenting them from fluent speech.

In other words, infants face a special speech processing problem because they don't know very many words. Without a doubt, word recognition facilitates word segmentation (McClelland & Elman, 1986; Bortfeld et al, 2005; Norris & McQueen, 2008), but in order to make use of this segmentation mechanism, infants must first learn the word they are recognizing: word recognition requires word learning. If it were the case that infants were exposed to many or most

words in isolation, it would be easy to explain word learning without word segmentation. However, as shown by both corpus (van de Wiejer, 1998; Brent & Siskind, 2001) and laboratory studies (Aslin et al, 1996), infants do not hear most word types in isolation. The fact that they learn them means that they must have segmented them out of fluent speech. In other words, word segmentation precedes and supports word learning (Davis, 2004) rather than the other way around. Or, more precisely, the earliest word segmentation appears to precede and supports the earliest word learning, though once word learning has begun in earnest it can support and improve word segmentation as well.

Thus, of these three interlocking speech problems, the evidence strongly suggests that infants make the greatest progress on word segmentation first, presumably by exploiting prelexical properties including stress (Jusczyk et al, 1999b) and segmental phonotactics (Mattys & Jusczyk, 2001). These facts suggest that word segmentation, or at least infant word segmentation, is a logically distinct problem from word learning. More broadly, these findings support a two-stage theory of speech processing which involve distinct sublexical and lexical representations, a distinction which is supported in current theories (Pierrehumbert, 2001; Vitevitch & Luce, 1998) on the basis of dissociations between sublexical phonotactic likelihood and lexical neighborhood density on measures such as object naming, word recall, nonword repetition, lexical access, and word learning (Frisch, Large, & Pisoni, 2000; Luce & Large, 2001; Luce & Pisoni, 1998; Storkel et al, 2006; Thorn & Frankish, 2005; Vitevitch, 1997; Vitevitch, 2002a; Vitevitch, 2002b; Vitevitch et al, 2004; Vitevitch & Luce, 1998, Vitevitch & Luce, 1999).

Proposal

On the basis of these facts, this dissertation investigates the proposal that infants' segmentation of speech arises in a prelexical parser (cf. Pierrehumbert, 2001; Hay, 2003). More specifically, it tests the proposal that infants recover word boundaries based on the identity of the surrounding diphone, hence the name DiBS – **Di**phone-**B**ased **S**egmentation. The use of diphones – sequences of two speech sounds such as [ba] and [pd] – is motivated by two facts. First, as reviewed in Chapter 1, there is clear psycholinguistic evidence that infants do attend to diphones (Mattys et al., 1999) and exploit them for word segmentation (Mattys & Jusczyk, 2001). Second, Hockema (2006) shows that diphones in English exhibit a bimodal distribution, whereby the majority of diphones occur almost exclusively word-internally, or almost exclusively across word boundaries. The clear implication of this result is that an infant could achieve relatively good segmentation simply by exploiting this cue, as already suggested by the results of Cairns et al (1997). Thus, DiBS is in clear contrast with earlier work on word segmentation.

One of the earliest proposals comes from Saffran and colleagues (Saffran et al, 1996; Aslin et al, 1998), who proposed that infants posit word boundaries at points of low *transitional probability* (taking the syllable to be the relevant domain for calculations). The transitional probability measure (and other proposals based on phonotactic coherence, such as Swingley (2005)'s *pointwise mutual information*) are appealing from a learnability perspective, because they only require that the infant be able to track local relationships between speech units. However, I show in Chapter 4 that when coherence-based proposals of this type are implemented, the resulting segmentation is inconsistent, and generally of low accuracy. As discussed in more

detail there, DiBS is generally similar to these coherence-based models, in that it too makes use of diphones; the primary difference is that DiBS explicitly models word boundaries and phonotactic context.

Another class of proposals for word segmentation comes from work on connectionist models. For example, Aslin et al (1996) trained a connectionist network to predict the next speech element in its input, including a special end symbol for the end of a phrase. They found that the end symbol was in general strongly activated at word boundaries, and less so word-internally, implying that the model had learned something about the statistical signature of word endings from observing the distribution of speech sounds at utterance endings. This research represents a major step forward in our understanding of infant word segmentation, because it indicates that *distributional information at utterance boundaries is useful for discovering word boundaries*. Unfortunately, an intrinsic property of learning connectionist models of this type is that their internal representations are opaque (see Chapter 1); that is, it is not clear what kind of distributional information the model made use of. This is part of the reason that more contemporary proposals make use of more explicit grammatical models.

For example, Goldwater's (2006) formulates word segmentation and word-learning as a joint optimization problem. She defines a prior distribution over wordforms (the *generator*) and a prior distribution over word frequency distributions (the *adaptor*); then a search procedure identifies the segmentation that maximizes the joint probability of these two distributions. This model makes crucial use of the 1-1 correspondence between the occurrence of a word and word boundaries. Thus, Goldwater's (2006) model differs from DiBS crucially in conflating the word

segmentation and word learning problems: under her model, infant word segmentation can only be explained through word-learning, rather than segmenting based on prelexical cues such as stress and phonotactics.

The most similar proposals to DiBS in the literature are Xanthos (2004) and Fleck (2008), which are to some extent incremental models that identify word boundaries on the basis of phonotactics. These models differ from DiBS chiefly in the cognitive assumptions that they make. For example, based on considerable psycholinguistic evidence (Saffran et al, 1996; Mattys et al., 1999; Mattys & Jusczyk, 2001), DiBS crucially assumes that infants can track sublexical sequences of length 2. In contrast, Fleck (2008) assumes that infants track sublexical statistics of length up to 5, an assumption which is unsupported by infant research to my knowledge. These models also differ from DiBS in terms of the larger processing framework assumed; specifically, they do not distinguish prelexical and lexical representations.

Baseline model

Chapter 2 developed baseline-DiBS, a supervised learning model in which the statistically optimal diphone statistics p(# | xy) are given by training the model on the same corpus it is tested on. Then, a phonetic transcription of the British National Corpus (BNC) was created using the CELEX pronouncing dictionary (Baayen et al, 1995). Corpus Experiment I replicated and extended the findings of Cairns et al (1997) by running baseline-DiBS on a phonetic corpus derived from a large (~100 MW) written corpus. The results showed very high accuracy (~92%) with a moderate hit rate (~75%) and an extremely low false positive rate (<5%), i.e. a pattern of

undersegmentation.

In a follow-up experiment, baseline-DiBS was tested on the Buckeye corpus (Pitt et al, 2007). The Buckeye is of special interest because it includes a phonetic transcription of naturalistic conversation representing natural pronunciation variation, in addition to a canonical transcription (with the canonical pronunciation for each word). Baseline-DiBS was run on each of these versions of the Buckeye corpus to determine the differences in predictions that arise owing to the use of a canonical transcription (in all other experiments). The results showed that although baseline-DiBS exhibited a considerable degradation in segmentation performance on the 'conversational' version, the extent and severity of degradation were considerably less than for other current-generation models of segmentation, Goldwater (2006) and Fleck (2008). Baseline-DiBS appeared to outperform these more complex models on the 'conversational' corpus owing to two factors. First, DiBS is a prelexical theory of segmentation and so does not explicitly model individual lexical items, whereas both Fleck (2008) and Goldwater (2006) achieve segmentation in whole or in part by recognizing words; the massive variability in word pronunciations in the 'conversational' corpus cannot be represented in these models, so they are forced to model each variant as a separate wordform. Second, DiBS only segments based only on phonotactic cues at word boundaries, and these cues appear to be the ones that are comparatively resistant to conversational reduction processes. I will return to these points in the open issues section later, so will turn to the next point for now.

Cross-linguistic applicability

As reviewed in Chapter 1, most word segmentation research, both psycholinguistic and computational, has focused on English. Thus, cross-linguistic research on this topic is intrinsically worthwhile in contributing to breadth in the field. More importantly for the theoretical issues at hand, word segmentation is a problem faced by infants learning every language. Thus, one of the most important tests for a theory of word segmentation acquisition is that it work for different languages. Chapter 2 addresses this issue by testing baseline-DiBS on Russian.

More specifically, a phonetic transcription of the Russian National Corpus (RNC) was created using Zalizniak's (1977) morphological dictionary and hand-crafted rules based on the grammar described in Avanesov (1967) and Hamilton (1980). Baseline-DiBS was tested on this corpus, with results strikingly similar to the English results from the BNC: overall accuracy of ~92%, and a characteristic pattern of undersegmentation, although the hit rate was substantially lower (~45%) owing to the generally greater length of words in Russian than in English and the existence of prepositions such as *po* which are also frequent prefixes. These results suggest that DiBS has at least some degree of cross-linguistic applicability.

It is entirely possible that the great similarity in segmentation performance between English and Russian owes to specific linguistic similarities between these two languages, or in other words, that these results might not generalize to all languages. With respect to word segmentation, two of the most aspects of Russian are its complex phonotactics and its morphology. As discussed in Chapter 2, Russian and English are alike in having generally complex phonotactics, e.g. allowing complex consonant clusters; however they differ in where

these complex clusters may occur. For example, Russian allows extremely complex onsets (e.g. *vstretit'c'a* 'meet up') but is more restrictive in codas; English allows complex onset (e.g. *strict*) but is especially permissive in codas (e.g. *sixths*). As discussed in Chapter 2 Russian and English are alike in having generally concatenative morphology, but they differ in the extent and richness of their morphology; for example Russian nouns have three different declension classes with 6 cases and 2 numbers (Davidson et al., 1997; Martin & Zaitsev, 2001). In sum, while Russian and English differ in many particulars, in terms of word segmentation they share a number of important similarities. Thus, these results are encouraging in that they show that the good segmentation of DiBS is not specific to English, but further cross-linguistic research is needed to determine whether DiBS is truly cross-linguistically applicable, or if it only succeeds on languages with rich phonotactics.

Learnability

The baseline-DiBS model is supervised, meaning that it is given access to the word boundaries in its training data. This is a reasonable assumption for fluent-listening adults, who after all correctly understand most of the words they hear. However, finding these word boundaries is precisely the problem that infants are trying to solve. Baseline-DiBS is therefore unsuitable as a model of infant word segmentation. Rather, the results of Chapters 2 and 3 serve as a proof-of-concept, illustrating that DiBS is at least a tenable theory. For DiBS to truly explain the acquisition of word segmentation, however, there must be a learnability story, by which infants can estimate good DiBS statistics even if they cannot obtain the statistic optimum. In fact,

part of the motivation for DiBS is precisely this learnability issue – owing to the small diphone domain, the number of parameters in a DiBS model is small, implying that the parameter should not require too much data to estimate, provided a reasonable estimation procedure can be found.

Chapter 4 develops exactly such a learning procedure. The first step is to rewrite the fundamental DiBS statistic p(# | xy) using Bayes' Rule:

$$p(\# \mid xy) = p(xy \mid \#)p(\#)/p(xy) \tag{6.1}$$

Assuming that the value p(#) and the distribution p(xy) are available to the infant, as reviewed in Chapter 1 or argued in Chapter 4, the infant is in a position to estimate p(# | xy) if they can estimate p(xy | #). The next step is to assume phonological independence across a word boundary:

$$p(xy \mid \#) = p(x \leftarrow \#)p(\# \rightarrow y) \tag{6.2}$$

where the terms p(x ← #) and p(# → y) represent the distribution of speech sounds word-finally and -initially, respectively. In *lexical-DiBS* these distributions are estimated from the learner's lexicon and in *phrasal-DiBS* they are estimated using from the distribution of speech sounds at utterance edges, a formally precise interpretation Aslin et al's (1996) insight that utterance-edge information is useful for word segmentation.

In Corpus Experiment III, these learning theories are tested; the results show that with a

full lexicon lexical-DiBS achieve performance nearly identical to baseline-DiBS, whereas phrasal-DiBS is slightly inferior but still generally comparable. The lexical-DiBS result in particular is both reassuring and surprising; reassuring as it offers a real validation of the learning theory, but surprising because the estimation mechanism is so different from baseline-DiBS and yet the results are almost identical. Crucially, both learning models exhibit *undersegmentation*, with the implication that if human infants do segment using a DiBS-like mechanism, then their prelexical parser should exhibit undersegmentation throughout the lifespan.

As reviewed in Chapter 1, the predictions of certain other prelexical proposals of word segmentation are not very clear, as they have not been satisfactorily modeled, or at least not modeled in a consistent way comparable with what is done here. Thus, in Corpus Experiment V, a number of coherence-based models are tested, such as one based on Saffran et al's (1996) proposal of forward transitional probability. The results show that these coherence-based measures achieve consistently inferior segmentation to DiBS; not only do they exhibit over+undersegmentation, but the overall level of accuracy is much lower. Thus, these models are not consistent with an efficiently designed processing system; rather, they predict that the lexical access system must repair both kinds of prelexical parsing errors.

Given these results, DiBS appears to be a better theory of prelexical word segmentation. However, Corpus Experiment IV tested lexical-DiBS in the best case, when it is equipped with a full lexicon. In infant acquisition, the parsing statistics should be gradually refined as the infant learns more and more words, but crucially, starting from a small lexical inventory. Thus, Corpus

Experiment VI investigated the effect of vocabulary size on lexical-DiBS parsing by randomly generating 100 lexicons for a range of vocabulary sizes. The results showed a steadily increasing level of segmentation which was better than chance even for very small vocabularies, and close to ceiling for vocabularies of 1000 words. Thus Chapter 4 provides a full learnability account for DiBS and illustrates that it achieves near-ceiling performance without requiring too much data.

Error patterns

The importance of the two-stage framework becomes clear when considering the error patterns exhibited by the prelexical segmentation mechanism. This is because the lexical access mechanism must 'catch' any errors made by the prelexical parser. As discussed at length in Chapter 1, there are two types of errors: the parser misses a word boundary when there was one (*miss*), or it falsely identifies a word boundary when there wasn't (*false alarm*). Thus, there are three possible error patterns: *undersegmentation*, in which the parser misses often but almost never false alarms; *oversegmentation*, in which the parser false alarms often, but almost never misses; and *over+undersegmentation*, in which the parser exhibits a substantial rate of both misses and false alarms. Which error pattern the prelexical parser exhibits therefore defines the nature of the input to the lexical access mechanism.

One of the most important findings of this dissertation is that all DiBS models undersegment. There are two significant implications of this fact. First, this implies that the lexical access mechanism need only ever cope with an undersegmentation error pattern in its input (if this prediction of DiBS is in accord with the true human facts). This is an efficient

design, as the prelexical parser correctly filters one error type, and the lexical access mechanism need only handle the other. Second, DiBS predicts that in the early stages of lexical acquisition, infants will learn undersegmented words. Anecdotally, this appears to be the correct behavior, although I am not aware of published studies which unambiguously indicate that children undersegment more than they oversegment. From a dynamics perspective, this is also a desirable behavior, as it is guaranteed to avoid an 'error snowball'.

An error snowball is the theoretically undesirable situation in which an initial segmentation error results in learning of an improper word, which then causes further mis-segmentation, which cause further improper word-learning, and so on. Of course, this does not happen in actual child language acquisition; the question is why not; and this is where DiBS model offers some insight. Given the fact that DiBS undersegments, there is a clear prediction that if a word learning error occurs, the improper word will consist of multiple words (e.g. *thehorse*). In terms of lexical-DiBS specifically, there is but a single effect of learning this improper word: it increases the 'boundary-hood' of the onset *th* and the offset *s*. This is not a bad outcome, as these really are word-boundary segments. More generally, when the model improperly learns a multi-word sequence as a word, the boundaries of that improper word are still legitimate word boundaries. Thus, there is no significant adverse effect on the boundary statistics estimated by lexical-DiBS. The only way to adversely affect the boundary statistics of lexical-DiBS would be to oversegment, which would incorrectly lead the model to treat word-internal diphones as boundary cues. The fact that DiBS undersegments is what prevents this from happening. Viewed from this perspective, undersegmentation is a conservative strategy that

prevents the learner from entering an error snowball.

Lexical access

In chapter 5, a theory of lexical access was developed in the context of the two-stage speech processing framework. The lexical access mechanism and the prelexical parser are designed to work together, with the prelexical parser identifying some word boundaries with high reliability; the lexical access mechanism then further decomposes the output of the prelexical parser by matching it against stored wordforms in the lexicon. More specifically, the lexical access mechanism identifies a range of possible decompositions and assigns a probability distribution to them using a unigram probability model. In case a decomposition includes an unmatchable word, a probability is assigned as the product of the prior probability of encountering a new word (p_v) and the lexical phonotactic likelihood that a novel word will have the target form. Both cases are illustrated in Equation 6.3 below:

input: *thebike*

decomposition 1: $p(\textit{the}+\textit{bike}) = p(\textit{the}) \cdot p(\textit{bike}) = f(\textit{the})/F \cdot f(\textit{bike})/F$

decomposition 2: $p(\textit{thebike}) = p_v \cdot p(\omega_v = [\text{Dib2k}]) = n_{hapax}/F \cdot (1-p_\#)^4 \cdot p_\# \cdot |\Phi|^{-5}$ (6.3)

where F is the expected frequency mass of the lexicon (including reserved mass for unseen items), n_{hapax} is the number of hapax types (types which occur with a frequency of 1), $p_\#$ is the prior probability of a word boundary, and $|\Phi|$ is the number of phones in the language (see

Chapter 5 for details).

Corpus Experiment VII assessed the utility of this theory of lexical access by examining its effect on segmentation. Input from the BNC was divided into samples representing roughly one days' worth of input (~30,000 wds, 4000 phrases, see Appendix 2B). A subset of these samples were selected for assessment, representing testing at one month intervals during the second six months of life. The model was equipped with the full lexicon of forms that occur in the BNC; hard parses from the prelexical parser served as the input to the lexical access system. The results showed that the lexical access system increased the hit rate to near-ceiling without substantially increasing the false-alarm rate. In other words, the prelexical parser and lexical access mechanism proposed here do indeed function together to achieve near-ceiling segmentation. This outcome did not depend strongly on the quality of the prelexical parser – as long as it did not oversegment, the lexical access mechanism exhibited excellent decomposition.

However, this outcome was predicated on the learner having access to the 'correct' lexicon, in which morphologically complex words are represented as single words for the purposes of statistical estimation and wordform matching. In order for infants to achieve this kind of performance, they must be able to learn exactly this kind of word – so a theory of word-learning theory must be developed.

Toward word-learning

As reviewed in Chapter 5, comparatively little is known about which word*forms* infants learn and why. A variety of cognitive factors such as verbal working memory and expressive

vocabulary are implicated in vocabulary development (e.g. Masoura & Gathercole, 2005); more specifically linguistic factors such as the phrasal position in which a word occurs (Tardif, Gelman, & Xu, 1999), conformance to the dominant stress pattern of the language (Swingley, 2005), and phonotactics and lexical neighborhood density (Storkel et al, 2006) also appear to matter. It is safe to say that there is no generally accepted theory of word-learning that predicts under what circumstances a wordform will be learned.

In the present case, exactly such a predictive theory is needed. That is because the ultimate goal of this dissertation is to gain insight by modeling the interlocking problems infants face in word segmentation, word recognition, and word learning. The ideal theory would specify under what circumstances infants learn wordforms, and allow for a close fit with observed developmental facts, such as the trajectory of infants' lexicon sizes. However, this is an area of very active research, and the basic facts are not fully known, though it is clear that word-learning is a complex behavior.

As a first, crude pass at this problem, I proposed that infants learn words based on the frequency with which they have segmented them out from their input. More specifically, I proposed that infants track lexical 'candidates' in their input. Every time the lexical access mechanism selects a winning decomposition that includes unmatched input (a novel word, i.e. lexical access failure), the unmatched form becomes a lexical candidate, or if it has already become a lexical candidate, its frequency is incremented. Once a candidate has been accessed 10 times, it is 'learned', i.e. transferred from the list of candidates and entered into the lexicon proper.

While crude, this learning theory models several important properties of word-learning. First, it models the property that words which occur more frequently are more likely to be learned (Storkel et al, 2006). Second, it models the property that words with better junctural phonotactics are more likely to be learned (Mattys & Jusczyk, 2001), a property which falls out naturally from DiBS parsing. Third, it models the property that shorter words are easier to learn, owing to the generally higher lexical phonotactic probability assigned to shorter words.

Of course, this model of word-learning is insufficient in a number of ways. For example, it does not take into account any of the social factors in word-learning (Baldwin, 1995; Baldwin et al, 1996), except in the indirect sense that socially important items and events will tend to be more frequent. Also, this model does not take into account lexical factors such as phonological similarity to known words, which appear to play an inhibitory role in word-learning during infancy (Swingley & Aslin, 2000; Stager & Werker, 1997) but a facilitatory role in more proficient word learners (e.g., Masoura & Gathercole, 2005).

A full bootstrapping model was created and tested in Corpus Experiment VIII. For a prelexical parser, this model used mixture-DiBS, a linear mixture of phrasal-DiBS and lexical-DiBS in which the weighting of lexical-DiBS grew as the model learned more and more words. The lexical access system included the theories of lexical access and word learning developed in Chapter 5. The results were similar to those of Experiment VII in that the combined system achieved a near-ceiling hit rate. However, they were unlike Experiment VII in that the combined system failed to achieve a near-floor false positive rate; in other words, the combined system exhibited aggressive oversegmentation.

Inspection of the acquired lexicon revealed that this was in part owing to a number of single-consonant 'words' such as [t], [d], [s], [z], [v], [n], [k], and [l]. Since many of these are indeed meaningful sublexical units of English (plural/possessive allomorphs: [t]/[d], [s]/[z]; pronominal clitics: [v]/*I've*, [l]/*I'll*), it is not necessarily a problem that the system acquired these units. However, the lexical access system was not designed to model morphological relationship between such sublexical units. Thus, once [s] was learned as a 'word', there was no constraint which forced it to be recognized only word-finally. As confirmed by mathematical analysis in the general discussion of Chapter 5, the result was that such single-segment 'words' were indeed segmented off word-initially. This further entrenched the single-segment words and led to more of them.

A final experiment was conducted to determine whether this issue could be addressed with word-learning constraints. Specifically, Corpus Experiment IX was exactly like the previous experiment, except with one additional constraint: a lexical candidate must contain a vowel to be added to the lexicon. The results showed that this simple constraint remarkably improved the ultimate performance of the bootstrapping model, both in terms of reducing its aggressive oversegmentation, and in terms of the lexicon that was acquired. Note that this experiment is not intended to involve the cognitive claim that infants actually make use of exactly this constraint; nor is it designed to achieve the best possible lexicon. Rather, it was intended to gain insight on what properties of the system were causing the failure in Experiment VIII.

Thus, taken together, the results of this dissertation suggest that *a relatively naïve statistical approach (DiBS) is able to achieve quite high performance on prelexical word*

segmentation (Chapters 2-4), but that a naïve statistical approach fails when it comes to word-learning (Chapter 5). Instead, a richer representational apparatus is needed for lexical access and word-learning. Some specific issues associated with these points are discussed in more detail in the following sections.

Outstanding issues and future directions

Lack of prosody

To my mind, one of the most surprising aspects of this work is the relatively high degree of segmentation that can be achieved without seriously grappling with issues of prosodic representation. The only levels of prosodic representation in this dissertation are the phone, the word, and the phrase, in the sense that the model is given phrases (sequences of phones) as its input and must partition them into words. The corpora used in this dissertation do not have an explicit representation of stress; stress is only represented indirectly through its segmental reflexes (e.g. absence of vowel reduction). Intonation and prominence relations are also not expressed – there is no representation of focus, information structure, relative prominence of different stressed syllables. In fact, even function words such as *the* are realized with a canonical (stressed) pronunciation in every experiment except Corpus Experiment III with the Buckeye corpus. There is no representation of syllable or foot or mora or any other intermediate level of representation from the prosodic hierarchy (Selkirk, 1984; Nespor & Vogel, 1986). It is highly surprising to me that such a high level of segmentation can be achieved without reference to these levels of representation, because there are many reasons to think that each of them can be

informative for word segmentation, as discussed below.

Absence of stress. Stress is a highly informative cue for word segmentation in English. As noted in Chapter 3, English exhibits grammatically-conditioned regularities in stress which may be useful for word segmentation. For example, Cutler & Carter (1987) found that over 90% of content word tokens began with a stressed syllable in a large corpus of spontaneous British English speech. Thus, as long as infants can distinguish the onsets of stressed syllables, and these onsets are aligned with word onsets (cf. Swingley, 2005), infants should be able to achieve a high degree of success in word segmentation by positing word boundaries before stressed syllables. Indeed, as reviewed in Chapter 1, this is precisely what English-learning 7.5-month-olds appear to do (Jusczyk, Houston, & Newsome, 1999). There are certain issues that arise with respect to stress, however.

First, syllabification is not transparently available in the signal, as listeners from different language backgrounds syllabify the same signal in different ways (Dupoux et al, 1999). This fact implies that syllabification too must be learned.

Second, it is not clear that syllable onsets are always aligned with word onsets. In fact, Swingley (2005) convincingly demonstrates the importance of *re-syllabification* phenomena for word segmentation.[42] He showed this by altering the syllabification according to a prelexical segmentation algorithm for varying percentages of the syllable boundaries, finding that his word-

42 Re-syllabification describes cases in which morphological structure does not align with syllable structure; this kind of misalignment is especially likely when a morphological unit with a final consonant precedes another morphological unit which begins with a vowel. For example, in Russian the nominative singular form of 'city' is *`go.rod* where the stem-final /d/ is syllabified into a coda; in the nominative plural *go.ro.`da* the same /d/ is syllabified into the onset. Similarly in my own speech the sequence *can't a* is sometimes realized [kæ.nə], i.e. the nasal is morphologically associated to the first vowel but syllabified with the second vowel.

finding algorithm (see Chapter 1 for details) was not robust against this variation. At present it is unclear to what extent re-syllabification actually occurs in spontaneous speech (for discussion see Swingley, 2005); but given that it occurs at all it is potentially a serious issue for syllable-based theories of word segmentation.

A final issue that arises with stress is that the accent system varies cross-linguistically. For example, French is reported to have phrase-final accenting (Rossi, 1980; Vaissiere, 1991) . This accenting is not lexically contrastive, so that 'stress' in French is purely demarcative rather than distinguishing words as it does in English (*`re.cord* vs. *re.`cord*). On the one hand, this suggests that stress is an even better cue for word segmentation in French than in English. On the other hand, this cross-linguistic variation means that the learner must first discover the stress system of their language before they can make use of it for word segmentation. How infants actually do this is an active area of research (e.g. Dupoux, Sebastian-Galles, Navarette, & Peperkamp, 2008; Skoruppa et al, in press); in fact, this is part of why I neglected stress in the phonetic transcriptions of the corpora. It is my hope that a DiBS-like account can be applied to stress learning, but developing such an account lies outside the scope of this dissertation.

Absence of intonation. Like stress, intonation is not represented in the phonetic transcriptions used in this dissertation. Intonation is largely a sentential property in both English (Pierrehumbert, 1980) and Russian (Davidson et al., 1997; Martin & Zaitsev, 2001), meaning that intonational contours are not associated with individual content words but with entire phonological phrases. As such, intonation is not likely to be as useful for word segmentation in

these languages as other aspects of linguistic structure that are more clearly associated with individual lexemes, such as stress. However, there are other languages in which the intonational structure is lexically-driven. For example, Japanese possesses lexically contrastive intonational patterns (Pierrehumbert & Beckman, 1988), so intonational structure is likely to be a more useful cue for word segmentation in Japanese. One of the formally attractive properties of DiBS is that it can easily be extended to model exactly such structure. The basic equations remain largely the same; only the prosodic domain and nature of the units change.

The Boston Radio News Corpus (Ostendorf, Price, & Shattuck-Hufnagel, 1995) is an corpus of radio broadcasts annotated according to the ToBI prosodic standard (Silverman et al, 1992). This corpus crucially differs from the British and Russian National Corpora used throughout this dissertation in that it explicitly marks prosodic organization (with break indices) and intonation. The DiBS theory of word segmentation and in particular the learning theories described here can be modified to include intonational and prosodic structure. Thus, the Boston Radio News Corpus is an invaluable source of data for pursuing this line of research, which I must leave to the future.

Absence of syllable structure. As remarked above in the discussion of Swingley (2005), syllabification and word segmentation are related, and both must be learned; the relationship is not simple, however. For example, it is not the case that every word boundary is truly aligned with a syllable boundary, as re-syllabification may induce morphological-prosodic misalignments. It is likely that word segmentation and syllabification can be learned jointly

(Johnson, 2008), but further research is needed to determine under how often and under what circumstances re-syllabification actually occurs and what kind of problem it presents for word segmentation.

Certainly some oversegmentation errors could be prevented by modeling syllable structure. For example, as noted in Chapter 5, sequences such as *with several* being parsed as *wi thseveral*. The sequence *ths* is parsed as word-internal because it occurs so frequently in fractions (e.g. *fourths, fifths*) and a few other words such as *depths*. Of course, the problem is that this sequence is only licit in English codas, not in English onsets. A richer prosodic structure that learned such syllabic constraints would avoid this kind of error.

Absence of other levels of prosodic hierarchy. Beyond syllable structure specifically, this dissertation has neglected other levels of the prosodic hierarchy, in particular the foot and prosodic word. In fact, as noted in Chapter 5, inspection of the output of the prelexical parser suggests that in many cases it identifies prosodic words or closely corresponding units, consisting of a content word and possibly one or two function words such as determiners or prepositions. These can be regarded as initially promising results.

Indeed, just as word boundaries can be identified by modeling their statistical signature, other levels of prosodic representation may be modeled as well. As remarked above with reference to stress, DiBS was designed to be extensible to other levels of representation. Joint optimization of word boundaries and other levels of prosodic structure is not only possible theoretically, but likely to result in better modeling at each level individually (Johnson, 2008). It

is for this reason that I left other levels of prosodic representation to future research.

However, this research has suggested one place in which a richer prosodic representation would be especially helpful – in learning new words. As discussed in Chapter 5, the word-learning mechanism and lexical access mechanism proposed in this dissertation are under-constrained. The lexical access mechanism does not represent dependencies between sublexical units, so that if it is set to search for such units (e.g. the plural allomorph [s]) it is currently unable to distinguish their occurrence in morphologically appropriate position from their occurrence elsewhere, resulting in an error avalanche as in Experiment VIII. One type of constraint that is likely to address this problem is a constraint on word learning (Pierrehumbert, p.c.) – novel words can only be admitted to the lexicon if they can be realized as full prosodic words, i.e. according to some minimal word constraint (McCarthy & Prince, 1986/1996). At present, this level of representation is lacking, and indeed, just as with syllabification, there are important learnability issues that must be addressed. Prosody is both an outstanding issue and a promising future direction for this research.

Absence of morphological structure. As noted in Chapter 5, the theories of lexical access and word-learning here do not represent complex morphological structure. For example, head-affix dependencies are not represented, so if the system learns (the plural) [s], this 'word' then becomes available to be segmented off from the beginning of words. Segmenting [s] from this morphologically inappropriate position leads to an error snowball in which other single-consonant 'words' are learned and become entrenched. To prevent this from happening, the

lexicon and lexical access system must be enriched to represent/assign morphological structure. Presumably the nature and types of words that can be learned can then be more appropriately constrained.

Mixture-DiBS

One issue with mixture-DiBS (the incremental version of DiBS used in the bootstrapping experiments, Experiment VIII and IX) concerns the mixture weights. Recall that mixture-DiBS is simply a linear mixture of phrasal-DiBS and lexical-DiBS:

$$p_{mixture}(\# \mid xy) = (\omega_{phrasa} \cdot p_{phrasal}(\# \mid xy) + \omega_{lexical} \cdot p_{lexical}(\# \mid xy))/(\omega_{phrasal} + \omega_{lexical})$$

$$\omega_{phrasa} = \text{number of input phrases} \qquad \omega_{lexical} = \sum_{\omega \in \Omega} f(\omega) \qquad (6.4)$$

As shown in Experiments VII, phrasal-DiBS achieves its near-ceiling segmentation with a minimum of training data, i.e. by or before the model has accumulated six 'months' of language exposure. Moreover, as shown by Experiments IV and VI, lexical-DiBS achieves superior performance to phrasal-DiBS with vocabularies of even a few hundred words. Thus, it makes sense to choose a weighting scheme which does not continue to 'reward' phrasal-DiBS after it reaches ceiling. but which does continue to reward lexical-DiBS as it continues to yield better segmentation.

The weighting scheme described in Equation 6.4 (which was used in Experiments VIII and IX) does continue to add weight to phrasal-DiBS as the model's language exposure increases.

In fact, it increases linearly with the model's language exposure as measured in number of input phrases. The intention behind this weighting scheme was that the phrasal-DiBS component be heavily weighted initially, but gradually yield to lexical-DiBS as a richer and richer lexicon was acquired. Specifically, since the average phrase contains about 7.5 words, the lexical-DiBS weighting should eventually be weighted almost 7.5 times as heavily as phrasal-DiBS, once the learner reaches a steady-state in which they recognize most of the words they hear.

The level of prelexical segmentation attained by the prelexical parsers in Experiments VIII and IX involves a hit rate of about 55%, which is approximately the ceiling level of performance achieved by phrasal-DiBS in Experiment IV, but below the maximum hit rate of 65% attained by lexical-DiBS in Experiment VI (which investigated segmentation as a function of vocabulary size). Thus, the maximum level of prelexical segmentation achieved in the bootstrapping experiments (VIII and IX) is driven by the phrasal-DiBS component, even though the lexical-DiBS component should be able to yield better segmentation by this vocabulary level. This suggests that the weighting scheme weights the phrasal-DiBS component too heavily relative to lexical-DiBS.

One way to address this would be to stop increasing the weighting of phrasal-DiBS once it has reached its maximum. For example, the weighting of the phrasal-DiBS component could be frozen at the value it takes at 6 'months' of language exposure, as phrasal-DiBS has evidently reached its near-ceiling level of segmentation already with this amount of language exposure. In contrast, the weighting of lexical-DiBS should continue to increase as more words are learned.

Pronunciation Variation

One of the most interesting results in this dissertation comes from Experiment II, in which baseline-DiBS was tested on the Buckeye corpus (Pitt et al, 2007) of conversational speech. Specifically, the segmentation performance on a canonical transcription of the corpus was compared to the segmentation performance on a more phonetically precise transcription representing a number of conversational reduction processes. The performance of baseline-DiBS was also compared against the current-generation gold-standard models of Goldwater (2006) and Fleck (2008). The most interesting result of this comparison was that although all three models exhibited degraded performance on the 'conversational' transcription, DiBS exhibited significantly *less* degradation than the other two models.

As discussed in Chapter 2, this effect is driven by two facts. First, conversational production processes give rise to multiple pronunciation variants. Second, the kind of pronunciation variation that occurs appears to selectively target word-internal sequences; in other words, the junctural phonotactic cues that signal word boundaries are relatively well-preserved by conversational reduction processes. In Goldwater's (2006) and Fleck's (2008) models the first fact can only be handled by positing each distinct variant as a separate word type, which for different model-internal reasons leads to a degradation in performance in each case. In contrast, DiBS is an entirely prelexical theory of word segmentation, so the distinction between distinct word types and pronunciation variants of the same type is essentially irrelevant. The second fact explains why DiBS is not as adversely affected by conversational reduction as the other two models – in fact, conversational reduction processes appear to selectively preserve exactly the cues which are

the most important for DiBS segmentation.

Pronunciation variation is a serious issue in a variety of speech domains. The above discussion illustrates why it is a problem for word segmentation. However, the same problem occurs in automatic speech recognition contexts (Jurafsky & Martin, 2008). In this context, there are two general types of solutions. The first is to store some canonical representation of a wordform, and use a generative model the kinds of variation that are likely to occur, e.g. with a string-distance algorithm. The second type of solution is to explicitly list all the pronunciation variants of a word. In either case, open-vocabulary scenarios in which the system is likely to encounter unknown words are a problem. Seen in this light, the results of Experiment II are quite promising, as they suggest that novel words' boundaries may be detected on the basis of prelexical phonotactics even when the novel words themselves cannot be recognized. That is, although DiBS is ultimately intended as a cognitive model of the acquisition of word segmentation, it may prove of use in speech technology applications. For example, it may help to identify unknown words in open-vocabulary contexts.

Toward word learning

According to the two-stage speech processing framework adopted in Chapter 1, word recognition is mediated by a prelexical phonological parser, which assigns a preliminary phonological parse to speech input. This parse is then used to facilitate downstream processing by initiating lexical access attempts at locations in the speech stream which are likely to correspond to word boundaries. Together, these two subsystems present an efficiently designed

system, in which the prelexical parser reliably identifies some word boundaries using phonological generalizations, and the lexical access system recovers the remaining boundaries through recognition of specfic lexical items. Chapters 2-4 developed the baseline and learning theories for DiBS, a diphone-based theory of prelexical word segmentation. Chapter 5 then concentrated on theories of lexical access and word-learning, the remaining components needed for a complete incremental/bootstrapping theory of word segmentation, word recognition, and word learning.

The results of Chapter 5's experiments suggest the naïve statistical approach taken here is not sufficient to support such a complete account. More specifically, while a naïve statistical approach appears to suffice for prelexical/phonological parsing, a richer model of morphological structure is needed than the one adopted in Chapter 5. Experiment VII showed that when the bootstrapping system is equipped with the 'correct' lexicon to begin with, the combination of prelexical parser and lexical access mechanism can achieve near-ceiling segmentation. However, the bootstrapping experiments (VIII and IX) showed that in the absence of more specific constraints, the word-learning mechanism would not acquire the 'correct' lexicon.

Rather, the naïve word-learning theory proposed here acquired a number of affixes which cannot occur on their own. For example, it acquired the affixes *re-* and *de-*, whose status in infant comprehension is unknown to my knowledge. Under the unigram theory of lexical access adopted in Chapter 5, the true dependencies between such affixes and their morphosyntactic heads cannot be represented. Thus, either the lexical access mechanism must be enriched to model these dependencies, or the word-learning mechanism must be further constrained to

distinguish these sublexical units from full words. In light of these findings, a richer model of the lexicon is called for, e.g. as a network of sublexical units (Bybee, 1995; Baayen, 2003). Instead, or possibly in addition, a richer prosodic representation may be called for, e.g. constraining the word-learning mechanism to learn only words which can be realized as full prosodic words.

Additionally, in order to instantiate this word-learning model I was forced to make a number of *ad hoc* decisions, such as an arbitrary frequency threshold of 10 after which a lexical candidate was admitted to the lexicon. To my mind, *these facts highlight crucial gaps in our knowledge of infant wordform learning*. Existing research has demonstrated that a wide variety of factors are important for word learning, including the social context, (e.g., Baldwin, 1995), cognitive properties of the learner (e.g., Masoura & Gathercole, 2005), and both sublexical and lexical properties of the word itself (e.g. Storkel et al, 2006). A growing body of research demonstrates that by 18 months infants exploit function words for grammatical categorization (e.g. Mintz, 2003; Peterson-Hicks, 2006). However if the learning theories instantiated in the bootstrapping experiments (Corpus Experiments VIII and IX) are even coarsely correct, the results suggest that function items may some of the *first* words to be segmented and learned, a possibility which is consistent with existing research on infant perception of function elements (Shi, Morgan, & Allopenna, 1998; Shi & Werker, 2001; Shi, Werker, & Cutler, 2006; Shi, Werker, & Morgan, 1999). In spite of this wealth of research on factors that influence word learning, we have far to go in understanding why infants learn the words they do, and why they do not learn the words they do not.

Summary of Contributions

This dissertation has made a number of theoretical contributions to our state of knowledge.

I regard the most important contribution of this dissertation as the devleopment of a learning theory for DiBS, whereby word boundaries can be estimated from information that infants almost certainly can observe and represent. In particular, phrasal-DiBS illustrates how to estimate word boundary probabilities using the distribution of speech sounds at utterance boundaries. Lexical-DiBS illustrates how the same word boundary probabilities can be estimated from word types and frequencies in the infants emerging lexicon. These two learning theories can be combined in mixture-DiBS, which dynamically weights the two components according to their reliability. These models define a fully incremental and cognitively plausible model of prelexical word segmentation.

Another contribution of this dissertation is to replicate and extend earlier findings which suggest that phonotactic approaches to word segmentation are cross-linguistically applicable (Batchelder, 2002; Fleck, 2008). Specifically, this dissertation investigated the segmentation performance of DiBS on English and Russian, and found a generally similar pattern of *undersegmentation* with high overall accuracy (92%) (although there was a cross-linguistic difference in that the overall hit rate was significantly lower in Russian, in part because Russian words are longer). Russian and English share a number of phonological properties, in particular their relatively complex phonotactics (see Chapter 3 for discussion); but they differ in morphological structure, with Russian possessing a much richer inflectional system (see Chapter

3 for discussion). The high degree of similarity of DiBS' segmentation on these two different languages is a promising result, as any learning theory must be cross-linguistically applicable if it is truly part of how humans learn language.

Perhaps most controversially, this dissertation has argued forcefully that word segmentation and word learning, while conceptually related, are indeed separate problems that are solved separately by infants. In part this claim is motivated by the review of acquisition literature in Chapters 1 and 5 suggesting that infants achieve some degree of prelexical word segmentation in order to start learning more wordforms. In part it is motivated by evidence supporting a distinction between prelexical and lexical processing in adults (Pierrehumbert, 2001; Vitevitch & Luce, 1998). I submit that the results of Experiment VII provide additional support for this claim: Experiment VII showed that the prelexical DiBS parser and lexical access system may function together to achieve near-ceiling segmentation. The prelexical parser identifies many word boundaries, filtering all or most false alarms; the lexical access system can efficiently search the lexicon, secure in the positive word boundary identifications made by the prelexical parser. In other words, this two-stage processing framework achieves efficient, near-ceiling performance specifically by separating the problems of word segmentation from word recognition and word learning.

An interesting result that merits further attention is the study of pronunciation variability in Chapter 2 (Experiment II). As remarked above, pronunciation variability is an important topic for both cognitive theories of speech perception and speech technology applications such as automatic speech recognition (Jurafsky & Martin, 2008). The results of Experiment II suggest

that phonotactic approaches may hold some promise for speech technology applications, as conversational reduction processes appears not to target junctural cues as much as word-internal structure. Without explicit models of pronunciation variation – which can be computationally costly – lexical models are forced to model pronunciation variants as distinct word types. At the least these results suggest the need for future research on phonotactic approaches to pronunciation variability.

Moreover, the results of the bootstrapping experiments (Corpus Experiments VIII and IX) highlight crucial gaps in our understanding of word*form* learning specifically. Acquisition research has demonstrated that word learning is a complex behavior which is affected by a variety of social, cognitive, and linguistic factors (Hall & Waxman, 2004); but despite this research, fundamental questions about wordform learning remain. In particular, the role of function items in infant speech processing is in its infancy.

References

"Декрет о введении нового правописания (Decree on intorduction of new orthography)." Известия В.Ц.И.К. 13 October 1918, #223 (487). 1917. http://bibliography.ufacom.ru/method/dekret.html.

Anderson, J.L., Morgan, J.L., and White, K.S. (2003). A Statistical Basis for Speech Sound Discrimination. *Language and Speech* 46(2-3), 155-182.

Arlotto, A. (1972). *Introduction to Historical Linguistics.* New York: Houghton-Mifflin.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9, 321-324.

Aslin, R.N., Woodward, J., LaMendola, N., & Bever, T.G. (1996). Models of word segmentation in fluent maternal speech to infants. In J.L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, Mahwah, NJ: Erlbaum, pp. 117-134.

Avanesov, R.I (1967) R*usskoje Literaturnoje Proiznoshenije.* (Russian Literary Pronunciation) Moskva: Prosveshchenie.

Baayen, R.H. (2001). *Word Frequency Distributions.* Text, Speech and Language Technology Series, Vol. 18. Dordrecht: Kluwer.

Baayen, R.H. (2003). Probablistic approaches to morphology. In Bod, R., Hay J. and Jannedy, S. (eds). *Probabilistic Linguistics.* MIT Press, pp. 229-288.

Baayen, R.H., Dijkstra, T., & Schreuder, R. (1997). Singulars and Plurals in Dutch: Evidence for a Parallel Dual-Route Model. *Journal of Memory and Language*, 37(1), 94-117.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX Lexical Database (Release 2). Linguistic Data Consortium, Philadelphia.

Baayen, R.H. & Schreuder, R. (1999). War and Peace: Morphemes and Full Forms in a Noninteractive Activation Parallel Dual-Route Model. *Brain and Language*, 68(1-2), 27-32.

Baayen, R.H., Schreuder, R., & Sproat, R. (2000). Morphology in the Mental Lexicon: A Computational Model for Visual Word Recognition. In F. Van Eynde and D. Gibbon (eds.) *Lexicon Development for Speech and Language Processing*, 267–293. Dordrecht: Kluwer.

Baldwin, D.A. (1995). Understanding the link between joint attention and language. In C. Moore & P. J. Dunham (Eds.), Joint attention: Its origins and role in development (pp. 131-158). Hillsdale, NJ: Lawrence Erlbaum.

Baldwin, D.A., Markman, E.M., Bill, B., Desjardins, R.N., Irwin, J.M., & Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, 67(6), 3135-3153.

Barbosa, P. A. (2002). Integrating gestural temporal constraints in a model of speech rhythm production. In S. Hawkins & N. NGuyen (Eds.), *Proceedings of the ISCA workshop on Temporal Integration in the Perception of Speech* (p.54). Cambridge: Cambridge University Printing Service.

Batchelder, E.O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, 83, 167–206.

Bauer, L. (1983). *English Word-formation.* Cambridge: Cambridge University Press.

Bauer, L. (1988). *Introducing Linguistic Morphology.* Edinburgh: Edinburgh University Press.

Bauer, L. (2003). *Introducing linguistic morphology (2nd ed.).* Washington, D.C.: Georgetown University Press.

Beckman, M. & Pierrehumbert, J.B. (1986) Intonational Structure in Japanese and English. *Phonology Yearbook III*, 15-70.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of Perceptual Reorganization for Nonnative Speech Contrasts: Zulu Click Discrimination by English-Speaking Adults and Infants. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3), 345-360.

Bien, H., Levelt, W. M. J. and Baayen, R. H. (2005) Frequency effects in compound pro-duction. *Proceedings of the National Academy of Sciences*, 102, 17876-17881.

Blanchard, D. & Heinz, J. (2008). Improving Word Segmentation by Simultaneously Learning Phonotactics. In A. Clark & Toutanova, K. (eds.), *Proceedings of the Conference on Natural Language Learning (CoNLL)*. pp. 65–72.

Bod, R. (1998). *Beyond Grammar: An Experienced-Based Theory of Language.* Stanford: CSLI Publications.

Bod, R. (2001). Sentence Memory: Storage vs. Computation of Frequent Sentences. In *CUNY Conference on Sentence Processing 2001*. Philadelphia.

Bod, R., Scha, R., & Sima'an, K. (2003). *Data-Oriented Parsing.* CSLI Publications, University of Chicago Press.

Booth, A.E. & Waxman, S.R. (2003). Mapping words to the world in infancy: Infants' expectations for count nouns and adjectives. *Journal of Cognition and Development*, 4(3),

357-381.

Borensztajn, G., Zuidema, W., & Bod, R. (2008). Children's grammars grow more abstract with age - Evidence from an automatic procedure for identifying the productive units of language. In *Proceedings of CogSci 2008* (47-51). Austin: Cognitive Science Society.

Bortfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech stream segmentation. *Psychological Science*, 16, 298-304.

Botvinick, M. & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, 113, 201-233.

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34, 71-105.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61, 93-125.

Brent, M.R. and Siskind, J.M., (2001). The Role of Exposure to Isolated Words in Early Vocabulary Development. *Cognition*, 81/82, 33-44 .

The British National Corpus, version 3 (BNC XML Edition) (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5), 425-455.

Cairns, P., Shillcock, R.C., Chater, N., Levy, J. (1997). Bootstrapping word boundaries: a bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, 33, 111-153.

Cassidy, K.W. & Kelly, M.H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, 30, 348-369.

Cho, T., & Keating, P. (2001). Articulatory strengthening at the onset of prosodic domains in Korean. *Journal of Phonetics*, 28, 155-190.

Chomsky, N. & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, 1: 97-138.

Christiansen, M.H., Allen, J. & Seidenberg, M.S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13, 221-268.

Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. L. (2003). Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics*, 31, 585-598.

Corbett, G.G. (1982): Gender in Russian: An Account of Gender Specification and its Relationship to Declension. *Russian Linguistics*, 6(2), 197-232.

Cutler, A. & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133-142.

Cutler, A. & Norris, D.G. (1988) The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113-121.

Dahan, D., & Brent, M. R. (1999). On the discovery of novel word-like units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General*, 128, 165-185.

Daland, R., Sims, A. D., and Pierrehumbert, J. (2007) Much ado about nothing: A social network model of Russian paradigmatic gaps. In *Proceedings of the 45th Annual Meeting of the*

Association for Computational Linguistics, pp. 936-943. Prague, Czech Republic.

Dale, P.S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, 28, 125-127.

Davidson, D.E., Gor, K.S., & Lekić, M.D. (1997). *Russian stage one : Live from Moscow!* American Council of Teachers of Russian. Kendall/Hunt Pub: Dubuque, Iowa.

Davidson, L. & Roon, K. (2008). Durational correlates for differentiating consonant sequences in Russian. *Journal of the International Phonetic Association*, 38 (2): 137-165

Davis, M. (2004). Connectionist modelling of lexical segmentation and vocabulary acquisition. In Quinlan, P. (Ed) *Connectionist Models of Development: Developmental Processes in Real and Artificial Neural Networks*. Psychology Press, Hove, UK.

De Pijper, J. R., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America*, 96, 2037-2047.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C. & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, 25: 1568-1578.

Dupoux, E., Sebastián-Gallés, N., Navarrete, E., & Peperkamp, S. (2008) Persistent stress 'deafness': the case of French learners of Spanish. *Cognition* 106, 682-706.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.

Fais, L., Kajikawa, S., Amano, S., & Werker, J. F. (in press). Infant discrimination of a morphologically relevant word-final contrast. *Infancy*.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, 146, 668–679.

Fernald A, Taeschner T, Dunn J, Papousek M, de Boysson-Bardies B, Fukui I. 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. J Child Lang. 1989.16(3):477-501.

Fisher, C., & Tokura, H. (1996). Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 343-363). Mahwah, NJ: Lawrence Erlbaum Associates.

Fleck, M. M. (2008). Lexicalized Phonotactic Word Segmentation. *Proceedings of the Association for Computational Linguistics* 2008.

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, 29, 109 –135.

Friederici, A. D., & Wessels, J. M. I. (1993) Phonotactic knowledge of word boundaries and its use in infant speech-perception. *Perception & Psychophysics*, 54, 287-295.

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language*, 42, 481–496.

Gathercole, S.E., Hitch, G.J., Service, E., & Martin, A.J. (1997). Phonological Short-Term Memory and New Word Learning in Children. *Developmental Psychology*, 33(6), 966-979.

Ghahramani, Z (1998). Learning dynamic Bayesian networks. *Adaptive Processing of Sequences*

and Data Structures, Lecture Notes In Computer Science; Vol. 1387, pp. 168-197.

Giegerich, H. J. (2004). Compound or phrase? English noun-plus-noun constructions and the stress criterion. *English Language and Linguistics*, 8, 1-24.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, 73, 135-176.

Goldwater, S. (2006). *Nonparametric Bayesian Models of Lexical Acquisition.* Unpublished dissertation, Brown University.

Goldwater, S., Griffiths, T.L, & Johnson, M. (2006). Contextual Dependencies in Unsupervised Word Segmentation. P*roceedings of Coling/ACL*, Sydney, 2006.

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, 18(3),

254–260.

Green, D. & Swets, J.M. (1966). *Signal detection theory and psychophysics.* New York: John Wiley and Sons Inc.

Guasti, M.T. (2004). *Language Acquisition: The Growth of Grammar.* The MIT Press, Cambridge, MA.

Hall, D. G., & Waxman, S. R. (2004). Weaving a lexicon. Cambridge, Mass: MIT Press.

Halle, M. and Marantz, A. (1993). Distributed Morphology and the pieces of inflection. In Hale, K. and Keyser, S. J., editors, *The View from Building 20*, pages 111-176. Cambridge MA: MIT Press.

Hamilton, W. S. (1980). *Introduction to Russian Phonology and Word Structure*. Slavica.

Harris, Z. (1955). From phoneme to morpheme. *Language*, 31(2).

Hay, J. (2003). *Causes and Consequences of Word Structure.* Routledge, New York and London.

Hay, J. (2007). The phonetics of 'un'. In Munat, J. (ed.), *Lexical Creativity, Texts and Contexts.* John Benjamins Publishing, pp. 39–57.

Hay, J. and Baayen, R.H. (2002) Parsing and Productivity. In Booij, G. and van Marle, J. (eds.), *Yearbook of Morphology 2001*. Kluwer Academic Publishers, pp. 203-235.

Hay, J., Pierrehumbert, J.B., & Beckman, M. (2004) Speech Perception, Well-formedness and the Statistics of the Lexicon. In Local, J., Ogden, R. and Temple, R (eds) *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Cambridge University Press, Cambridge, pp. 58-74.

Hayes, B. (1984). The phonetics and phonology of Russian voicing assimilation. In M. Aronoff and R. T. Oehrle (eds.), *Language Sound Structure*. MIT Press, Cambridge, Mass., 318-328.

Hayes, B. (2000). Gradient well-formedness in Optimality Theory. In J. Dekkers, van der Leeuw F., & van de Weijer, J. (eds.), *Optimality Theory: Phonology, Syntax, and Acquisition*. pp. 88-120. Oxford University Press

Hayes, B. & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39, 379-440

Henry, L. & MacLean, M. (2003) Relationships between working memory, expressive vocabulary and arithmetical reasoning in children with and without intellectual disabilities. *Educational and Child Psychology*, 20(3), 51-63.

Hockema, S.A. (2006). Finding words in speech: An investigation of American English.

Language Learning and Development, 2(2), 119-146.

Hyman, L. M. (1975). *Phonology*. New York: Holt, Rinehart, and Winston.

Ito, J. (1986). *Syllable Theory in Prosodic Phonology.* University of Massachusetts, Amherst: PhD Dissertation.

Johnson, K. (2004) Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (eds.) *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium.* Tokyo, Japan: The National International Institute for Japanese Language. pp. 29-54.

Johnson, M. (2008). Using Adaptor Grammars to Identify Synergies in the Unsupervised Acquisition of Linguistic Structure, *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* pp. 398-406.

Johnson, M., Griffiths, T.L, & Goldwater, S. (2007) Adaptor Grammars: a Framework for Specifying Compositional Nonparametric Bayesian Models. *Advances in Neural Information Processing Systems*, 19, 2007.

Johnson, E.K. & Jusczyk, P.W. (2001). Word segmentation by 8-month-olds: when speech cues count more than statistics. *Journal of Memory and Language*, 44(4), 548-567.

Jones, D. (1969). *An Outline of English Phonetics.* Cambridge: W. Heffer & Sons, Ltd.

Jurafsky, D. (2003). Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In Bod, R., Hay J. and Jannedy, S. (eds). *Probabilistic Linguistics.* MIT Press, pp. 39-96.

Jurafsky, D. & Martin, J.H. (2008). *Speech and Language Processing: An Introduction to*

Natural Language Processing, Speech Recognition, and Computational Linguistics (2nd edition). Prentice-Hall.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993) Infants' preference for the predominant stress patterns of English words. *Child Development,* 64, 675-687.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993) Infants' sensitivity to the sound pattern of native language words. *Journal of Memory and Language*, 32, 402-420.

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999) Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, 61, 1465-1476.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999) The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39, 159-207.

Jusczyk, P. W., Luce, P. A. & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33, 630-645.

Kajikawa, S., Fais, L., Mugitani, R., Werker, J.F., & Amano, S. (2006). Cross-language sensitivity to phonotactic patterns in infants. *Journal of the Acoustical Society of America*, 120(4), 2278-2284.

Kazlauskiene, A. & Velickaite, K. (2003). Pastabos del lietuviu klabejimo tempo. *Acta Linguistica Lithuanica*, 48, 49-57.

Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. (2003). Domain-initial articulatory strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in Laboratory Phonology, VI: Phonetic Interpretation* (pp.143-161). Cambridge: Cambridge

University Press.

Kelly, M. H., & Bock, J. K. (1988) Stress in time. *Journal of Experimental Psychology: Human Perception and Performance*, 389-403.

Kingdon, R. (1958). *The Groundwork of English Stress.* London: Longman.

Kochetov, A. (2002). *Production, perception, and emergent phonotactic patterns: A case of contrastive palatalization.* New York, London: Routledge.

Kohler, W. (1967). Gestalt psychology. *Psychological Research*, 31(1), 18-30.

Kuhl, P K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S. & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9, F13-F21.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044), 606-608.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

Lehiste, I. (1960). An acoustic-phonetic study of open juncture. *Phonetica Supplementum ad* 5, 1-54.

Liberman, M. & Sproat, R. (1992). The Stress and Structure of Modified Noun Phrases in English. In I. Sag (ed.), *Lexical Matters*, pp. 131-181, CSLI Publications, Chicago, University of Chicago Press.

Lieber & Stekauer (2009). Introduction: status and definition of compounding. In Lieber, R. & K. Stekauer (eds.) *The Oxford Handbook of Compounding*, pp. 3-18.

Luce, P., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, 16, 565–581.

Luce, P.A. & Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, 19, 1-36.

Lyapunov, A. M. (1900). Sur une proposition de la théorie des probabilités. *Izv. Imp.Acad. Sci. St. Pétersb.*, sér. 5, t. 13, pp. 359-386.

Lyapunov, A. M. (1901). Nouvelle forme du théorème sur la limite de probabilité. *Mém. Imp. Acad. Sci. St. Pétersb. Cl. phys.-math.,* t. 12, No. 5, separate paging.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Majerus, S., Poncelet, M., van der Linden, M., & Weekes, B.S. (2008). Lexical learning in bilingual adults: The relative importance of short-term memory for serial order and phonological knowledge. *Cognition*, 107, 395–419.

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science*, 6, 314-317.

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT: Cambridge.

Marchand, H. (1960). *The Categories and Types of Present-Day English Word-Formation.* Wiesbaden: Otto Harassowitz.

Marcus, G.F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German Inflection: The Exception that Proves the Rule. *Cognitive Psychology*, 29, 189-256.

Martin, C. & Zaitsev, A. (2001). *Russian stage two: Live from Moscow!* American Council of Teachers of Russian. Kendall/Hunt Pub: Dubuque, Iowa.

Masoura, E.V. & Gathercole, S.E. (2005). Contrasting contributions of phonological short-term memory and long-term knowledge to vocabulary learning in a foreign language. *Memory*, 13(3), 422-429.

Mattys, S.L. & Jusczyk, P.W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78, 91-121.

Mattys, S.L., Jusczyk, P.W., Luce, P.A., & Morgan, J.L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38, 465-494.

McCarthy, John J. (1982). Prosodic Structure and Expletive Infixation. *Language*, 58(3), 574-590.

McCarthy, J. & Prince, A. (1986/1996). *Prosodic Morphology 1986*. Technical Report #32, Rutgers University Center for Cognitive Science.

McClelland, J. L. & Elman, J. L. (1986). The TRACE Model of Speech Perception. *Cognitive Psychology*, 18, 1-86.

Mehl, M.R., Vazire, S., Ramirez-Esparza, N., Slatcher, R.B., & Pennebaker, J.W. (2007). Are Women Really More Talkative Than Men? *Science*, 317(5834), 82.

Mester, A. (1990). Patterns of truncation. *Linguistic Inquiry*, 21:478–485.

Mintz, T.H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1), 91-117.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Towards an

understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 756 –766.

Nazzi, T., Jusczyk, P.W., and Johnson, E.K. (2000). Language Discrimination by English-Learning 5-Month-Olds: Effects of Rhythm and Familiarity. *Journal of Memory and Language*, 43, 1–19.

Nespor, M. & Vogel I. (1986). Prosodic Phonology. Dordrecht: Foris.

Norris, D. & McQueen, J.M. (2008) Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2), 357-395.

Olsen, S. (2000). Compounding and stress in English: A closer look at the boundary between morphology and syntax. *Linguistische Berichte*, 181, 55-69.

Ostendorf, M., Price, P.J., & Shattuck-Hufnagel, S. (1995). *The Boston University Radio News Corpus.* Technical report, Boston University.

Padgett, J. (2003). Russian Voicing Assimilation, Final Devoicing, and the Problem of [v] (or, The mouse that squeaked). *Ms. University of California, Santa Cruz.* [ROA #528.]

Papousek, H. & Papousek, M. (1991). Innate and cultural guidance of infants' integrative competencies: China, the United States and Germany. In M. Bornstein (Ed.), *Cultural approaches to parenting* (pp. 23-44). Hillsdale, NJ: Erlbaum.

Peterson-Hicks, J. (2006). *The Impact of Function Words on the Processing and Acquisition of Syntax.* Unpublished dissertation, Northwestern University.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation.* PhD thesis, MIT.

Pierrehumbert, J. (1994). Syllable Structure and Word Structure. *Papers in Laboratory*

Phonology III, Cambridge Univ. Press, Cambridge, UK. 168-188.

Pierrehumbert, J. (2001) Why phonological constraints are so coarse-grained. In J. McQueen and A. Cutler (eds) *SWAP special issue, Language and Cognitive Processes*, 16 5/6, 691-698.

Pierrehumbert, J. (2002) Word-specific phonetics. *Laboratory Phonology VII*, Mouton de Gruyter, Berlin, 101-139.

Pierrehumbert, J. (2003) Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, 46(2-3), 115-154.

Pierrehumbert, J. and M. Beckman (1988). *Japanese Tone Structure*, Linguistic Inquiry Monograph 15, MIT Press, Cambridge.

Pierrehumbert, J. and D. Talkin, (1991) Lenition of /h/ and glottal stop. *Papers in Laboratory Phonology II*, Cambridge Univ. Press, Cambridge UK. 90-117.

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech (2nd release)* (URL: www.buckeyecorpus.osu.edu). Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Plag, I. (2006). The variability of compound stress in English: Structural, semantic, and analogical factors. *English Language and Linguistics* , 10(1): 143-72.

Polka, L., Colantonio, C. & Sundara, M. (2001). Cross-language perception of /d-ð/: Evidence for a new developmental pattern. *Journal of the Acoustical Society of America*, 109(5), 2190-2200.

Poser, W. (1990). Evidence for foot structure in Japanese. *Language*, 66:78–105.

Rietveld, A. C. M. (1980). Word boundaries in the French language. *Language and Speech*, 23, 289-296.

Roach, P. (1983). *English Phonetics and Phonology: A Practical Course.* Cambridge: Cambridge University Press.

Rose, R. L. (2005). The phonological optimization of nicknames in Japanese: Why kids don't sing 'Sachi-chan wa ne'. *Proceedings of Linguistics Society of Japan*, 131, pp. 228-233.

Rossi, M. (1980). Le francais, langue sans accent? [French: A non-stressed language?]. In I. F'onagy & P. Leon (Eds.), *L'accent en francais contemporain* [Stress in modern French] (pp. 13–51). Paris: Didier.

Rumelhart, David E., Geoffrey E. Hinton and Ronald J. Williams (1986). Learning internal representations by error propagation. In D.E. Rumelhart, McClelland, J.L., & the PDP Research Group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1, Foundations)*, 318–362. Cambridge, MA: MIT Press.

Rytting (2004). Greek word segmentation using minimal information. In *Proceedings of the Student Research Workshop. HLT-NAACL 2004 Companion Volume*, pp. 207-212. Boston, Massachusetts: The Association for Computational Linguistics.

Saffran, J.R., Aslin, R.N., & Newport,, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926-1928.

Sampson, R. (1980). Stress in English N+N phrases: a further complicating factor. *English Studies: A Journal of English Language and Literature,* 3: 264-70.

Saussure, Ferdinand de (1983). *Course in General Linguistics.* Eds. Charles Bally and Albert

Sechehaye. Trans. Roy Harris. La Salle, Illinois: Open Court.

Scheibman, J. (2000). I dunno... A usage-based account of the phonological reduction of don't in American English conversation. *Journal of Pragmatics*, 32, 105-124.

Selkirk, E. (1984). *Phonology and Syntax: The Relation Between Sound and Structure.* Cambridge: The MIT Press.

Shady, M.E. & Gerken, L. A. (1999). Grammatical and caregiver cues in early sentence comprehension. *Journal of Child Language*, 26, 1-13.

Shattuck-Hufnagel, S., & Turk., A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25, 193-247.

Shi, R., Cutler, A., Werker, J.F., & Cruickshank, M. (2006). Frequency and form as determinants of functor sensitivity in English acquiring infants. *Journal of the Acoustical Society of America*, 11(6). EL61-EL67.

Shi, R., Morgan, J., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. Journal of Child Language, 25, 169-201.

Shi, R., & Werker, J. (2001). The basis of preference for lexical words in 6-month-old infants. Developmental Science, 6, 484-488.

Shi, R., Werker, J.F., & Cutler, A. (2006). Recognition and representation of function words in English-learning infants. *Infancy*, 10(2), 187-198.

Shi, R., Werker, J., & Morgan, J. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. Cognition, 72, B11-B21.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, J. Pierrehumbert, J. Hirschberg, & P. Price (1992). TOBI: A Standard Scheme for Labeling Prosody, *Proceedings of the International Conference on Spoken Language* 92, Banff, Oct 12-16, 1992

Skoruppa, K., Pons F., Christophe, A., Bosch, L., Dupoux, E., Sebastián-Gallés, N., Limissuri, R.A., & S. Peperkamp (in press). Language-specific stress perception of stress by nine-month-old French and Spanish infants. *Developmental Science*.

Soderstrom, M., Kemler-Nelson, D. G., & Jusczyk, P. W. (2005). Six-month-olds recognize clauses embedded in different passages of fluent speech. *Infant Behavior and Development*, 28, 87-94.

Soderstrom, M., White, K.S., Conwell, E. & Morgan, J.L. (2007). Receptive grammatical knowledge of familiar content words and inflection in 16-month-olds. *Infancy*, 12, 1-29.

Soderstrom, M., Jusczyk, P. W., & Wexler, K. (2001). English-learning toddlers' sensitivity to agreement morphology in receptive grammar. In *Proceedings of the 26th Annual Boston University Conference on Language Development*, Boston.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388, 381–382.

Stekauer, P., Valera, S., and Diaz, A. (2007). Meaning predictability and conversion. *ms*.

Storkel, H.L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research,* 49, 1175-1192.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive*

Psychology, 50: 86-132.

Swingley, D. and Aslin, R.N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, 76(2), 147-166.

Syrett, K. & Lidz, J. (2005). Children Want to Access Every Interpretation Adults do. *Proceedings of North East Linguistics Society 35*. GLSA: Amherst, MA.

Tardif, T., Gelman, S.A., Xu, F. (1999). Putting the 'noun bias' in context: A comparison of English and

Mandarin. *Child Development*, 70(3), 620-635.

Thorn, A. S. C., & Frankish, C. R. (2005). Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 729–735.

Truss, L. (2003). *Eats, Shoots & Leaves.* London: Profile Books.

Tsao, F.-M., Liu, H.-M., Kuhl, P. K. (2006). Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *Journal of the Acoustical Society of America*, 120, 2285-2294.

Vaissiere, J. (1991) Rhythm, accentuation and final lengthening in French. In: J. Sundberg, L. Nord, & R. Carlson (Eds.), *Music, language, speech and brain* (108–120). Wenner–Gren International Symposium Serie, vol. 59.

van de Weijer, J. (1998). *Language Input for Word Discovery.* Ph.D. thesis. Max Planck Series in Psycholinguistics 9.

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech.

Computational Linguistics, 27, 352–372.

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech*, 40, 211–228.

Vitevitch, M. S. (2002a). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 735–747.

Vitevitch, M. S. (2002b). Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear. *Language and Speech*, 45, 407–434.

Vitevitch, M.S., Armbruster, J. & Chu, S. (2004). Sub-lexical and lexical representations in speech production: Effects of phonotactic probability and onset- density. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 30, 514- 529.

Vitevitch, M.S. & Luce, P.A. (1998). When words compete: Levels of processing in spoken word perception. *Psychological Science*, 9, 325-329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory of Language*, 40, 374–408.

Vitevitch, M.S. & Luce, P.A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers*, 36, 481-487.

Vitevitch, M.S., Luce, P.A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, 40, 47-62.

Werker, J.F., Fennell, C.T., Corcoran, K., & Stager, C.L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, 3, 1-30.

Werker, J.F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, 103, 147-162.

Werker, J. & Tees, R. (1984) Cross-language speech perception evidence for perceptual reorganization during the first year of life. *Infant Behavior and Developmen*t, 7:49-63.

Wexler, K. (1994) Optional infinitives, head movement and the economy of derivations in child grammar. In D. Lightfoot & N. Hornstein (Eds.) *Verb movement* . pp. 305-350. Cambridge, MA: Cambridge University Press.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 91, 1707-1717.

Xanthos, A. (2004). Combining Utterance-Boundary and Predictability Approaches to Speech Segmentation. *Proceedings of the Psycho-computational Models of Language Acquisition Workshop at COLING 2004*, pp. 93-100.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, 8(10).

Zalizniak, A. A. (1977). *Grammaticheskii slovar russkogo jazyka* [Grammatical dictionary of the Russian language]. Russkii jazyk: Moskva.