PRESERVING SUBSEGMENTAL VARIATION IN MODELING WORD SEGMENTATION (OR, THE RAISING OF BABY MONDEGREEN)

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Christopher Anton Rytting, B.A.

*****

The Ohio State University
2007

Dissertation Committee:

Dr. Christopher H. Brew, Co-Advisor

Dr. Eric Fosler-Lussier, Co-Advisor

Dr. Mary Beckman

Dr. Brian Joseph

Approved by

_____

Co-Advisor

_____

Co-Advisor
Graduate Program in Linguistics

ABSTRACT

Many computational models have been developed to show how infants break apart utterances into words prior to building a vocabulary—the "word segmentation task." Most models assume that infants, upon hearing an utterance, represent this input as a string of segments. One type of model uses statistical cues calculated from the distribution of segments within the child-directed speech to locate those points most likely to contain word boundaries.

However, these models have been tested in relatively few languages, with little attention paid to how different phonological structures may affect the relative effectiveness of particular statistical heuristics. This dissertation addresses this issue by comparing the performance of two classes of distribution-based statistical cues on a corpus of Modern Greek, a language with a phonotactic structure significantly different from that of English, and shows how these differences change the relative effectiveness of these cues.

Another fundamental issue critically examined in this dissertation is the practice of representing input as a string of segments. Such a representation implicitly assumes complete certainty as to the phonemic identity of each segment. This runs counter both to standard practice in automatic speech recognition (where "hard decisions" are eschewed) and, more crucially, overestimates the ability of infants to parse and identify those segments from the spoken input. If even adult native speakers (with the benefit of higher-level linguistic knowledge, such as a

vocabulary) do not always agree with with one another on the identity of a particular segment (as demonstrated not only by mishearings at the word level, but also by annotator disagreement when transcribing corpora), then infants cannot be expected to do so.

Furthermore, transcriptional (or "sequence-of-segments") input implicitly assumes that all instances of a given phoneme (or even word, for word-level transcriptions) are identical, which is clearly not the case in speech. Hence, models using transcriptional input, regardless of symbol set, make unrealistic assumptions which may overestimate the models' effectiveness, relative to how they would perform on variable input.

Some connectionist models of word segmentation aim for a more realistic input representation by using phonological feature vectors rather than atomistic symbols for segments. Inasmuch as these representations reflect the greater acoustic similarity of closely-related phonemes, they better approximate the acoustic input available to the child. However, based as they are on essentially *binary* vectors, they still treat instances of the same phoneme identically, and hence fail to represent the great subsegmental variability found in speech. Attempts to approximate subsegmental variation have relied on untested statistical assumptions about the nature of subsegmental variability, instead of using an actual acoustic signal.

This dissertation proposes an improved representation of the input that preserves the subsegmental variation inherently present in natural speech while maintaining sufficient similarity with previous models to allow for straightforward, meaningful comparisons of performance. The proposed input representation uses an automatic speech recognition system to divide audio recordings of child-directed speech into sectors corresponding to the segments expected in a canonical pronunciation, and, using an automatic phone classifier, converts these

sectors into phone probability vectors. The resulting phone probability vectors are then transformed into real-valued vectors with the same dimensions used by an influential connectionist model of word segmentation, in order to facilitate comparison with this model. By comparing this real-valued, highly variable input with the original model's binary input, the effect of input representation on this model of word segmentation can be measured, and the true contribution made by segmental, distributional cues vis-à-vis other cues in the signal better assessed.

The results reported here suggest the real-valued inputs used here present a harder learning task than idealized inputs, resulting in poorer performance for the model when tested using only "segmental" (or vector-based) cues. This may help explain why English-learning infants soon gravitate toward other, potentially more salient cues, such as lexical stress. However, the model still performs above chance even with very noisy input, consistent with studies showing that children can learn from distributional segmental cues alone.

To my parents, who taught me how to wonder

ACKNOWLEDGMENTS

To begin, some thanks are in order to those who helped me discover my chosen field of study, long before this dissertation was started. In particular, I thank Bob Spencer for introducing me to my first two foreign languages, and to Deryle Lonsdale and Alan Melby for introducing me to computational linguistics—and for teaching me early on how to present and write my research for a larger audience. More general thanks are also due to those in the Linguistics Department, the Classics program, and several other institutions at Brigham Young University for encouraging me in bringing together unusual combinations of interests and in beginning to think like a scholar and researcher.

Thanks to my advisors: Chris Brew for believing in me from the start, and for his incredible patience in continuing to believe in me till the very end, and for insisting on clarity and simplicity throughout, while giving me the freedom to pursue my own ideas; Eric Fosler-Lussier for bringing a needed second perspective to this project, and particularly for bridging the gap between linguistic viewpoints and conceptions of speech and those of the automatic speech recognition community. Thanks to both for much-needed advice on how to limit the dissertation sensibly, and tackle one problem at a time.

The contributions of the other two members of my committee have been no less valuable along the journey. Brian's boundless enthusiasm for and knowledge of all aspects of the Greek language have been extremely helpful along the way, as

have both his patience in watching this project unfold, and his encouragement to take the time to push forward my work on other aspects of Greek. Mary Beckman's hard questions (most memorably, "What is a syllable?") have helped propel the research forward, even when they seemed to be diverting it or distracting from it. More crucially, they have kept me from taking too much for granted and from losing sight of the infants in the model-what they hear and perceive in midst of the "buzzing, blooming confusion" about them.

Thanks to Shari Speer and her 615 Psycholinguistics class, for introducing me to the research that provided the inspiration for this dissertation, for reading drafts of the NSF GRF proposal that got this dissertation off the ground, and for guiding me (along with Keith Johnson) through the IRB process for my second pregenerals paper. Thanks also to my second pregenerals and generals exam committees: Keith Johnson (along with others already mentioned), for encouraging me to consider the speech signal more carefully, and Dieter Wanner, for asking (along with Brian), "What is a word?" Many other professors both at Ohio State and elsewhere have contributed in numerous ways to my growth as a linguist and as a researcher.

I also wish to acknowledge the assistance of Morten Christiansen, in answering questions concerning his research and granting permission to quote his materials; Erik Thiessen, for sending me copies of his research articles and helpful Matlab scripts; Brian MacWhinney, for granting me access to the CHILDES/TalkBank corpus; Michael Brent and Jeff Siskind, Myron Korman, and Ursula Stephany for collecting the relevant corpora in the first place; Brian Pellom, for allowing use of the SONIC speech recognizer; Eric Fosler-Lussier and Soundararajan Srinivasan (and the creators of HTK itself!) for the HTK-based automatic phone classifier; and Douglas Blank for the Conx neural network toolkit.

I owe a similar debt of gratitude to my fellow students, and here especially I cannot help but to overlook some, for I am indebted to many. To those not mentioned, who should have been: please forgive the oversight. First of all, I thank my cohort: Wes Collins, Robin Dodsworth, and Markus Dickinson. The first few quarters of graduate school was an intense and often intimidating experience, but I can't think of any other group I'd rather share it with. I thank Markus also for introducing me to CGSA. Their faith and fellowship, as well as the faith and fellowship and of my own congregation and of the linguistics department prayer group, have been a much-appreciated source of support and comfort. Thanks to all the members of these groups, past and present, who are too numerous to be named here. Above all, thanks to Him who answers all earnest prayers.

I thank Rich Christiansen for his support and encouragement, and for remembering what it was like to be a graduate student still learning to juggle disparate responsibilities. *Shukran jazilan* also to Nathan Toronto for his example, showing that it's possible to juggle more and larger responsibilities than I can now imagine, and still finish!

A big *efkharistó* to Giorgos Tserdanelis for his insights as a native speaker, for the books and recordings, and for his connections to the Greek community here in Columbus as well as abroad-as well as for the many wide-ranging conversations about Greek linguistics, and linguistics generally.

Many thanks are due also to the computational linguistics and phonetics/phonology discussion groups, the cognitive science program and its graduate student discussion group, and the Speech and Language Technology lab, for helpful comments and questions on preliminary stages of this research. Several members of these groups have kindly given their time for more informal discussion, technical assistance, and general support: Ilana Bromberg, Robin Dautricourt, Paul

Davis, DJ Hovermale, Emily Jamison, Craig Hilts, Grant McGuire, Dennis Mehay, Jeremy Morris, Crystal Nakatsu, Salena Sampson, and Elizabeth Smith. I thank Soundararajan Srinivasan in particular for help with HTK and Matlab, Jeff Mielke for answering numerous questions about his dissertation and research, and Kirk Baker for reading and commenting on early versions of chapter 5.

Others have been very helpful in the preparation of the document itself, particularly those who helped me get up to speed in LaTeX and R: Kirk Baker, Adriane Boyd, Mike Daniels, Jon Dehdari, Markus Dickinson, Anna Feldman, Soyoung Kang, Xiaofei Lu, Detmar Meurers, and lastly Shravan Vasishth (for *not* answering all of my questions, but teaching me how to find the answers on my own). Thanks to Tim Watson for answering questions of formatting and protocol, and to Pauline Welby for giving me inspiration for a title.

Finally, a huge debt of gratitude is owed to my family. My parents, for raising me to appreciate school for what it is—a resource and an opportunity to and help others do the same—and for understanding when the responsibilities that come with an academic life pulled me away and cut short time with them. My sister, dr. jenny rytting, for blazing the trail before me—and for spending long hours putting her hard-won English-major editing skills to work (along with Liz and my sister-in-law, Lauren Monson) in proofreading nearly this entire manuscript and pointing out my numerous infelicities, inconsistencies, and stylistic errors. Any mistakes that remain are my own responsibility, of course—as are the sentence fragments in this paragraph. (Good literary device . . . thanks, David Moser and DRH.) Kirk for help with statistics. Liz for periodically asking me, "Are you making this up?" Last, and certainly far from least, my dear wife Megan, who became engaged to this work in progress rather late in the game, with little idea of what it

would entail, but accepted it whole-heartedly. Thank you for your patience, support, and unconditional love.

Portions of this and closely-related research were conducted with the monetary support of several scholarships: the National Science Foundation Graduate Research Fellowship, the Distinguished Dean's University Fellowship, the Phi Kappa Phi National Fellowship, as well as support from NSF-ITR grant #0427413, granted to Chin-Hui Lee, Mark Clements, Keith Johnson, Lawrence Rabiner, and Eric Fosler-Lussier for the multi-university Automatic Speech Attribute Transcription (ASAT) project.

Preliminary versions of this research were presented to audiences at SIGPHON 2004, the HLT-NAACL 2004 Student Research Workshop, the 2004 Midwest Computational Linguistics Conference, CogSci 2006, and Interspeech 2006 (with another presentation coming very soon to the Linguistics Society of America), as well as several events local to Ohio State. This would not have been possible without the generosity of the NSF GRF cost-of-expense allotment, the Ohio State Linguistics Department graduate travel committee, and other funding sources.

These things and many things besides would not be possible without the tireless and cheerful support of the secretaries, fellowship coordinators, and other staff at Ohio State: Frances Crowell, Jane Harper, and Claudia Morettini in the Linguistics Department; Ewana Witten in the Department of Computer Science and Engineering; and Gail Griffin and Jo Wittenauer in the Graduate School office. The same is true for the amazing technical support supplied by the Linguistics Department system administrator, Jim Harmon, and his able assistants through the years (in alphabetical, and I believe reverse chronological, order)—Jon Dehdari, Jeff McCune, and Matt Hyclak. They are unsung (or at least undersung) heroes.

VITA

3 January, 1976	Born - Columbus, Indiana
2000	B.A., Linguistics, *summa cum laude* Brigham Young University
2000 - 2001, 2004 - 2006	Dean's Distinguished University Fellow, The Ohio State University
2002 - 2005	NSF Graduate Research Fellow, The Ohio State University
2004 - 2006	Graduate Research Associate (Speech and Language Technology Lab) The Ohio State University
Summer 2005	Graduate Teaching Associate, The Ohio State University

PUBLICATIONS

1. Rytting, C.A. and Lonsdale, D.W. 2006. An operator-based account of semantic processing. In *The Acquisition and Representation of Word Meaning: Theoretical and Computational Perspectives*, ed. by Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli, Vol. XXII-XXIII in series *Linguistica Computazionale*, pp. 117-137, Pisa/Roma: Istituti Editoriali e Poligrafici Internazionali.

2. Rytting, C.A. 2005. An iota of difference: Attitudes to *jod* in lexical and social contexts. *Journal of Greek Linguistics*, **6**, pp. 145-179. Amsterdam/Philadelphia: John Benjamins.

3. Rytting, C.A. 2003. Review of "Psycholinguistic Studies of Modern Greek: L1 Acquisition and L1/L2 Sentence Processing." *Journal of Greek Linguistics*, **4**, pp. 152-157.

4. Rytting, C.A. 2002. An empirical test of Analogical Modeling: The /k/$\sim \emptyset$ alternation. In *Analogical Modeling: An Exemplar-based Approach to Language*, ed. by Royal Skousen, Deryle Lonsdale, and Dilworth B. Parkinson. Amsterdam/Philadelphia: John Benjamins.

FIELDS OF STUDY

Major Field: Linguistics
Specialization: Computational Linguistics

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF ACRONYMS

ADS adult-directed speech
ANN artificial neural network
APC automatic phone classification
APR automatic phone recognition
ASR automatic speech recognition
CDS child-directed speech
CWBP conditional word boundary probability
FST finite-state transducer
HMM hidden Markov model
MDL minimum description length
MFCC mel frequency cepstral coefficient
MI mutual information
MLP multi-layer perceptron
MSS metrical segmentation strategy
RNN recurrent neural network
SRN simple recurrent net
TP transitional probability
UBM utterance boundary marker
VOT voice-onset time
WIT window-in-time
WST word segmentation task

CHAPTER 1

INTRODUCTION

Lady Mondegreen never lived, but that didn't stop reports of her death from being greatly exaggerated—or at least widely discussed. In 1954, Silvia Wright reported a now-famous mishearing of a stanza from Percy's (1765) poem *Reliques*:

> Ye highlands and ye lowlands
> Where hae ye been
> They have slain the Earl of Murray
> And laid him on the green.

Young Silvia misheard the last two lines as 'They have slain the Earl Amurray and Lady Mondegreen.' Since that time a *mondegreen* has come to refer to a novel word (or an improbable phrase) arising from a mishearing. The topic of this thesis is neither Lady Mondegreen's death, nor such mishearings *per se*, in either adults or even children of Silvia's age at the time of the mishearing. (For more details on those two topics, see Welby (2003), which discusses intonational cues to word segmentation in French, among many other sources.) Rather, it is about how infants during their first year begin to learn words in the first place. Hence, one might say that it is not the slaying of Lady Mondegreen that is discussed, but her birth.

1.1 What is the word segmentation task?

Before children can begin to acquire the syntax of their language, they first must learn some words. In order to learn these words, first they must be able to separate them out from the speech signal they hear around them. As speakers do not normally mark the boundaries between words (as is conventionally done with spaces in written English), this task is non-trivial. Nevertheless, while no acoustic marker of spoken word boundaries is perfectly reliable and unambiguous on its own, a number of partially usable cues, when used together, enable children to break apart utterances into words they can then match with possible meanings. The task of learning the language-specific cues and segmenting utterances into words is known as the word segmentation task (WST).

Before one can sensibly discuss what cues babies might use to find boundaries between words, it helps first to be clear about what is meant by a word. Unfortunately, this is not as easy as one might suppose! One account (found in Werker and Curtin 2005) informally defines words as "recognizable forms that have meanings shared by speakers of a community or language group" (p. 204). However, this definition is a bit too broad, as it could logically include larger units of language, such as phrases or multi-word idioms, as well as subword units such as the present progressive *-ing* in English, the meaning of which, though perhaps hard to articulate or explain, is nevertheless implicitly shared across speakers.

Another unsatisfactory definition, designed to exclude such subword units (called *affixes* or *bound morphemes*), might be the smallest unit that can appear between two pauses (or changes of speaker). For example, 'skiing' can be an utterance (e.g. in answer to the question 'What are you doing?'), as can 'ski' (in answer

to 'What do you want to do next?'), but '-ing' is not a felicitous answer to any normal question (outside of grammatical or other meta-linguistic discussions about words and subword units themselves).

Unfortunately, such a definition also excludes words like *the* or *a*, which also cannot normally stand alone between pauses. This proposed diagnostic test does not provide ways to distinguish these types of words (sometimes called *function words*, since they communcate a grammatical function or relation rather than semantic content) from affixes like *-ing* or *un-*. By extension, such a test would lead one to consider *a ski* or *the ski* as single words, in the same way that *skiing* is considered a single word. While other tests could be devised to distinguish function words from affixes, the definition of "word" would quickly become very complicated, requiring the application of a number of tests that detract from the main point of the question at hand.

For this reason, it is advisable to appeal to convention, by adopting the simple (or even simplistic) answer that a word is whatever comes between two white spaces on a page. While this definition is admittedly a useful (and much used) approximate operationalization, it is not quite correct, for at least three reasons: first, babies learn language by speaking, not by writing; second, the acoustic stream of speech sound, as mentioned above, generally does not include obvious and consistently reliable correlates to white space in writing; and third, white space does not always align with word boundaries in conversational spoken language, including that spoken to infants. For example, *wanna* and *kinda* (the spoken English forms that cover some but not all meanings of written English's 'want to' and 'kind of') are generally pronounced, and may well be learned by infants, as a single word.

Similarly, compounds such as 'blackbird' may not be distinguished from phrases such as 'black bird' (cf. Vogel and Raimy 2002).

The lexical status of other spoken English idioms such as *y'know*, *Idunno*, and *get'm* (for written English 'you know', 'I don't know', and 'get them/him'), along with child-speech reduplications like 'bye-bye', 'night-night', and 'boo-boo', are less clear.

Nor do these difficulties diminish as one moves to other languages: see Joseph (2002) and other papers in that volume for a lengthy discussion of similar ambiguities in Modern Greek. However, a rigorous definition of 'word' that covers all cases unambiguously is not the point of this thesis. Fortunately, there are quite a number of common, uncontroversial examples of words that children do learn, and most studies confine themselves to these. Furthermore, since many if not most English-speaking parents intend for their children to learn to read and write as well as speak English, and indeed begin practicing the beginnings of literacy skills with them at an early age (for example, pointing at the words in a book while reading aloud to them), the conventional (writing-based) definition of words as units of language delimited by spaces does represent an eventual goal for these learners.

It has long been recognized that words consist of arbitrary pairings between some concept that one speaker wishes to communicate to another, and some form of communicative gesture used to denote that meaning. Saussure ([1916] 1983) refers to the former of these as the *signifié* or that which is signified; the latter is the *signifiant* or signifier. The pairing of signifier and signified constitutes the *sign* as a whole; words are one type of sign.

The child's task in learning the language of its caretakers is to learn not just forms or signifiers, but the pairings. However, before this can be done, the infant must know what forms are available to participate in these pairings; thus, learning

to segment out possible forms from running speech is a necessary first step. A word's *form* may be defined (again, informally) as the range of sound patterns that might occur when someone utters the word—barring egregious mispronunciations, but including all the variety that may arise from different voices, genders, ages, and (to a certain extent) dialects. Since its object of consideration is normally hearing infants of pre-literate age (or rather abstract models of how they might learn words), this thesis excludes written forms from its discussion, focusing on spoken forms only.[1] However, for convenience, liberal use is made throughout much of the research here presented, of symbolic transcriptions representing how a word was said. This approximation rests on the assumption (not entirely uncontroversial, but adopted here for convenience) that spoken language—and hence the spoken form of a word—can be described as a sequence of smaller units of sound. These units may be referred to as phones (or, when the focus rests on particular abstract qualities of these sounds, phonemes). However, for the purposes of this thesis it is usually sufficient simply to refer to these symbolized units of sound, or any instances of them, as *segments*.

This thesis proceeds on the assumption that in the course of language learning, a form cannot be associated with a particular meaning until the form itself is known, and that a form cannot be said to be fully known until it is properly segmented, meaning that at least one instance of the word is correctly separated from the rest of the spoken utterance in which it was uttered, without being further split into smaller parts. The discussion that follows does not address the question of how these forms are then mentally represented, stored in memory, or associated

[1]Thus this dissertation has nothing in particular to say about the other "word segmentation task": namely, the engineering task of separating strings of written characters, letters, or glyphs into words or equivalent meaningful units, for those languages that do not already use white space to guide the reader. For an excellent introduction to that task as applied to written Chinese, see the recent dissertation of Lu (2006) , another computational linguist from The Ohio State University.

with meanings. Nor does it consider how children learn to say these words for themselves, mapping what they hear to the motor commands of their own throat, tongue, and lips that allow them to reproduce these sounds for themselves (though these attempts may well interact with and hasten the receptive learning process; for a model that does address this issue, see Plaut and Kello 1999).

Likewise, the internal structure of words (beyond the level of the segment) and how that might be learned is outside the focus of this dissertation. Hence, obviously related forms like *cat* and *cats*, or *child* and *children*, are regarded as separate word forms (or *word types*) in the lexicon the children are building. For it is not clear that the relationships between these words is at all obvious to young infants, as it is to adults. The various restrictions listed above, along with the use of standard orthographic word boundaries (or close variants as defined and used by the corpus annotators), allow for a clear, unambiguous statement of the word segmentation task by which the various proposed models may be evaluated and compared.

1.2 Modeling word segmentation: An overview

Obviously, this thesis is far from the first (and likely not the last) foray into the topic of how infants pick out word forms from running speech. Indeed, a plethora of models for word segmentation already exists, more than can be done justice to here. After a brief note on terminology, this section outlines a number of previously articulated models.

1.2.1 A note on terminology

The majority of the models discussed in the following section and throughout this dissertation have been developed with reference to text—either written language or symbolic transcriptions of speech as opposed to audio recordings or other acoustic representations of continuous speech. This outlook is revealed in the very name given to the problem: the word segmentation task, rather than the *speech* segmentation task. As a result, a number of terms are used in ways that may seem unusual to members of the speech-processing community. For example, reference is made, particularly in Chapters 3 and 6, to *phonological features*. These refer to various distinctive traits that can be used to distinguish one type of speech sound (or phoneme) from another. These usually refer to articulatory phenomena, such as tongue placement or lip rounding. Traditionally these features are binary, although Christiansen et al. (1998) makes occasional (and insignificant) use of a feature that allows 0.5 as a possible value in addition to 0 and 1. Hence they are very different from the *acoustic* features commonly used in the automatic speech recognition community. Where acoustic features (in the signal-processing sense) are meant, they are specifically referred to as acoustic features.

Similarly, since most of these models are wedded to a representation of speech as a sequence of symbols (e.g. segments or syllables), they typically make use of the convenient fiction that such segments occur in sequence (rather than overlapping each other), and take roughly comparable amounts of time to occur. Hence, when the connectionist models discussed in Section 1.2.2.3 and Chapter 3 refer to "time-steps," they are referring not to fixed increments of time (e.g. the 10-ms time-slice nearly ubiquitous to modern automatic speech recognition and

signal processing), but to much longer (and widely variable) units of time corresponding to the duration of an entire segment. (Models for which the syllable is the underlying unit of input naturally would make use of even longer basic units of time.)

1.2.2 Transcript-based computational models

1.2.2.1 Clustering approaches

Olivier (1968) and BootLex Perhaps the earliest computational model of vocabulary acquisition from running text is Olivier (1968). Starting with a corpus of English text in standard orthography, but with all spaces and punctuation removed, Olivier's algorithm attempts to build up a lexicon of possible words, and then restore the spaces to their proper places. At the start of Olivier's simulation, the lexicon begins with each letter of the alphabet as its lexical entries, each with an initial frequency of 1. The training corpus is divided into smaller units of 480 characters, and each of these units is parsed using the words in the current lexicon, choosing the segmentation that maximizes the likelihood of the section given the probabilities (frequencies) associated with the current lexical entries.

At the end of each parse, each pair of adjacent words is considered as a new possible word. Since no provision is made for limiting the number of words in the lexicon (beyond what was expedient to avoid overloading the very limited electronic memory store then available to Olivier) or constraining the length of these words, Olivier's algorithm has a marked tendency to generate a large number of multi-word long phrases as it is run on large corpora.

BootLex (Batchelder 1997, 2002) is closely based on Olivier (1968), but addresses the issue of overgeneration by adding an "optimal word length" constraint.

Batchelder tested this algorithm much more extensively than Olivier, using a variety of languages (English, Spanish, Japanese), texts (both written and spoken; both child-directed and adult-directed), and transcription systems (orthographic and phonemic; segment-based and mora-based). As expected, transcriptions of spoken language were more easily segmented than written texts; this was largely due to shorter utterances and words, fewer word types, and a greater degree of repetition of those word types (token-type ratio). No clear and consistent differences among the languages or the transcription systems is reported. Although English appears to perform better than the other two languages, this is most likely due to the relative sizes of the training corpora.

PARSER Perruchet and Vinter (1998) also propose an iterative clustering algorithm, but one based more directly on cognitive principles such as human memory limitations. As with the two previous approaches, the PARSER algorithm also starts with a lexicon consisting only of "primitive" elements—the symbols of the encoding alphabet. However, since PARSER was tested on a corpus consisting of a set number of consonant-vowel (*CV*) syllables, PARSER's symbol set uses these syllables rather than segments as its primitive *shaping units* (much like the mora-based transcription used in one variant of BootLex). PARSER could in principle be used with segment-based primitives, though this was not tested explicitly.

Rather than parsing batches of text, and then mechanically combining pairs of adjacent words at the end of each batch, PARSER processes the training corpus incrementally, a few shaping units (or SUs) at a time. It randomly selects a "percept length" of one, two, or three SUs, and treats the corresponding (longest-match) sequence from the corpus as a single "percept." If this percept (and its constituent

SUs) are already found in the lexicon of percepts, their associated weights are increased; otherwise they are added to the lexicon. Other entries that contain this percept have their weights reduced (as a model of "interference"). At periodic intervals, the weights of all the lexical entries are reduced, in order to simulate gradual memory decay. Those entries whose weights fall below zero are removed from the lexicon.

Although the PARSER algorithm has the advantage of incorporating basic mechanisms inspired by the workings of human memory, it has only been tested on "mini-language" corpora corresponding to particular psycholinguisti experiments, such as Saffran et al. (1996b). It has not been tested on actual child-directed speech.

Swingley (2005) A more recent clustering approach was proposed by Swingley (2005) and applied to English and Dutch corpora. Like PARSER, it assumes syllables as its basic unit. Unlike previous approaches, however, it does not iteratively build up a lexicon. Rather it considers, over one pass of the entire training corpus, statistics of word-likeness for different sequences of syllables. Monosyllables are considered words if their frequency is larger than a certain threshold, and if they are not part of any bi- or tri-syllable words. Bisyllables are considered words if frequency and pointwise mutual information (see Section 4.2.3) are above the threshold, and they are not part of any tri-syllable words. Trisyllables are considered words if their frequency and the pointwise mutual information of both of their (overlapping) bisyllable parts are above threshold. Results are reported for all possible thresholds, for both actual syllables (as judged by adult native speakers) and for syllables resulting from a syllabification heuristic.

1.2.2.2 Hierarchical groupings

One issue that arises when considering the segmentation or parsing of a text or utterance is how to treat multiple levels of linguistic structure. For language is not just a sequence of words any more than it is merely a sequence of sounds. Words have subparts (e.g. the bound morphemes referred to above) and words group into larger clusters (such as phrases) below the level of the utterance. Obviously, the infant needs to learn about all these levels of structure, but as the different levels have somewhat different properties, they are usually treated as separate tasks. Embedded units are either ignored during evaluation, treated as "interfering patterns" (as in PARSER) or explicitly excluded (as in Swingley 2005).

A few models, on the other hand, explicitly embrace the notion of nested structure, and do not privilege any particular level of the resulting hierarchy as corresponding directly to the word segmentation task. This feature is intuitively appealing, but makes quantitative evaluation and comparison with other models rather difficult. Very early work on hierarchical clustering models is reported in Wolff (1977) and Collet and Wolff (1977). However, a clearer exposition of a hierarchical approach may be found in de Marcken (1996).

Carl de Marcken (1996), like Olivier (1968) and Wolff (1977), uses sequences of orthographic symbols (letters) as the main input to his model, and initializes his model by assuming these single letters as terminals in his hierarchy. The frequency of each of these is calculated, along with the co-occurrence of adjacent pairs. Then an algorithm for calculating minimum description length (MDL) is used to determine which new words would need to be added to the lexicon in order to allow the most efficient encoding of the lexicon and the corpus together. These words are added, and the frequency and co-occurrence statistics are recalculated given

the new lexicon. Finally, the MDL algorithm is invoked again to see if deleting any of the learned words would shorten the overall encoding. These four steps (calculation, lexicon augmentation, re-calculation, lexicon trimming) are repeated to build additional levels of the hierarchy until the model converges to a stable solution, where no shorter representation may be found.

Although de Marcken's (1996) experiments primarily use letters in standard English orthography, he also presents an interesting side experiment using the output of a "noisy" HMM-based automatic phone recognizer. Again, because de Marcken does not attempt to evaluate his model with metrics directly comparable to other models, it is difficult to interpret his results. Nevertheless, he demonstrates that it is possible in principle to devise a model of word discovery that uses input derived from actual speech, without the intervention of human transcribers.

1.2.2.3 Analytic models

The clustering and hierarchical models discussed above view the infant's task as one of synthesis or concatenation: starting from some primitive set of segments of minimal length (such as a phone or a syllable), words are found by joining adjoining segments together. This may the thought of as a bottom-up approach. The models discussed below take a top-down approach. For these models, the initial unit of analysis is the utterance, defined as the stretch of speech between two clear demarcatory events (such as silent pauses or changes of speaker). The task of the infant is to break up that utterance into smaller units that correspond to words. Some of these models, such as those developed by Michael Brent and colleagues, develop an explicit lexicon like the clustering models examined above. Others, such as Hockema (2006) and the connectionist approaches, focus on learning to identify areas of the utterance that correspond to word edges. In these models, a

vocabulary may be extracted after the corpus is parsed (or the word boundaries identified) for evaluation purposes, but the words learned early on are not explicitly stored and do not directly contribute to the segmentation of other words.

INCDROP The INCDROP model (Brent 1999; Brent and Siskind 2001) employs a very simple but effective idea: when no other information is available, assume that each utterance consists of a single word, and store that new word in the dictionary. However, when an utterance does contain a known word, use that word to break up the utterance into smaller pieces (i.e. the material before and after the word) and treat those as separate utterances. When multiple parses are possible (i.e. two or more overlapping words could be extracted out of the same utterance), the parse that would yield a more compact description of the corpus and lexicon is chosen (in accordance with the MDL principle).

Hockema (2006) A recent article (Hockema 2006) presents a non-incremental approach to word segmentation by examining the characteristics of word edges in English. Insofar as it uses global statistics to determine segmentation points rather than an incremental parsing (or clustering) algorithm, Hockema's top-down approach may be seen as roughly analogous to Swingley's (2005) bottom-up model, which uses global statistics to perform a single clustering step.

Hockema claims that the (ordered) pairs of phonemes that straddle word boundaries form a distinct class from those that do not: that is, that a given ordered phoneme pair either almost always or almost never straddles a word boundary. This is not in and of itself an algorithm for discovering words or word boundaries, since the statistic it depends on requires knowledge of word boundaries to begin with. However, it is useful as a type of theoretical "upper bound" by which other approaches might be measured. It might perhaps be combined with some other

approach to augment its performance. Hockema's observations will be examined in more detail in Section 4.2.4.

Connectionist models The computational models discussed above have characterized the input available to infants as a sequence of symbols—whether segments, morae, or syllables—rather than in terms of a continuous acoustic speech signal. Furthermore, they by and large explore the performance of a single heuristic operating on this symbol sequence, rather than a combination of heuristics acting in concert.

Another class of computational models uses artificial neural networks (or ANNs) in order to gain more flexibility in their input mechanisms. This framework allows these models to consider multiple, possibly simultaneous cues—which is in fact the hallmark of one of these models, proposed by Christiansen and colleagues. There are a number of different artificial neural network (ANN) architectures that have been used for this task. For example, the Playpen model (Gasser and Colunga 1998) uses a generalized Hopfield network with Hebbian learning. Hammerton (2002) uses a self-organizing map in his model of speech segmentation. However, the models that are of most direct concern to this thesis use a type of simple recurrent net (SRN) pioneered by Elman (1990). These models, and particularly the Christiansen model, are discussed in some detail in Chapter 3.

1.2.2.4 Models of audio-visual word learning

Finally, some researchers have developed models that consider not only the (acoustic) linguistic input, but also the visual field of the infant (or of a robot in a corresponding task of language acquisition). Representative models include CELL (Roy 1999; Roy and Pentland 2002) and subsequent work by Yu and Ballard (2002).

These studies address questions of matching form with visual correlates of meaning, rather than focusing on word segmentation in the strictest sense. However, since some of the techniques used are similar, a brief summary is given below.

CELL The CELL model (Roy 1999; Roy and Pentland 2002) starts with an audio-visual display consisting of visual images of some object combined with recordings of a person talking about that object, in the way that one might do if presenting that object to an infant, or interacting with an infant observed to be looking at that object. The speech is processed in 10-ms frames converted into acoustic feature vectors using RASTA-PLP Hermansky et al. (1991). Each of these 10-ms frames is then associated with a probability vector over a pre-determined set of phones using a recurrent neural network (RNN). From this representation, the most likely phone sequence (along with the most probable phone boundary points) is determined using the Viterbi algorithm on a hidden Markov model (HMM) (Rabiner 1989). Repeated words are then detected using a recurrence filter, with allowances made for the variation in pronunciation among instances of a phone or sequence of phones by calculating a symmetric distance metric. Frequently recurring words are then stored in a "short-term memory" module and matched to the visual display using mutual information.

Although this model appears much more sophisticated than previous models, it is difficult to compare with other models, as results are reported only for "groundable words"—that is, words that refer to concrete objects such as the objects displayed in the visual component of the input. Results are not reported for all words, unlike most other models. In addition, the evaluation metrics for word segmentation are much more lenient than those used elsewhere: missegmentations involving articles or inflectional affixes are not counted against the model, nor are

misalignments involving the insertion or deletion of a single phone at the edge of a word.

Yu and Ballard (2002) Subsequent work by Yu and Ballard (2002) uses a similar approach, but extends it to recognizing not just objects, but actions, such as folding or stapling paper, or lining up objects. These actions are arguably more abstract than the objects recognized by the CELL model, and the relevant chunks of language are longer (multi-word verbal phrases) and potentially more variable, using, for example, different tenses of the verb. However, the techniques of analyzing and segmenting the sound are similar, using RNN- and HMM-based automatic phone recognition (defined in Section 5.2.4). The automatic phone recognition system used by Yu and Ballard is actually somewhat simpler than CELL, in that the automatic phone recognition system makes "hard" decisions on the sequence of segments, rather than keeping track of phone probability vectors. In order to equate two instances of the same verb phrase, it performs similarity matching on these phone sequences, using a distance metric based on phonological features similar to those used by Aslin et al. (1996) or Christiansen et al. (1998).

1.3 The current proposal

This thesis focuses primarily on modeling the combination of segmental distributional cues, specifically patterns of segmental predictability (see Section 2.3.3.2) and the distribution of segments before pauses or utterance boundaries (discussed in Section 2.3.3.3). These two cues are examined and combined in two distinct ways. Chapter 4, using a corpus of child-directed Modern Greek, examines these cues directly, by contrasting various statistical formulations of these cues, and comparing their performance separately, in combination, and against theoretical upper

bounds. By extending the examination of these cues to a less-frequently studied language which has a rather different set of patterns relevant to resolving the segmentation problem, it may be seen whether these cues are useful across languages, or are specific to English.

Chapter 6 examines the interaction of these two cues within a connectionist framework, using a previously studied corpus of child-directed British English speech for comparison purposes. It then extends the chosen connectionist model in a novel way, by replacing the idealized, transcriptional (or "string-of-phonemes") input used by nearly all the (unimodal) models here reviewed with real-valued input data taken from an automatic phone classifier's output on audio recordings of American English child-directed speech. This enables us to determine whether the model is robust to different assumptions about the input actually available to infants, or whether it crucially depends on particular types of input.

While this thesis does not directly examine the interaction between segmental cues and suprasegmental cues (such as lexical stress), it lays the groundwork critical to examining this interaction from a viewpoint that is not limited to idealized abstractions of these cues, but is able to examine them as they arise from the acoustic signal. Along these lines, a new interpretation of the role of suprasegmental cues is suggested on the basis of an novel approximation of stress derivable from real-valued segmental cues. The modeling framework as extended here is shown to be flexible enough to accommodate the type of input needed to conduct a full-scale examination of segmental, suprasegmental, and even subsegmental cues.

1.4 Outline of the dissertation

Before we examine the various computational approaches to the word segmentation task in great detail, it is essential to know as precisely as possible the infants'

abilities these models are attempting to explain. In order to set the stage, as it were, Chapter 2 provides a brief survey of empirical findings with regard to the WST. After a brief overview in Section 2.1, Section 2.2 describes relevant research on the acoustic discrimination abilities of infants less than one year old, as they distinguish different voices and languages, hear differences in rhythm, and finally begin learning to differentiate the sounds of their soon-to-be-native language—and also learn not to differentiate between equivalent sounds (instances of the same phoneme) in their language. Section 2.3 lists several important classes of cues that infants learn to use to find word boundaries: periods of silence (Section 2.3.1), properties of the speech (such as stress) that may span over multiple segments (e.g., applying to whole syllables), called *suprasegmental* cues (Section 2.3.2), *segmental* cues, or cues that deal with the way individual segments or phones are distributed (Section 2.3.3), and cues that involve subtle variation in the way a particular sound (or phoneme) is produced in context, called *subsegmental* cues (Section 2.3.4). Finally, Section 2.4 provides a brief timeline of important stages in English-learning infants' development with regards to acoustic discrimination and word segmentation abilities.

Although Section 2.3 describes several types of cues that infants learn to use in concert in order to find word boundaries, most computational models of word segmentation, particularly the uni-model models described in Section 1.2.2, are based primarily if not exclusively on segmental, distributional cues such as those discussed in Section 2.3.3. Chapters 3 and 4 examine these cues from a computational perspective.

Chapter 3 presents a detailed overview of connectionist models of the WST, starting in Section 3.1 with Elman (1990); Cairns et al. (1997), and Aslin et al. (1996). Section 3.2 outlines a model which addresses the interaction between multiple

cues, described and elaborated in Christiansen et al. (1998), and a preliminary attempt in Christiansen and Allen (1997) to model subsegmental variation. Section 3.3 begins to explore two facets of the Christiansen model and the other connectionist models: recurrence (Section 3.3.1) and ways of representing input to the models (Section 3.3.2).

Chapter 4 makes an excursus to consider the relative performance of segmental, distributional cues in Modern Greek, a language where the WST has never before been examined or modeled computationally. Section 4.1 reviews two major heuristics based on such distributional cues, enumerating (in Section 4.2) several variant mathematical formulations used to operationalize these cues in computational and empirical models, and discussing their relative performance in English. Section 4.3 reviews what is not well understood concerning these cues. Three simulations, described in Sections 4.5, 4.6, and 4.7, seek to answer these open questions. Section 4.8 sums up the results of these simulations.

Chapter 5 examines in more detail a crucial limitation of the connectionist models reviewed in Chapter 3 and the statistical heuristics employed in Chapter 4: namely, the assumption of invariance at the segmental level derived from the use of symbolic (and particularly word-level) transcriptions of child-directed speech. By abstracting away from all subsegmental information, these models may underestimate the difficulty of the segmentation task. Alternative methods of modeling the input infants receive are discussed in Section 5.1, which reviews some problems with using transcriptions alone (5.1.1), some previous attempts in the literature to account for subsegmental variation (5.1.2), and proposes some desired characteristics of a more realistic input representation for the WST (Section 5.1.3).

Section 5.2 argues that a more realistic way to approximate subsegmental variation is to use corpora with audio data, coupled with automatic speech recognition technology. Section 5.2.1 specifically defines a proposed representation that has the desired characteristics listed in Section 5.1.3. and Sections 5.2.2 and 5.2.3 show how they may be applied to the statistical heuristics used in Chapter 4 and the connectionist models discussed in Chapter 3, respectively. Section 5.2.4 shows how this representation may be approximated through the use of a particular variant of automatic speech recognition known as automatic phone classification (APC).

Chapter 6 replicates and extends crucial portions of the Christiansen model (excluding lexical stress cues) using the new input representation proposed in Chapter 5 and comparing it with the original input representation and related variants. Simulation 4, described in Section 6.1, replicates the basic model of Christiansen et al. (1998) and compares it to a recurrent neural network (RNN) version of Aslin et al. (1996), in order to verify the claim that the interaction of cues as modeled in Christiansen et al. (1998) improves performance. Simulation 5 (Section 6.2) tests the Christiansen model on data derived from an automatic phone recognizer, to see how robust the model is to the sorts of perception errors made by such a recognition system when provided with naturalistic, spontaneous speech such as an infant is exposed to. Section 6.3 describes the results of Simulation 6, which implements the input representation proposed in Section 5.1. Finally, Section 6.4 discusses the results of Simulations 4-6 and their implications for further extensions and applications of the Christiansen model and other computational models.

Chapter 7 reviews the goals of the dissertation (Section 7.1), the findings (Section 7.2), and the broader implications (Section 7.3) of the research presented here. It then briefly explores in Section 7.4 ways that suprasegmental cues such

as lexical stress could be modeled more realistically from actual audio data. Finally, Section 7.5 lists some potential points of departure for future work within the framework presented here.

CHAPTER 2

REVIEW OF EXPERIMENTAL FINDINGS

2.1 Overview

This chapter reviews the experimental literature relevant to the word segmentation task in English, setting the stage for a later in-depth examination of a selected strand of computational models of word segmentation. A number of properties of spoken English are examined according to cues they may provide to infants seeking to segment spoken language into word forms for vocabulary acquisition and the eventual pairing of these word forms with their meanings. Among the potential cues examined in the literature here reviewed are: (1) the distribution of periods of silence surrounding and demarcating utterances; (2) suprasegmental cues, such as stress; (3) segmental distributional statistics, including phonotactics, troughs in transitional probability, and correlation of segments/features with utterance-final position; and (4) subsegmental cues such as allophonic variation and coarticulation. Stress (considered together with accent and 'full' vowel quality, with which it is strongly confounded in typical American and British English child-directed speech) is found to be a very robust cue for English, and also correlated with a variety of acoustic manifestations (e.g. duration). In English it triggers a trochaic bias (an increased ability to segment words with SW patterns) as early as 7.5 months, and this bias becomes a dominant factor between the ages of 9 and

11 months, overriding other conflicting cues. By about 11 months, children seem to have learned to consider multiple cues in order to correctly segment iambic (WS) words; however, stress continues to play a role in segmentation even in adult speech processing, particularly in conditions of noise.

Before one can sensibly consider what cues babies might use to find boundaries between words, one must know what types of sounds babies can reliably distinguish, and where possible, what they cannot. An extensive literature exists on this topic alone; here it will be touched on only briefly. This chapter will address only auditory cues—that is, cues found within the speech signal that could reasonably be made available to the infant by hearing alone.

2.2 Acoustic discrimination in the first year

2.2.1 Fundamental types of speech sound

Within the first two months after birth, infants are able to tell the difference between speech and non-speech sounds, displaying a preference for speech over non-speech (Vouloumanos and Werker 2004). Speech itself is not a single type of noise (in the way a sine wave or a single sustained chord on an organ could be said to be) but rather the modulation of a variety of types of sound with different characteristics. Some regions of speech (e.g. vowels) are periodic, like a chord on an organ, meaning that a particular pattern of pressure on the eardrum repeats itself with a high degree of regularity. The frequency with which this basic pattern of pressure repeats itself (i.e. how many times per second) is called the fundamental frequency, or f_0. In speech, this is caused by the rate of vibration of the vocal folds,

and it is closely related to the perceived pitch of the sound.[1] Regions of speech with relatively large amounts of periodic noise are sometimes referred to as *sonorant*.

Other regions of speech consist of aperiodic noise. Like the 'hiss' of a snare drum, these vibrations exert pressure on the eardrum in no clear pattern, and are perceived as 'percussive' or 'hissy.' Short burst of aperiodic noise many be produced with a short burst of pent-up air rushing out of the mouth after being 'bottled up' by a 'stopped up' mouth; these are appropriately known as *stop bursts*. More sustained aperiodic noise can be produced by forcing air through a narrow opening, causing the airstream to become turbulent. These are known as fricatives. The shape of the mouth, and which parts of the vocal tract caused the closure or the narrowing, control which bands of the frequency spectrum are most prevalent in the aperiodic noise. (These spectral bands of frequency do not give rise to quite the same sense of pitch that periodic noise does, but they do enable one to distinguish different types of fricatives, as one can distinguish a bass drum from a snare drum.)

Still other regions of speech consist of silence. Usually these regions are short enough not to be confused with utterance boundaries or other pauses, as defined above—at least by native speakers, though they may be confused with pauses by speakers (or listeners) more accustomed to another language. A very common type of brief silence is the *stop closure* when the tongue, lips, or other parts of the mouth have sealed it off and interrupted the airflow. Sometimes the vocal folds will continue to vibrate during part of this interruption, which gives rise to

[1] Other sub-patterns of regularity, with shorter periods and hence higher frequencies, also arise and may be amplified from the particular shape of the mouth during speech. The most important types of resonant frequencies for certain types of sounds, particularly vowels, are called formants. The formant with the lowest frequency is called f_1, and the next-lowest is called f_2. These formants are correlated with the position of the tongue and jaw during the production of the vowel, and are crucial for distinguishing one type of vowel from another.

a *voiced stop*. While stop closures generally sound very similar to one another (i.e. mostly silence), there are sometimes subtle cues at the edges of that silence—such as transitional formants from adjacent vowels—that help the listener determine where the closure of the mouth took place—the *place* of the stop consonant. These are not always reliable, however, and when these cues are the only ones available (e.g. when a stop is unreleased at the end of an utterance or before another stop), the listener may misidentify the place of the stop. If the stop is released, however (e.g. before a vowel), the acoustic cues available in the stop burst are usually sufficient to distinguish the place of the stop (Winitz et al. 1972, cited in Ohala 1990). (See Section 6.3.2 for the implications of this in modeling speech input from an acoustic signal.)

Regions of sound that are perceived to possess a certain coherent structure to them (e.g. periodic noise with constant formants (see footnote 1), aperiodic noise with constant amplitude and spectral qualities, or a stop closure predictably followed by a stop burst) may be referred to as *segments* of speech. However, it must be noted that the segmentation of speech into various regions or units is by no means deterministic, and both naïve speakers (or infant learners) of a language and expert phoneticians may disagree amongst themselves as to what constitutes a segment of speech. Moreover, this underdetermination holds whether the level of segmentation is the phone (as here), the syllable, or even the word or the phrase.

2.2.2 From types of sound to a sense of rhythm

The alternation between different types of speech sounds—sonorant regions, aperiodic regions, and short regions of silence, as well as regions of greater or lesser amplitude, or higher or lower pitch—may set up certain perceptions of rhythm which an infant can (and in fact does) perceive and respond to. Certain languages

are felt to have very different rhythms than others (e.g. Japanese vs. German) and infants, as early as four days after birth, can use these rhythmic patterns to tell their own language from a foreign language with a different rhythmic pattern (Mehler et al. 1988). Precisely which cues are used in making this distinction is debatable: the methodology in such studies has been criticized in Lieberman (e.g. 1996) as not properly filtering out cues besides "rhythmic" or prosodic cues alone.

It is also known that even 3-day-old infants can detect specific instances of lengthened vowels, of the type used to signal word endings in French (Christophe et al. 1994). Neonates can also distinguish different pitch contours from Japanese stimuli (Nazzi et al. 1998). Infants are also able to hear the acoustic correlates of stress very early on. Italian-learning newborns can distinguish two-syllable utterances with stress on the first from those with stress on the second syllable (Sansavini et al. 1997). By two months of age (perhaps much sooner), English-learning infants can do the same in their language (Jusczyk and Thompson 1978).

The exact acoustic correlates to lexical stress differ from language to language (for those languages that have it at all). Some common acoustic correlates with lexical stress are duration of the vowel (for those languages that do not have a phonemic vowel length distinction), spectral tilt (or the balance of intensity between the low-frequency and high-frequency parts of the acoustic signal), and peak amplitude of the vowel. In cases where the word is also accented (as it often is in child-directed speech), the word will have a distinctive pitch contour on or near the stressed syllable. (Pitch accents typically involve a high pitch on the stressed syllable in English statements. More generally, the pitch contour's overall shape and timing vary across languages and intonation types within languages).

Vowel quality (the frequencies of the vowel's formants) also plays a role in English, though not in all languages. Further discussion of these cues is given in Section 7.4.2.

2.2.3 Development of segmental discrimination

Naturally, children are able to detect other types of cues for distinguishing certain types of segments from others at very early ages. For example, voice-onset time (VOT)—the relative timing of the release of a stop closure with the onset of vocal-fold vibration—is perceived by very young infants, particularly when the difference in VOT crosses a phonemic boundary in the adult language (Eimas et al. 1971) but also when it does not (McMurray and Aslin 2005). Young infants are able to perceive a wide variety of fine-grained acoustic differences, including ones which are used distinctively in languages other than their own. However, between 6 to 12 months they gradually form categories of sound appropriate to their native language, sacrificing those fine-grain distinctions not used in their own language.

For example, at 6 months of age English infants discriminate English bilabial and alveolar place distinctions (/ba/ vs. /da/), Hindi retroflex and dental place distinctions (/ʈɜ/-/t̪ɜ/), and the Salish glottalized velar-uvular distinction (/kʔɜ/-/qʔɜ/), as well (Anderson et al. 2003; Werker and Tees 1984). As English infants hear more and more examples of English /t/ on both sides of the retroflex-dental boundary, however, they lose the ability to distinguish Hindi /ʈɜ/ and /t̪ɜ/. (Losing the (/kʔɜ/-/qʔɜ/) distinction takes longer, perhaps because /t/ is more frequent in English than /k/—see Anderson et al. 2003.) By 10-12 months, English

infants no longer distinguish [d] from an unaspirated [t] extracted from the consonantal cluster *st*, which they could at 6 months (Pegg and Werker 1997).[2]

Conversely, the ability to use fine-grain distinctions (even ones native to English such as /ba/ vs. /da/) in complex tasks such as matching two nonce words as labels for two novel objects, may not come till much later, around 17 months (Werker et al. 2002). Similarly, although infants may be able to recognize very common words (like their name) by four months of age (Mandel et al. 1995; Tincoff and Jusczyk 1999), they may not recognize two instances of a previously unfamiliar word form as being the same word when said by a different voice, or in a different pitch or speed, until 11 months (Houston and Jusczyk 2000, 2003).

Taken as a whole, these experimental findings suggest that, while infants may display an impressive degree of acoustic perception at remarkably early ages, including abilities that seem perfectly tuned for learning language (and speech segmentation in particular). Nevertheless, their abilities during the first year of life are still quite different from those of adults, which must be borne in mind when trying to understand how speech segmentation abilities develop during the second half of that first year.

2.3 Speech segmentation at 6-12 months old

2.3.1 Silence and one-word utterances

Among the auditory cues hypothesized to aid infants in segmentation, one of the most obvious, and most frequently cited, is the silence between streams of speech. Naturally, if audible pauses were *always* found between words (just as white space

[2]This is not to say that older infants and adults lose the ability to distinguish non-native contrasts completely. In certain situations many adults can still make such discriminations, but it may be harder than for native categories. See e.g. McMurray et al. (2002); McMurray and Aslin (2005) for further discussion.

is always found between words in English print), there would be no word segmentation problem. It stands to reason that if speakers vary considerably in the the frequency and salience of their natural pauses between words, the baseline difficulty of the task will vary accordingly. For this reason, some computational models (e.g. Cairns et al. 1997; Allen and Christiansen 1996) explicitly remove information about pauses from some of their testing conditions.

On the other hand, periods of silence (or pauses) are certainly a cue to *utterance* boundaries. Indeed, for some intents and purposes, they may be used as the definition of an utterance boundary. The transcribers of some speech corpora (e.g. Brent and Siskind 2001, in transcribing the Brent corpus) have operationally defined an utterance as a stream of speech bounded by a particular duration of silence (e.g. 300 ms) on either side. This dissertation will use the term 'utterance boundary' to refer to an acoustic cue marking the beginning or end of speech so obvious that it seems of little doubt that even a newborn could reliably detect it, and an 'utterance' to be a stream of speech by a single speaker between two such utterance boundaries. When referring to specific corpora, however, the definition of the transcribers will be assumed (with their definition noted when explicit).

Understandably, silence (or noticeable pauses in speech) has widely been rejected as a sole cue to word learning for (at least) three reasons: first, most words (and particularly function words, though children may not learn these until later anyway) seldom appear in isolation; second, words may sound sufficiently different when pronounced within an utterance from their isolated rendition as to be unrecognizable as the same word to a young infant; and third, so many other cues have been shown to be successful (both in models and in experiments with infants) that it would be foolish to ignore them. Brent and Cartwright 1996, add a

fourth reason: adults can obviously segment unknown words from multi-word utterances, and clearly they have to develop this ability sometime. While it might be argued that adults may segment utterances into words simply as a by-product of lexical access over the known words in the sentence, such an explanation ignores the question of how the adults built that vocabulary in the first place, given the three reasons stated above.

Nevertheless, certain researchers have argued that the importance of silence (or of one-word utterances) has been underrated as a basic bootstrap to word segmentation. While it is clearly not the only cue needed, the types of wordforms observed alone in one-word utterances may be acquired and serve as cues to help segment other words. This may well be true for a limited number of words: for example, Bortfeld et al. (2005) show that children can use their own names to segment off words that appear adjacent to them in an utterance. Brent and Siskind (2001) demonstrate a strong correlation between words found in one-word utterances and those learned earliest by children (determined by a receptive vocabulary test: Fenson et al. (1993, 1994)). Finally, silence (in signaling utterance boundaries) interacts with segmental, distributional cues to play a somewhat more nuanced role in speech segmentation—further discussed in Section 2.3.3.3.

2.3.2 Stress and other suprasegmental cues

Another important type of cue includes *suprasegmental* (or prosodic) cues—aspects of the speech signal that may range over multiple segments—such as rhythm or stress. Rhythm has been discussed in Section 2.2.2 above and loosely defined as the relative duration of segments (or larger units such as syllables). Other prosodic phenomena include relative prominence of segments or syllables (e.g. patterns of

stress or accent) and also intonation (the "melody" or overall pitch contour over an utterance).[3]

English, as a language with a high degree of variance in the durations of both "vocalic" and "consonantal" stretches of the speech signal, is often claimed (along with many other languages with similar properties) to use stress-units (or prosodic "feet") as its basis for segmenting words (see e.g. Demuth 1996). In addition, the types of cues associated with stress in English (duration, amplitude, vowel quality, and the likelihood of a pitch accent) are all assumed to be fairly easy for a baby to hear.

However, at this point some differences in terminology ought to be made clear: in defining the relationship between stress and word boundaries in English, "stress" has often been used as a cover term for a variety of (albeit strongly correlated) phenomena. Some scholars (e.g. Anne Cutler and colleagues) have distinguished between "strong" vs. "'weak"' syllables. These are distinguished primarily by vowel quality—a strong syllable is one with a "full" or unreduced vowel, though it may (or may not) also bear lexical stress or even a pitch accent. (For example, both syllables in 'pancake' are strong, even though only the first is stressed.) A weak syllable has a reduced vowel (e.g. /ə/, /ɨ/) and cannot be lexically stressed or receive a pitch accent.[4] Other scholars (e.g. Christiansen and Allen 1997) have simply referred to this as lexical stress, even when they have operationally defined it in terms of vowel quality.

[3]Naturally, these interact considerably: in English, lexical stress, pitch accents, and intonational markers of phrase boundaries all influence duration of segments and syllables; a syllable marked with some level of lexical stress may or may not receive a pitch accent, but is more likely to in child-directed speech than in adult-directed speech; pitch accents, in turn, influence the overall "melody" of the sentence. The exact details of these interactions are not addressed here, but be relevent to future research examining supersegmental cues, as briefley sketched in Section 7.4.

[4]If a typically weak syllable were so accented for e.g. contrastive stress, it would also feature an unreduced vowel, as in the /i/ in accented 'the' or the /e͡ɪ/ in accented 'a'.

Still others, e.g. Mattys et al. (1999), use the terms "strong" and "weak" to refer to primary and secondary stress, since all the syllables in their stimuli have full vowels. Morgan (1996) skirts the issue entirely by using full vowels throughout with no pitch accent and manipulating vowel duration alone. However, the effects all run the same direction: by the age of 9 months, "full" (unreduced) vowels, increased duration, and primary stress all seem to cause the stimuli that must then be interpreted as word-initial.[5]

In natural child-directed speech, where all these cues are free to occur together naturally, this tendency to interpret stressed syllables as word initial is observed as early as 7.5 months. Words like 'castle' are correctly segmented from natural speech, but words like 'guitar' are not (Jusczyk et al. 1999b). This may be referred to as the *trochaic bias*, borrowing the poetic term "trochee" for strong-weak patterns in prose. Two adjacent strong syllables (e.g. 'roll dice') are also correctly segmented as separate words at this stage (Mattys and Jusczyk 2001a).

While the trochaic bias is known to be a salient cue for word segmentation in Germanic languages, less is known about languages with different prosodic patterns. One study found that Canadian French-learning babies show an opposite (iambic) bias, segmenting out only weak-strong bisyllables as words (Polka et al. 2002). However, it must be noted that the distinction between "strong" and "weak" syllables is different in French than in English, and so these results must be interpreted cautiously.

[5]While sorting out these distinctions is not a primary purpose here, this chapter will use (unless otherwise noted) "strong" and "weak" to refer to syllables with full vs. reduced vowels, and "stress" to refer to the underlying specification of certain syllables of certain words as possible (and in child-directed speech, probable) bearers of a pitch accent.

2.3.3 Segmental cues

In Section 2.2.1, a segment was defined as a region of sound with a certain degree of coherence. A phone may be said (for the moment) to be a category of segments that sound sufficiently alike that they may be treated as essentially equivalent members of a class, and statistics tabulated over them (in some implicit way). A phoneme may be said to be a class of phones (or allophones) which adult speakers (as the infant may be gradually realizing) seem to treat as functionally equivalent (though they use different allophones in different places). Alternatively, a phoneme may be viewed as a bundle of distinctive features, which in some models (e.g. Christiansen et al. 1998—cf. Section 3.2.2) are assumed to be primitives. In other models (e.g. Cairns et al. 1997—cf. Section 3.1.2) they are assumed to be directly derivable from the acoustic signal.[6]

We have noted in Section 2.2.3 that infants are not yet adult-like in recognizing the phonemic categories found in their native language (sometimes seeming to over-generalize, and in other tasks to under-generalize). Nevertheless, it seems that on some level they are able to recognize that certain types of sounds rarely if ever appear together in a sequence, and others do all the time. The probability of hearing a particular sound B, given that one has just heard the sound A, is the transitional probability (TP) $\Pr[B|A]$, calculated as $\Pr[A, B] / \Pr[A]$. Further discussion of transitional probability and related statistical cues is found in Chapter 4.

[6]See Beckman (2003) and Werker and Curtin (2005) for a discussion of the emergent relationship between phones and phonemes before a full, adult-like lexicon is in place, and why using these terms may not always make much sense at the ages under discussion. Pierrehumbert (2003) also presents some strong arguments against this use of phonemes, or IPA-style phones for that matter, as categories, or units for building categories. As convincing as these arguments may be, we will adopt this definition for the time being, in order to describe the assumptions some researchers seem to be working under. It is worth noting that it is not always clear what assumptions are truly part of a researcher's model and what are simply made for convenience or simplicity of design.

Constraints that forbid two sounds (or phonemes) from appearing together in a particular order within a larger unit (say, a syllable or a word) are called *phonotactic constraints*. Two-phone sequences that violate phonotactic constraints are likely to have a low transitional probability. Conversely, other sounds sequences appear together more often than would be predicted by chance; they may have a relatively high transitional probability.

2.3.3.1 Phonotactics

Transitional probability may be used as a cue to word segmentation in two different ways. First of all, as mentioned above, transitional probability may be a cue to discovering phonotactic constraints in the language. For example, if an infant hears a sequence of two sounds that normally do not occur together in that order, for example, [zf] as in *has fixed*, it is very likely that the [z] and the [f] do not belong in the same word, so the child may posit a word boundary between them. Children apparently become aware of these tendencies sometime between 6 and 9 months: Jusczyk et al. (1993a) found that 9-month-old American infants listened longer to words whose sound sequences were commonly found within English words than to words with "illegal" phonemic sequences; 6-month-old infants showed no preference between the two lists. Mattys and Jusczyk (2001b) showed further that 9-month-olds who hear a novel word in a context that creates unlikely sequences around its boundaries (e.g. 'gaffe' in the word-sequence *. . . bean gaffe hold . . .*—[ng] and [fh] are not common English sequences) are better able to pull out that word than in a context with common sequences at the word boundaries (e.g. 'gaffe' within the sequence *. . . fang gaffe tine . . .* [ŋg] and [ft] being more common within English words).

2.3.3.2 Transitional probabilities

A second and more general way transitional probability can be used involves an old insight traceable back to Harris (1954, 1955), who observed that if one finds, at a particular point in a sequence of phonemes, that a large number of phonemes could come next, then one is likely to be at a morpheme (or for our purposes, word) boundary. Stated in terms more akin to modern uses of information theory, those positions in the sound stream where it is very hard to predict what sound should come next are said to have high perplexity. All else being equal, if a large number of phonemes could come next, then the probability of any particular one coming next is bound to be small. So points with low transitional probability (relative to surrounding points) are likely to be boundary points, even if they don't involve particularly unusual clusters or ones that violate phonotactic constraints.

This insight is the starting point for a wide variety of models, perhaps most famously Saffran et al. (1996b), who found that infants 8 months old could segment three-syllable pseudo-words in an artificial 'micro-language' in the absence of any other acoustic cue (e.g. stress, coarticulation,[7] and pauses). The more general notion of predictability is the driving force behind the "immediate task" in the computational models discussed in Chapter 3, particularly for Elman (1990) and Cairns et al. (1997), where higher error in predicting the next phoneme corresponds to lower conditional probability of that phone given the preceding one. It is also used directly as a baseline cue in Brent (1999)—and will be revisited at greater length in Sections 4.2.2 and 4.6.

[7] See Section 2.3.4.2 below for a definition.

2.3.3.3 Segments on the boundaries

Finally, segmental statistics can be anchored to utterance boundaries. Aslin et al. (1996) show that a multi-layer perceptron (MLP) trained to learn the probability of two- or three-phone sequences preceding the end of an utterance (modulo their representation of the phonemes within the MLP as bundles of phonological features) manages not only to learn to predict utterance boundaries, but also shows higher-than-average activation for utterance-internal word boundaries as well. Experimentally, they also show that both English- and Turkish-speaking mothers, when asked to teach a novel word to their infants, very frequently use a strategy of placing the novel words in utterance-final position. More recently, Seidl and Johnson (2006) have found more direct evidence that words next to boundaries (i.e. at the beginnings and endings of utterances) are easier to learn. More will be said about this cue in Section 4.5.

At first glance, it may seem odd that Aslin et al. (1996) did not also try to extrapolate beginnings of utterances to beginnings of words. Few studies have paid much attention to the beginnings of utterances as cues (one exception being Brent and Cartwright 1996). The Aslin et al. study may have missed this because of the uni-directional "prediction" paradigm common to connectionist approaches (cf. Chapter 3). However, Pierrehumbert (2003) suggests another reason: namely, that "human languages tend to have maximal contrast sets (maximal statistical perplexity) at the word onset" (p. 144). Utterances do not necessarily share this property, however: Many utterances will begin with determiners like 'the', 'this', 'that' or the existential 'there', which all share the phoneme /ð/. However, extrapolating that many words begin with /ð/ would be wrong in English—the better strategy is to skip over the 'the' and pay attention to what comes after it. Shi et al.

(2003) show that 8-month-olds do exactly that—they remember words better when they are preceded by 'the' or other common function words.

2.3.4 Subsegmental cues

2.3.4.1 Allophonic variation

Finally, we come to acoustic cues below the level of the phoneme. As was alluded to above, certain sounds (phonemes) that we think of as equivalent are actually pronounced quite differently in different contexts within a word, foot, or syllable. For example, /t/ is aspirated (pronounced with a puff of air and a long VOT) at the beginning of a word or stressed syllable, unaspirated after /s/ (in the same word), flapped foot-medially between vowels (i.e. after a stressed vowel and before an unstressed one), and often unreleased or even glottalized at the end of a word. For example: 'stuttering tot' may be realized as [ˈstʌɾəɹɪŋˈtʰatˀ] or as [ˈstʌɾəɹɪŋˈtʰaʔ(t)].

Similarly, /l/ is much more likely pronounced with the tongue tip touching the alveolar ridge at the beginning of a syllable than at the end of one: e.g. 'bright' [l] vs. 'dark' [ɫ] in [ˈlaʊdˀˈbɛɫ] 'loud bell'.

Although adults, partially influenced by their orthography, may consider the various forms of /t/ somehow to be manifestations of "the same sound" (and this author confesses to struggling against "hearing a /t/" when listening to a word-final glottal stop in Arabic), for preliterate, prelexical children, there is no reason to perceive these segments as the same. We should expect them to be treated as distinct entities (see Pierrehumbert 2003). However, the fact that [tʰ] often appears word-initially, whereas [t] and [tˀ] never do, makes [tʰ] a possible cue for a word boundary immediately before it, and [t] and [tˀ] generally good cues against

an immediately preceding word boundary. (See Church 1987 for a thorough treatment on the possibilities of using these and other allophonic variants as cues for word segmentation.) A famous example of phonemically identical sequences that are commonly disambiguated by these allophonic variations is the example of the 'nitrate' (pronounced [n͡aɪt(ʰ)˳ɹ͡eɪtˀ] vs. 'night rate' (pronounced [n͡aɪʔ(t)ɹ͡eɪtˀ].

Experiments testing allophonic effects on segmentation are rarer than for other cues. Hohne and Jusczyk (1994) show that infants can hear the difference between 'night rate' and 'nitrate' by 2 months of age. However, while infants may be able to use allophonic variation in conjunction with other cues for segmentation at 9 months, allophony is not a sufficient cue on its own until 10.5 months (Jusczyk et al. 1999a).

2.3.4.2 Coarticulation

When two phonemes appear together in a word, people naturally seek to move their articulators from one phoneme's vocal-tract configuration to the other as efficiently as possible. This changes the realized pronunciation of each phoneme. For example, a /k/ said before an /i/ (as in 'key') is pronounced (in English as well as many other languages) with the tongue contact considerably closer to the front of the mouth than the /k/ before /u/ (as in 'coo') (Ladefoged 1993). Curtin et al. (2001) show that 7-month-old infants do not recognize words with the wrong coarticulatory cues, although adults do. This finding provides further evidence that children do not categorize speech sounds in the same way adults do. However, a recent study suggests that while adults do not use perceived coarticulation for segmentation in either "clean" speech or extremely noisy conditions, they may use it in mild noise (Mattys et al. 2005).

2.4 A timeline for word segmentation cues

The preceding information is summarized in a short timeline, showing the various cues at the ages at which they have been experimentally observed, with crucial changes in the relative strengths of cues marked in bold. A more detailed review is available in Werker and Curtin (2005).

6 months

- Infants are not yet able to segment novel words from the speech stream (Jusczyk and Aslin 1995).
- English-learning infants do not yet demonstrate a preference for listening to lists of trochaic words (Jusczyk et al. 1993b).
- English-learning infants are able to be primed to hear either trochaic or iambic rhythmic patterns as cohesive. However, they are not yet "fixated" on trochaic patterns Morgan and Saffran (1995).

6.5 months

- When stress and transitional probabilities are pitted against each other, transitional probabilities override stress (Thiessen and Saffran 2003).

7 months

- Shifting the stress to another syllable disrupts recognition of previously familiarized words (Curtin et al. 2005).

7.5 months

- Infants are able to segment words from the speech stream (Jusczyk and Aslin 1995), including two adjacent strong syllables as separate words (Mattys and Jusczyk 2001a).

- English-learning infants at this age successfully segment only strong-weak (SW) words, showing a trochaic bias (Jusczyk et al. 1999b).
- Canadian French infants segment only weak-strong (WS) words, showing a language-appropriate iambic bias (Polka et al. 2002).

8 months

- Infants can segment words out of child-directed speech, but not out of adult-directed speech (Thiessen et al. 2005).
- Infants are able to segment three-syllable pseudo-words from an artificial micro-language, using segmental transitional probability cues alone (Aslin et al. 1998; Saffran et al. 1996b).
- English-learning infants are better able to segment and remember novel words when they are preceded by frequently occurring function words such as 'the' (Shi et al. 2003).
- Stress and coarticulation each override conflicting transitional probabilities (Johnson and Jusczyk 2001), reversing the trend found in infants 6.5 months of age (Thiessen and Saffran 2003).

9 months

- English-learning infants prefer to listen to trochees generally, treating them as single units (Echols et al. 1997; Jusczyk et al. 1993b; Morgan 1996).
- Segmentation based on phonotactic and allophonic cues is first observed (e.g. Friederici and Wessels 1993; Mattys and Jusczyk 2001b).

10 months

- Infants perceive two-syllable words with common transitions between syllables (e.g. [ŋk] as in 'monkey') as more cohesive than bisyllables with infrequent inter-syllable transitions (e.g. [pt] as in 'reptile') (Morgan 1996).

11 months

- English-learning infants no longer missegment words with an iambic (weak-strong pattern)—apparently able to combine metrical knowledge with other types of cues, such as phonotactics and allophonic variation (Jusczyk 1997).

CHAPTER 3

OVERVIEW OF PREVIOUS CONNECTIONIST MODELS

In Section 1.2 a number of models for the word segmentation task (WST) are discussed. This chapter focuses on the connectionist strand of models. Section 3.1 gives a brief historical overview of various connectionist models for the WST. Section 3.2 discusses Christiansen's model specifically. Section 3.3 discusses two facets of the connectionist models that distinguish them from other proposed models here reviewed.

3.1 Early connectionist models

3.1.1 Elman (1990)

The first widely known connectionist model of the WST is found in Elman (1990). He developed a novel type of simple recurrent (neural) net (SRN) which allowed for an implicit representation of memory within the net's architecture.

These nets, which have come to be known as Elman nets, resemble so-called Jordan nets (Jordan 1986) in that they utilize as their base a MLP with one hidden layer, and then add *recurrent* connections which allow activations from one time-step to influence the hidden layer of the next time-step. In Jordan nets, these recurrent connections copy the activation levels of the output units (via connections with fixed weights of 1.0) to a set of "state units" which then feed into the

next step's hidden layer. In Elman nets, the hidden layer itself is copied to a set of "context units" which then also feed into the next step's hidden layer. The network has an equal number of hidden nodes and context nodes. At each time-point in the training (corresponding to the duration of a single phone segment, not to a constant time interval) the values from the hidden nodes are copied (again, via connections whose weights are fixed at 1.0) to the context nodes. The weights on the connections from the context nodes back to the hidden nodes are not fixed, but are subject to training like other arc weights.

Elman (1990) then trained these nets on a number of prediction tasks. In one of these tasks, an artificial mini-language with one-syllable "words" of various length (all corresponding to the regular pattern $//CV+//$, i.e., one consonant followed by one or more vowels) were presented to the net, one phone per time-step.[1] The net was then asked at each step to predict the identity of the next phone. (Phones were represented in both the input and the output as vectors of five binary, non-independent phonological features.) Using prediction tasks such as these, rather than training directly on the segmentation task itself, is crucial to the model's plausibility, as well as to the plausibility of other prediction-task models. The target values for the chosen prediction task are assumed to be fully observable at the next time-step, allowing the model to note its errors and train itself accordingly (using some form of back-propagation). If the model had been trained directly on the word boundaries themselves (which are assumed *not* to be consistently observable), the "self-supervision" of this network would not be plausible.

[1]These time-steps do not correspond to any consistent time scale, but are assumed to vary according to the duration of the phone. In Elman's model, each phone takes one time-step, no matter how long its actual duration is.

The network displayed greater uncertainty (as measured by the root mean squared error) at the ends of words than within words, just as might be predicted by Harris' original observation regarding spikes in "successor count" marking morpheme boundaries. Elman also constructed a mini-language with sentences generated from a small subset of real English words and a highly constrained grammar, with similar results.

3.1.2 Cairns, Shillcock, Chater, and Levy (1997)

Cairns, Shillcock, Chater, and Levy (1997) (henceforth CSCL97) built on Elman's approach, using a more elaborate set-up and including a more detailed analysis of their findings, so as to shed more light on certain scientific questions of the WST.

Method Like Elman (1990), CSCL97 reported neural networks trained on a derived prediction task rather than directly on the word boundaries. However, the derived task used is not merely the prediction of the next phone, but also includes remembering the previous phone and reconstructing the present phone in the presence of artifically created "noise" caused by randomly changing the value of the binary phonological features. It is not clear why CSCL97 used this more complicated task, since they calculated the likelihood of a word boundary between the current and upcoming phones the same way as Elman did: by observing the error in predicting the next phone only (although CSCL97 measured this error using cross-entropy on the next-phone prediction output units rather than using Elman's root mean squared error). The difficulty of remembering the previous phone at the next step was not used, though that might have been an interesting additional cue.

Since CSCL97 did not have access to a sufficiently large corpus of child-directed speech (CDS), they used a corpus of adult-directed speech (ADS), the

London-Lund corpus (Svartvik and Quirk 1980). One might expect ADS to be more challenging than CDS, having a larger vocabulary, longer words and utterances, and greater complexity generally. For example, the London-Lund corpus has on average six contiguous words between pauses or turn changes, while the mean length in the Korman corpus is only three words long. (See Section 6.1.1 for more details on the Korman corpus.) This must be borne in mind as results are considered.

Results Results were reported for two conditions: a "worst-case" scenario where pause and speaker-change information (essentially all utterance boundaries) were stripped out, and a more realistic scenario where these were added back in and automatically assigned word-boundary status (on the assumption that children have no trouble parsing pauses as utterance boundaries). In both testing scenarios, the input was presented segment-by-segment as feature bundles, though without the random bit-flipping "noise" used in training.

In the worst-case scenario, the model only identifies 21% of the boundaries correctly, with a 1.5:1 hit:false-alarm ratio. If these results seem to be somewhat unsatisfying, one may note that when pauses and speaker changes are restored, the model improves to 32% completeness with a 2.4:1 hit:false-alarm ratio. Furthermore, the model evidences interesting patterns, in that it is more successful at finding open-class words than closed-class words. Although the model does not correctly extract more words than chance would allow, it does favor open-class words at a greater-than-chance level. (Only 40% of the word tokens in the corpus are open-class, but 59% of the words correctly extracted are open-class.)

Discussion The point of CSCL97's simulation was to demonstrate that a RNN system with an architecture similar to Elman's (1990) net (though more elaborated)

could segment a corpus of actual language in a way that roughly corresponds to what we might expect of children. The fact that the model is more successful at finding open-class words than closed-class words is consistent with this claim.

Another interesting property of CSCL97's system is that it appears to learn the phonotactics rules of English by generalizing the patterns it observes in the corpus, particularly as they apply to syllable structure. Although its success at finding multi-syllabic words and function words was not particularly striking, around 80% of the proposed boundaries were phonotactically legal syllable boundaries, according to the patterns found in the corpus. In fact, the reported false-alarm rate was statistically equivalent to what a syllable detector operating at 80% precision would have obtained. These results suggest that the strategy adopted by CSCL97—training the net to predict the next phone—is better suited to learning syllabic structure and segmentation than full-blown *word* segmentation. It seems that, for adult-directed English at least, the difficulty of predicting the next phone (given the cues they used) is the same for either type of syllable boundary, whether word-internal or across words. Of course, given that a large majority of English word tokens are monosyllabic (See e.g. Pitt et al. 2005, Figure 1), learning syllabic structure is a significant step that may be expected to go a long way towards bootstrapping the WST. For example, it could in principle feed into a model more directly approximating Saffran and colleagues' experiments (such as Swingley 2005), where syllable boundaries are the primary unit.

However, for languages with complex and sometimes ambiguous rules for syllabification such as English, finding *phonotactically legal* syllable boundaries is not the same as finding the *actual* syllable boundaries—i.e. those that coincide with the boundaries of the intended words. There may be several legal boundary points between two given syllables (e.g. 'I scream', 'Ice cream', or even 'Aysk ream', filling

an accidental gap in the language with a phonotactically plausible nonce word). Determining which of the legal boundaries is in fact the intended boundary point is most directly achievable by paying attention to subsegmental cues such as the allophonic realization of the /k/ (aspirated, unaspirated, or unreleased) and degree of coarticulation between elements. (See Section 2.3.4.1, Church 1987 for further discussion.) A sufficiently detailed or "narrow" phonetic transcription may preserve these as differences by representing 'I scream', 'Ice cream', and 'Aysk ream' as distinct strings. But a system that represents them identically has ignored a crucial cue to word segmentation that children have available to them (cf. Hohne and Jusczyk 1994).[2]

Despite the deficiencies in the model, perhaps brought about by the inadequacy of the corpus' transcription and other issues of input representation, the CSCL97 model constitutes an important milestone in testing the abilities of RNN-based models on corpora of actual language, as opposed to the artificially generated corpora often used up until that point. The CSCL97 artificial neural network (ANN) is particularly effective at finding both the onsets of open-class words and the onsets of "strong" syllables (which CSCL97 operationally defines as syllables with vowels other than /ə/, /ʌ/, and /ɪ/). From this last result, Cairns and colleagues claim that their model is in principle able to bootstrap the trochaic bias called for in the English version of Cutler's metrical segmentation strategy (MSS). This it does just from the difficulty of predicting the first segment of the strong syllable's onset—not from observing the identity of the vowel (unless the vowel itself is syllable-initial). The prediction that follows from this is that, if an infant's

[2]For those languages that have a tendency, to a greater or lesser degree, for resyllabification, such as Spanish, the acoustic cues that signal a syllabic boundary may not match up as consistently with the lexical boundaries. Hence, syllabification may not be as directly helpful for word segmentation in these languages. However, since little is known about how word segmentation proceeds in these languages at this point, this must remain a caveat worthy of future investigation.

segmentation system is at least as good as the one modeled here (and likely better, having more cues with which to work), then the trochaic bias for the MSS can be learned from statistical phonotactic cues alone. This account of the initial bootstrapping of a trochaic-biased metrical segmentation strategy from English statistical, segmental/featural cues, is an interesting one, particularly in light of later findings by Thiessen and Saffran (2003).

3.1.3 Aslin, Woodward, LaMendola, and Bever (1996)

Whereas Elman (1990) and CSCL97 (in the tradition of Harris 1955) focus on the predictability of the following phone, the Aslin model (Aslin, Woodward, LaMendola, and Bever 1996, henceforth AWLB96) examine a different distributional cue: namely, the segmental context immediately preceding utterance boundaries. Since very few utterances end in the middle of a word, it stands to reason that utterance boundaries are properly a subset of word boundaries. Since utterance boundaries are presumed to be immediately accessible to children (having the fairly obvious cues of a clear intonational boundary followed by some substantial period of silence) but many other word boundaries are not, it seems worthwhile to investigate how much generalization is possible from utterance boundaries to all word boundaries.

Obviously, utterance boundaries do not constitute a completely representative sample of word boundaries: many types of words (e.g. "proclitic" function words such as 'a', 'the', or 'your', which typically precede a noun or the rest of a noun phrase) occur exceedingly rarely, if ever, at the end of an utterance. However, these are not the words we expect infants to be learning first anyway. Besides, as with the cue above (which is more properly a cue for syllables or morphemes), it

is not necessary that the cue be faultless, simply that it be good enough to bootstrap from in certain cases.[3]

In fact, AWLB96 found in their experiments that when mothers are asked to teach their children specific words, they often place them utterance-finally, so that the end of the word is clear. (Other cues, such as an initial strong syllable, perhaps aided by a deliberate stress on the word, may provide cues to the word's start.) However, the point of their connectionist model was to demonstrate that the segmental, distributional cues near utterance endings can be usefully generalized to other word boundaries, even in absence of other cues.

Method Rather than using a SRN, Aslin and colleagues used a feed-forward MLP with previous context represented with a fixed "window in time" of one, two, or three segments of context. This network consisted of 19, 37, or 55 input nodes (18 phonological features for each phone of within the context window, plus one marking the utterance boundary) and a single output node, trained on the presence or absence of an (immediately) upcoming utterance boundary. (Utterance boundaries were operationally defined as pauses at least one second in duration.) The number of hidden units varied from 20 to 100; increased numbers of hidden units did not improve the model's performance. Due to the small size of their corpus, they trained on two-thirds of the total corpus (using two or three passes), and tested on the remaining third.

[3]The case of Cantonese, where utterances are usually ended by a very restricted set of discourse particles, may constitute a case where the utterance-boundary extrapolation cue does not hold, or at the very least would need additional context to function. (I thank Mary Beckman for this observation.) If such particles are just as frequent in CDS as in ADS in Cantonese, then this might be taken as evidence against this cue's universality. One appropriate tactic that infants may learn is to strip off these ending particles before using a cue such as AWLB96 propose. However, such a proposal, taken to its logical conclusion, is precisely the strategy embodied in the INCDROP model, which has been successfully applied to Chinese text segmentation (Brent and Tao 2001).

# of Segments	Precision	Recall	F-score
Three	74%	62%	68%
Two	70%	53%	60%
One	51%	35%	41%
Random	25%	5%	8%

Table 3.1: Results reported in Aslin et al. (1996, adapted from Fig. 8.8, Page 132)

Results Table 3.1 shows the results reported for boundary detection for an MLP with 30 hidden units (as estimated from their bar graph, Figure 8.8, p. 132). From these results, AWLB96 claimed that two- and three-segment sequences (but not single segments) are sufficient for their system to learn boundary locations, though their threshold for determining sufficiency is not clear, given that the network given a single segment context also performed above their random baseline.

AWLB96 also argue for the importance of encoding phonemes with a phonologically motivated feature set. They compared a network trained on such input with one trained on input using an arbitrary encoding scheme based on alphabetical order rather than features motivated by phonological theory. The former clearly (and unsurprisingly) performs better. However, they did not explicitly test a feature set that assigns a separate input node to each phoneme, a more plausible control condition for testing the importance of phonological features. While input vectors based on phonological features may still be preferred on grounds of plausibility (assuming that a child's set of features is the same as adults', or at least the same as the phonologist's), the demonstration in AWLB96 is not fully convincing. The question of appropriate input representations will be discussed further in Section 3.3.2 and in Chapters 4 and 6.

3.1.4 Summary of the three previous models

The three models reviewed in this section capture aspects of the WST, focusing on particular cues thought relevant to the problem. While one may detect a certain progression of increasing sophistication and realism, none of them are completely satisfactory. Elman (1990) does not report results for actual language, but only for artificially created corpora. CSCL97 does use an actual corpus of natural language, but not child-directed speech. Furthermore, while it performs above chance on finding content words, and shows qualitatively a plausible type of behavior, its mechanism seems overly complicated for the results obtained. Aslin et al. (1996) shows promising results for a very simple model but does not show how that model relates to Elman's original work or the findings in CSCL97. It also makes claims about phonological features that are not substantiated by a thorough and fair examination of alternatives.

The progression of connectionist models naturally leads to an influential model that examines the interaction of multiple cues to word segmentation. While it does not address all of the issues raised in the preceding paragraph, it does provide a model that simplifies some of the unnecessary complications found in CSCL97 (i.e. jettisoning the extra tasks of recalling the current and previous phone) and is arguably more successful in its performance on the task.

3.2 The Christiansen model

The Christiansen model combines both of the distributional cues discussed above: namely, Harris' cue as implemented by Elman (1990) and CSCL97, and Aslin's utterance-boundary cue. It does so most simply in an early study (Allen and Christiansen 1996, henceforth AC96), by using an artificial mini-language reminiscent of

the mini-languages that Saffran and colleagues use as stimuli. In the most seminal study (Christiansen, Allen, and Seidenberg 1998, henceforth CAS98), an additional cue corresponding to lexical stress is added to the model, which is then trained and tested on a corpus of actual child-directed speech. A companion study (Christiansen and Allen 1997, henceforth CA97) examines the effect of variation in the input on the model's performance.

In addition to these three works, a number of later studies have applied the model to morphological and syntactic aspects of acquisition and language evolution, including grammatical categorization (Reali et al. 2003), case and word order (Lupyan and Christiansen 2002), subjacency (Ellefson and Christiansen 2000), and other aspects of grammar (e.g. Christiansen and Dale 2001). Christiansen, Conway, and Curtin (2000) showed that the Christiansen word segmentation model can account for certain empirical results thought to challenge the connectionist approach (Marcus et al. 1999). These later studies demonstrate that the Christiansen model is a viable computational model not only for the WST but for a variety of subtasks within the general process of language acquisition. It is primarily in these later stages of language acquisition that it has received its most recent development; however, only the first three papers dealing directly with the WST will be immediately relevant to the present study. The next three subsections detail each of them in turn.

3.2.1 Allen and Christiansen (1996)[4]

As mentioned above, the AC96 model represents the combination of the original Elman WST model and the Aslin model in its most basic form. As in the Elman model, the network used is an SRN (specifically, an Elman net), the input is a single

[4]This section is based on the introduction found in Rytting (2006a).

segment and the target task is to predict the identity of the next phone. This target task is "self-supervised" in that the next input provides immediate feedback as to what the preceding output should have been. Also like Elman (1990), the "words" in this mini-language consist of a very simple pattern of alternating consonants and vowels—although instead of the one-syllable /*CV*+/ pattern, a three-syllable /*CVCVCV*/ pattern is used. The crucial difference (which brings in the "Aslin cue") is that the set of possible symbols from which the next segment is predicted includes not only consonants and vowels, but also an utterance boundary marker (UBM), which may be thought of as the silence marking the end of a turn.[5]

Conceptually, if the task of predicting utterance boundaries is thought of as an additional task, separate from predicting the next phone (rather than as just an additional symbol within that prediction task), then one may think of the interaction between these two tasks as a novel property of the network. Multiple output nodes used to train the network on helpful, related tasks are referred to as "hints" or "catalyst" nodes (see Suddarth and Kergosien 1990). The intuition (following Aslin et al. 1996) is that since utterance boundaries are a subset of word boundaries with more or less representative properties of word boundaries generally (which is quite accurate in this mini-language devoid of any syntax), the utterance boundary output unit should be activated not only at utterance boundaries, but (to a lesser degree) at all word boundaries. This enables the AC96 model to measure the SRN's level of expectation for an upcoming word boundary not by using the network's error (whether measured with cross-entropy or root mean squared error) but simply by using the activation of the UBM output node, following the

[5]Of course, in real human dialogue, there are many cues besides silence which may mark the end of a turn or utterance—e.g. prosodic markers such as "boundary tones", gestural cues, and so forth—and of course many (adult) speakers cut short the silence with overlap or even interruptions. However, for these experiments a convenient simplification is to operationalize utterance boundaries by defining them as a silent pause of some pre-specified duration.

Aslin model. Above-average activation of the utterance boundary output unit signals an expected word boundary; below-average activation signals the expectation that the word continue. The other output units, corresponding to phones, can be thought of as catalyst units to the utterance boundary unit; while they are trained through back-propagation just like the UBM node, their output activation is not used at test time.[6]

A major point of AC96 is that word boundaries can be learned in this way only for languages with particular statistical properties. It is straightforward to construct artificial languages that lack these properties. Specifically, for a language where each syllable is equally likely to follow any other (or to end a word or an utterance), the network can only learn syllable boundaries. But if certain segments or groups of segments (such as syllables) are more likely to appear at the ends of words than others (as seems to be the case with most human languages), and a sufficiently large sample of the language is available to the learner, then the network can learn to distinguish word boundaries from word-internal syllable boundaries.[7]

Method To show this point, Allen and Christiansen constructed two artificial "mini-languages" to train the net: a "variable transition probability" (VTP) language and a language with "flat" transitional probabilities between syllables. Each

[6]This may lead one to wonder if the UBM is superior in all respects to the error measurements used by Elman and CSCL97, or if some voting combination both cues would improve performance further. However, such an examination is never mentioned, and is beyond the scope of this study.

[7]Naturally, for a language where words are almost exclusively monosyllabic, such as Cantonese, distinguishing between word boundaries and word-internal syllable boundaries is not a concern. Still, it seems highly likely in such a case that various syntactic and semantic restrictions, even if these are statistical tendencies rather than hard-and-fast rules, will still create sufficient variation in transitional probability to allow an SRN to acquire dectect these patterns and generalize them to discover interesting aspects of syntactic structure (cf. e.g. Reali et al. 2003, for relevant extensions into syntactic acquisition).

of these languages used the same twelve syllable types: /b/, /d/, /p/, or /t/ followed by /a/, /i/, or /u/. Each language consisted only of three-syllable, six-segment ($CVCVCV$) words. The "flat" language contained 12 such words constructed to maintain a word-internal transitional probability of 0.667 from one syllable to another; the VTP language used 15 words following different restrictions: for example, no word begins in /b/ or ends in /u/.

Both languages were trained using a SRN with 8 units each in the input and output layers and 30 each in the hidden and context layers. Word boundaries were not explicitly marked, but utterance boundaries (corresponding to the last input unit) were placed at intervals ranging from 2 to 6 words long. Training was done for seven iterations over a corpus with 120 instances of each word in the mini-language; testing was done on a corpus without marked utterance boundaries.

Results For all experiments, the dependent measure was activation of the utterance boundary output unit. On the VTP language, the network predicts a significantly higher activation for word-boundary positions than for word internal positions, including other syllable-boundary positions. The network trained and tested on the flat language showed higher activation at syllable-boundary positions, but these did not differ significantly between word-boundary and word-internal syllables. In addition, the network trained on the VTP language only performed above chance when catalyst nodes were used; when the UBM node was the only output node in training, performance was similar to that in the flat language.

While these findings may seem of themselves to be simply a minor variant of the original Elman and Aslin models, they establish "catalyst" nodes as a viable method of combining multiple cues relevant to the task. This is a crucial point, for all models up until this time (whether connectionist or otherwise) had relied on a

single cue or heuristic. It is still far from obvious how to combine multiple cues in (for example) a minimum-description-length model. The next section will show how this applies to actual examples of child-directed language.

3.2.2 Christiansen, Allen, and Seidenberg (1998)

Method Since the AC96 model was designed for and tested on a mini-language only, the SRN used was relatively simple. In order to expand the model for use with real English data, a larger network was needed. CAS98 tested a number of different models with different combinations of cues. The main model, which incorporated all three cues (phonological, UBM, and lexical stress), used a 14-80-39 SRN. Of the 14 input nodes, 11 were used for phonological features, two for marking (primary and secondary) lexical stress, and one for signaling an utterance boundary. Of the 39 output nodes, 36 are used to identify the phoneme (1 node each), two for predicting the lexical stress, and the last node for predicting an upcoming utterance boundary. A later study (Christiansen et al. 2000) modifies the basic structure of the CAS98 SRN by adopting a larger feature scheme (17 features rather than 11) for encoding the phonemes, for a total of 20 input features. A full chart mapping these features to the MRC phoneset is found in an expanded, book-chapter version of this article (Christiansen, Conway, and Curtin 2005, henceforth CCC05).

The CAS98 network was trained and tested on a corpus of child-directed British English collected by Korman (1984) and found in the CHILDES collection (MacWhinney 2000). More details about this corpus are given below in Section 6.1.1; a few brief details are given below as they directly relate to CAS98's use of the corpus for their training.

The Korman corpus consists of word-level transcriptions, given in standard orthography, of spontaneous speech from parents to their infants. To prepare this corpus for use in their network, Christiansen and his colleagues transformed these word-level transcriptions into phonemic transcriptions using the canonical pronunciations of the transcribed words as listed in the MRC lexicon (Wilson 1987). They represented each phoneme as a vector of binary phonological features.

The CAS98 net was trained on a single pass of the training corpus (90% of the Korman corpus) with a learning rate of 0.1, momentum of 0.95, and initial weight randomization ranging from -0.25 to 0.25.[8] The SRN was tested on the remaining 10% of the Korman corpus, as well as on a set of words and pseudowords not directly relevant here.

Results In evaluating the model in terms of precision and recall calculated over word boundaries, CAS98 stipulate (following AC96 and AWLB96) that a greater-than-average activation of the utterance boundary output node at a given point in time indicates the prediction of a word boundary. When the activation for the utterance boundary output node is greater than the mean for that node (as calculated across the whole corpus) at a lexical boundary, a "hit" or true positive is recorded. Similarly, a lower-than-average activation of this unit at a lexical boundary constitutes a "miss" or false negative, and an above-average activation at any other location constitutes a false positive.

The performance of the CAS98 model using this evaluation metric is shown in Table 3.2. The conditions that combined the cues of phonological features and utterance boundary information performed quite well, with and without the stress

[8]It is not explicitly mentioned whether the output nodes are competitive (as with a "softmax" output), but one may assume not, since both phoneme- and stress-predicting output nodes are expected to fire for segments in stressed syllables.

Training Condition	Words		Boundaries	
	Prec.	Recall	Prec.	Recall
phon-ubm-stress	42.71	44.87	70.16	73.71
phon-ubm	37.31	40.40	65.86	71.34
stress-ubm	8.41	18.02	40.91	87.69
utterances as words	30.79	10.15	100.00	32.95
pseudo-random	8.62	8.56	33.40	33.15

Table 3.2: Percent precision and recall for the three nets trained with the utterance boundary cue, for an algorithm that treats utterances as words, and for a pseudo-random algorithm that predicts lexical boundaries given the mean word length (from Christiansen et al. (1998), Table 3)

cue. The condition without phonological information performed noticeably worse, but still better than the two baseline conditions.

3.2.3 Christiansen and Allen (1997)[9]

The CAS98 model articulates a plausible explanation for how children may combine cues of limited provenance in order to learn word boundaries with greater accuracy than they could with any single cue or heuristic. However, it was tested with clean input only. That is, all the observable cues used as input for detecting word boundaries (i.e., phoneme identities, level of stress, and utterance boundary locations) were provided free of any errors or variation.

Since the input is derived by transforming a word-level transcription into a string of phonemes, any variation in the actual speech signal below the level of the word is abstracted away from the input presented to the SRN. One reason for this

[9]This section is based on the introductory section found in Rytting (2006b).

Corpus	Utterances	Inter-pause Regions*	Word Tokens	Word Types*
training	1,438	3465	10,371	1,682
test	159	376	1,147	457
total	1,597	3841	11,518	1,770

Table 3.3: Size of the training and test corpora in terms of utterances, stretches between pauses, and word tokens (cf. Christiansen and Allen 1997). Asterisked values calculated by the present author.

abstraction was due to necessity: at the time of their study, no phonetically transcribed corpus of CDS was available. Nevertheless, a parallel study, Christiansen and Allen (1997), or CA97, addresses the issue of natural variation in speech.

Method The CA97 study used as input a portion of the Carterette and Jones 1974 corpus which provides both an orthographic and a phone-level transcription of informal speech between adults. CA97 use both of these transcriptions in turn, referring to them as the "Citation Form" and "Coarticulation" conditions, respectively. Since pauses were also transcribed in the corpus and used in their study, they were also used as a cue for the network, being marked the same way as utterance boundaries were. (Since pauses were marked only in the phonological transcription, an equal number of pauses were added at randomly chosen word boundaries in the Citation Form condition.) Although the mean utterance length may be calculated as 7.2, the mean number of words between pauses is only 3.0, in line with the mean utterance length for the Korman corpus. Table 3.3 shows the size of the Carterette and Jones corpus, showing a $90\% - 10\%$ split between training and test corpora like that used by CA97, with values for the pauses and word types calculated by the present author.

In addition to an initial experiment comparing the Citation Form and Coarticulation conditions, a second experiment was conducted in which some small amount of artificial noise was added to the activation levels of certain features. Unlike CSCL97, however, only "peripheral" features of each phoneme (defined as those features whose change would not result in another phoneme found in the language) were flipped . For example, [*voice*] is a "core feature" to /b/, since a change in it results in a /p/; however, [*continuant*] is considered "peripheral" to /b/ since the changing [*continuant*] on /b/ would result in a bilabial fricative, which is not among the phonemes of English, but would (presumably) be perceived as an odd mispronunciation of /b/ (or a closely related phoneme). The input values for these peripheral features were changed at random with a probability of 0.1, resulting in at least one feature being flipped at a probability of about 0.18 for a given segment and 0.41 for a given word (as calculated for a word length of three segments; the mean word length was 3.22 segments). No phones suffered deletion, insertion, or substitution with another canonical phone except as recorded in the human-transcribed corpus. This approach to modeling subsegmental variation is discussed further in Section 5.1.2.

Since the Carterette and Jones corpus contained no marking of realized stress, the stress cue was operationalized simply using vowel quality: all vowels transcribed as schwa (/ə/) were marked as unstressed. All other vowels were marked as stressed. (Only one unit corresponding to stress was used; there was no unit for secondary stress in this model.) In essence, stress here is hardly a separate cue, but simply another (possibly redundant) phonological feature for vowels. The authors note that the stress cue, even with this operationalization of it, improved performance significantly; however, as no results without the stress cue were reported, its exact contribution for this corpus cannot be evaluated.

Results Christiansen and Allen report no significant differences between the Citation Form and Coarticulation conditions in either of their experiments. Furthermore, the degradation caused by the bit-flipping of the "peripheral" features was not significant. While it is possible that this null result arises partially from the small size of their corpus, it nonetheless suggests that the sort of variation they added does not greatly affect the model's performance—and certainly it does not cause catastrophic failure of the model. Their reported results for the first experiment (without added subsegmental variation) and the second experiment (with peripheral feature bit-flipping at 0.1 probability) are shown in Table 3.4. (Precision and recall are reported only for words, not for boundaries.)

However, since CA97 did not include baseline figures for the Carterette and Jones corpus (as was done for the Korman corpus in CAS98) it is not certain how much of this robustness depends on the relatively frequent occurrence of pauses. Baseline figures using these pauses, but no phonological information, would be helpful in knowing how large a role is really played by the crucial aspects of the model—the interaction of the phonological, UBM, and stress cues.

3.3 Two facets of the connectionist models

Besides those elements that are inherent to the architecture of ANNs generally, there are two facets of the strand of connectionist models of language acquisition (and WST specifically) represented by Elman (1990); Cairns et al. (1997); and Christiansen et al. (1998) which distinguish them from the other types of models discussed in Section 1.2 above. One (which distinguishes these models from the Aslin model and its predecessors) is recurrence, and particularly the Elman net's use of recurrence to model temporal memory. The second, also shared by AWLB96, is the

Training	Words	
Condition	Prec.	Recall
Simulation 1:		
No subphonemic variation		
Citation Form	24.33	40.24
Coarticulation	25.27	37.07
Simulation 2:		
Subphonemic variation (0.1 prob.)		
Citation Form	23.35	37.72
Coarticulation	23.99	34.48

Table 3.4: Percent (word) precision and recall from (from Christiansen and Allen 1997)

use of sets of binary (or near-binary) phonological features in the input representation, as opposed to purely symbolic or "atomistic" input (that is, input where each segment is represented by means of a symbol that has no internal representational structure, as in e.g. the string-discovery models of Brent 1999) on the one hand, or purely acoustic input (as in e.g. the syllable onset detector described in Wu et al. (1997) or more recent work in Gold and Scassellati 2006) on the other.

These two design elements are of course not without theoretical import. For the use of recurrence, Elman (1990) argues for the superiority of SRNs over window-in-time (WIT) MLPs, using essentially two points: first, the greater elegance in representing time implicitly rather than explicitly, and second, the ability of SRNs (in theory at least) to represent the gradual decay of memory over time, instead of the (rather implausible) sharp cutoff of short-term memory implied by the time window of WIT MLPs. Indeed, the exact amount of memory needed to represent and solve a particular problem is learned in the SRN's training; thus, the model of memory that emerges is intertwined in the nature of the task at hand. While both the Cairns and Christiansen models use the SRN variant that Elman

pioneered, it is not clear whether they espouse Elman's claims concerning temporal memory (and the best or most elegant way to represent it abstractly), or whether Elman nets are simply a convenient, but non-essential, way of dealing with the need to consider some ill-defined (and theoretically unbounded) amount of phonological context.

For AWLB96, the ambivalence was made a bit more explicit: SRNs were acknowledged as a possible architecture for their model, but WIT MLPs were felt to be simpler and sufficient for the task. The implication was that the choice was not of great theoretical importance. Nevertheless, if Elman-style recurrence were shown to be clearly inappropriate and neurally implausible (independent of back-propagation and other aspects of connectionist models), it would be good to know whether or not recurrence was essential to the performance of such models.

For the use of distributed, or feature-based, representations of segments as opposed to symbolic ones, the argument seems at first glance to be obvious, at least from a linguist's point of view: some pairs of phonemes are more similar than others, and a system that represents /b/ as equally different from /p/ as from /k/ or even /a/ seems to be missing something crucial about the nature of the input. Whether the phonological features themselves are psychologically real, or simply a useful shorthand for certain acoustic similarities between phonemes, it seems intuitively reasonable that representing them in the input should help. Furthermore, if people find features helpful, then one would expect a realistic model to benefit from them as well.

Practically speaking, however, whether these two design choices, separately or together, actually make a difference in a model's performance on a particular task is an empirical question. For the former, it stands to reason that any WIT MLP equipped with a sufficiently large time window will have the wherewithal to

capture whatever generalizations the SRN could capture for the same task—at the price of the inelegance of choosing that window size. This in essence becomes a trade-off in plausibility: clearly, the idea of a fixed time window inside the brain is bound to raise some eyebrows, but neither is there any guarantee of finding recurrent connections within the areas of the brain associated with the WST. However, if the use of recurrence does not greatly affect the model's performance, it can be regarded as a non-essential aspect of the model. The next section will examine the utility of recurrence on the original mini-languages used in AC96.

There may be more at stake in the latter design choice: whether to use feature-based or symbolic input representations. Two issues arise from this choice, besides the ability (discussed above) to model the interaction of multiple cues: first, the fitness of a particular phonological feature set simply as an approximation of auditory input; second, the role that features might play in the phonotactic structure of the language being learned. The theoretical implications of using features in the input representations will be discussed further in Section 3.3.2.

3.3.1 Considering recurrence

The first distinctive facet of the Elman, Cairns, and Christiansen models, the use of recurrent Elman nets over feed-forward (non-recurrent) MLPs, may seem like a minor issue. One advantage of the Elman net is that it frees the researcher from worrying about the exact amount of memory needed to represent and solve a particular problem. On the other hand, in cross-linguistic comparisons, there are reasons to want that aspect of the model to be made explicit. For example, the relative performance of various window sizes may still reveal some characteristics of the distributional cues present in the language in question. Some languages may have a quite restricted set of phonemes that are allowed to appear word-finally.

Such languages would be predicted to get considerable mileage out of even a single segment's worth of context. (See Chapter 4 for more discussion of this point.) Other languages (like English) allow many segments to appear word-finally and are likely to need longer segmental contexts (or the help of other cues). This was demonstrated by AWLB96 (as discussed in Section 3.1.3 above), who showed that time-windowed MLPs trained with two or three phones of context (on English CDS) generalized from utterance boundaries to word boundaries much better than those given just one phone of context. Such knowledge is likely to be stored (implicitly) in the SRN's hidden layer as well, but examining it will be less straightforward.

In order to investigate further the practical import of the first of these design choices—recurrence vs. fixed time windows—Rytting (2006b) replicated the crucial portions of Allen and Christiansen (1996), using artificial mini-languages with the same properties as those in the original AC96 study. One of these languages was designed such that each syllable type was equally likely to end a word and transitional probabilities between each syllable were held constant, like the "flat" mini-language in AC96.[10] The other mini-language, which corresponded to AC96's VTP language, had properties more akin to that of natural languages, such as variable transitional probabilities between different syllables and segments (with respect to each other and to word boundaries). As may be expected, the net was not able to learn to distinguish word boundaries from other syllable boundaries when trained on the first language, but it could do so for the second, more naturalistic language.

[10]The transitional probabilities were kept invariant not only by carefully constructing the words themselves, but also by making each word appear with equal frequency. This property of the language, in addition to its lack of syntactic structure, made it very unlike natural human languages, which display great variation in word type frequency—cf. e.g. Zipf 1965; Mandelbrot 1966.

This replication (performed using an SRN) was then compared with three WIT MLPs with time-window sizes of 1, 2, or 3 segments' worth of previous context, corresponding to the three window sizes used in AWLB96. These networks were trained and tested 16 times each, with different initial starting weights for each run.

All of the networks distinguished the true word boundaries from other syllable boundaries above chance when trained on the naturalistic language—unsurprising when one considers that the distribution of the final segment alone provided some clue to word boundaries. The various runs of the SRN ranged in performance greatly; the worst run was no better than the single-segment MLP (equivalent in design to the SRN with the context nodes removed), and the best run was almost as good as the average run for the three-segment MLP. The mean performance for the SRNs was statistically equivalent to that of the two-segment MLP, suggesting that, on average, the SRNs learned about two segments' worth of context (one segment beyond that given in the input), and never benefited from more than three segments' worth of context for that language. In this toy-sized problem, not much additional context was needed for a reasonably well-performing solution; nevertheless, it does not seem likely that this is due to a ceiling effect, as the discrimination performance for the three-segment case still had room for improvement.

One possible explanation for the SRNs' difficulty in consistently making use of the more distant (less recent) information in the input stream is the well-known problem that SRNs trained with gradient descent methods have with "latching on" to long-range information (see e.g. Bengio et al. 1994). As the temporal distance between the cue and the relevant training example increases, SRNs tend to become less and less efficient at learning to associate them.

If the most relevant cues to the WST are indeed far away from the end of the word (say, for example, a language where word boundaries are cued primarily by consistent antepenultimate stress, using the segment as the temporal unit of training), then training SRNs (or WIT MLPs, for that matter) with gradient descent would be a poor choice. However, if we assume that most child-directed speech has ample cues to word segmentation near the ends of words, then these concerns should not be an obstacle to using either SRNs or MLPs.

It must be borne in mind that SRNs have shown more variation between runs than WIT MLPs in learning the same task (Rytting 2006b). Thus, if results are based on a single run of the SRN, it may not be the learning task that determines how much previous context is kept in memory, but rather the randomized initial weights. Hence, it is important to examine multiple runs over the same data, so as not to generalize from the results of a single and possibly unreliable simulation.

Nevertheless, this minor issue does not clearly outweigh Elman's (1990) original arguments for the SRN. For these reasons and for ease in comparison with previous connectionist approaches, the simulations reported in the rest of this chapter use simple recurrent networks, with the note that recurrence is not an essential part of the model.

3.3.2 Considering input representations

A second noteworthy feature of the connectionist strand of models is the use of (sequences of) feature vectors rather than symbols in their representations of the input. It is worth noting that this style of representation is not inherently and exclusively connectionist: other machine learning paradigms (e.g. nearest-neighbor

approaches) also are able to use featural representations. However, the present author is not aware of any non-connectionist approaches to the WST specifically that have used distributed, featural representations rather than symbols.

Conversely, it is possible to use symbolic input representations within a connectionist framework, by using a "one-hot" (or localist) representation: each symbol in the relevant phonemic set is represented by a vector consisting of an activation of 1 for one feature uniquely associated with that symbol, and 0 for every other feature. It is worth noting that while the other connectionist approaches herein reviewed use distributed featural representations everywhere in the model, the Christiansen model differs from Elman (1990) and CSCL97 in using localist representations also. The preliminary model (AC96, discussed above) used localist representations for both input and output; CAS98 and CCC05 use both: they return to using phonological feature vectors for input, but continue to use localist representations for the segmental portion of their models' output and target units. (The features corresponding to lexical stress are handled differently.) CAS98 does not claim any theoretical motivation for this design choice: these localist representations are used "to facilitate performance assessments and analyses of segmentation errors" (p. 236).

Two questions arise in considering the merits of feature-based representations relative to localist ones. One is the relative appropriateness of distributed versus localist representations generally, divorced from details of the feature set. This question hinges on properties of ANNs and the mathematics of the training methods used. For example, it may well be that localist representations are simpler to learn in particular situations for mathematical reasons removed from the fitness of any particular feature set. This concern is a real one and cannot be excluded

from consideration when evaluating and comparing the relative performance of different models. However, for now it will be left as an open question.

The other question deals with the appropriateness of a distributed feature presentation, including the specific feature set used, to representing linguistic input. As mentioned above, there are two dimensions on which a representation may be judged. The first regards features as a substitute for acoustic similarity. Phonological features seem to be a better approximation of the actual auditory input than a mere sequence of symbols, insofar as certain pairs of phonemes, e.g. /n/ and /m/, are more similar to each other (and more easily confusable) than other pairs. Of course, a direct acoustic or auditory representation, if it were made practical, would also preserve this similarity. However, some feature systems may encode auditory similarities as perceived by native or soon-to-be-native speakers better than others. The second is similar to the mathematical issue above, but focuses the properties of the system being learned (i.e. a specific language's phonotactic rules) rather than on the learner (e.g. ANNs). An assumption widely (though not universally) held in theoretical phonology is that phonological features play a substantial role in the structuring of human languages. For example, if a certain language allows only a certain set of phonemes to appear word-finally, one would expect the members of this set to have some feature or features in common, e.g. [−*consonantal*] or [+*consonantal*][−*voice*].

Some languages, however, do not follow such clear patterns. For example, the allowable word-final consonants in Modern Greek (discussed at length in Chapter 4) are /n/ and /s/, not counting loan words and other peripheral phenomena. This pair, a coronal fricative and a coronal sonorant, do not form a natural class, at least according to most phonological feature sets. Moreover, Modern Greek is not alone here: Mielke (2004:172) cites fourteen other languages where

fricatives and sonorants pattern together in some type of morphological or phonological alternation, such as Bukusu, where nasals delete before nasals and fricatives. In Ancient Greek the class of possible word-final consonants is /n/, /s/, and /r/, an equally disjoint class (Thrax 1883). In Kolami (Mielke, 2004:164) every sound in the language's inventory may be word final, except /b, tʃ, dʒ, u, uː, o, oː/—another unnatural class. Although these examples may seem marginal, Mielke extimates that 24% of the classes of phonemes that participate in phonological or morphophonological alternations are not predicted to be a natural class by *any* standard phonological feature set. While Mielke did not explicitly examine word boundary phonotactics, it seems likely that similar degrees of exceptionality would be found there as well.

If one happens to be learning a language where phonological features are relevant (either by hard phonotactic rule or statistical tendency) to learning word boundaries, then it stands to reason that including some information about the relevant features in the input should help the network to learn these generalizations, resulting in better predictions of word boundaries. AWLB96 makes such a claim, demonstrating that a phonologically relevant feature set facilitates performance compared to an arbitrarily assigned feature set of equal length. This is not quite the same as using "one-hot" representation, where each phoneme has its own unique input feature, and so these findings do not directly compare to the symbolic input representations used in non-connectionist models. On the one hand, using an arbitrary feature set as a control avoids the above-mentioned issue of potential differences in how ANNs handle distributed versus localist representations; on the other, this choice has left it unclear whether AWLB96's main feature set was particularly well-suited to the problem, or that study's arbitrary (control) feature set particularly ill-suited to it.

Similar ambiguities arise when trying to understand the implications of empirical psycholinguistic findings. It seems that (at least some) phonological features—or the acoustic similarity that they entail—can play a role in making certain phonotactic patterns easier (or harder) to learn than others. Saffran and Thiessen (2003) find that children are able to learn a phonotactic rule involving a contrast in behavior between the sets of phonemes {/p/, /t/, /k/} and {/b/, /d/, /g/}—generalizable as $[-voice]$ versus $[+voice]$—but not when {/p/, /d/, /k/} and {/b/, /t/, /g/} (sets that cut across the $[voice]$ feature) are given contrastive behavior. However, they note that this does not prove that $[voice]$ as a phonological feature is psychologically real. An alternate explanation is that phonetically similar sounds are more easily grouped together, and that $[voice]$ is simply a convenient "shorthand" for referring to the phonetic similarity shared by sets of sounds.[11]

The point of this discussion is not to make strong claims about the the connection between a particular feature set's performance on a particular task and the relevance of phonological representations as a whole to language acquisition. Quite the opposite: in situating the choice of an input representation in context of the many factors that might effect a particular input representation's performance, one should gain a healthy skepticism for categorical claims extrapolating from this choice.

In this same vein, one should not read too much into the performance of the Christiansen input representation schemes vis-à-vis other representations, nor

[11]In order to distinguish between these two explanations, one may have to replicate this experiment on groups of sounds where distinctive feature theories and segment confusability make different predictions. For example, Mielke (2004) cites Graham and House (1971) as reporting that the pairs /f/-/θ/ and /r/-/w/ were particularly confusable to children, even though by many traditional phonological feature geometries, they differ by more than one feature. If children learn a phonological rule that combines /f,θ/ or /r,w/ into a single (morphophonemic) natural class more readily than a group defined in standard theories as a natural class by virtue of differing on only a single phonological feature, then acoustic similarity would be preferred as an explanation.

view them as representative of phonology as a whole. On the one hand, the feature vector representations used for the phonemes in CCC05 better represent typical usage of features in theoretical phonology than the one used in CAS98. On the other hand, even in the CCC05 representation there are a number of idiosyncratic or questionable divergences from the current norm.[12]

Given these uncertainties, it seems wise to control for input representation where possible, but not to read too much theoretical significance into its details. From the plausibility standpoint, the best input representation would logically be that which most closely represents the auditory nature of the input. However, as making that determination is non-trivial, one may turn to performance as a guide for feature-set selection—keeping in mind that issues of the machine learner may also play a role. On the other hand, it stands to reason that if changing feature sets does not significantly change the model's performance, then whichever aspects of the input representation were changed are not crucial aspects of the model. That is, if distributed features perform no better than a localist representation, it does not make sense to argue that that set of phonological features is necessary to the model. However, neither does this show that phonological feature systems are

[12]One such divergence is the failure to distinguish between negative and unspecified values for features (e.g. $[-laminal]$ coronal phonemes versus non-coronal phonemes for which $[laminal]$ is not relevant). While such a feature set may be consistent with Trubetzkoy's and Jakobson's original conceptions of distinctive features havig 'marked' and 'unmarked' values (cf. e.g. Trubetzkoy 1939; Jakobson and Halle 1956), it risks grouping certain sounds into classes distinguished only by their lack of a particular feature, rather than by their true similarity. This risk becomes particularly high when applying a feature system to a phoneset (or a language) for which they were not originally designed.

Alternatives to this would involve either postulating additional input features to distinguish negative from unspecified features (e.g. using one binary input feature for $[+laminal]$ and another for "true" $[-laminal]$, with both of these features set to zero for phonemes unspecified for [laminal]) or using intermediate input values (between 0 and 1) for the unspecified features. An elaboration of this latter scheme is suggested in Li and MacWhinney (2002). While a comparison of this variant input representation would perhaps be instructive, it would be more interesting yet to use a feature set extracted from the feature system itself, e.g. from the iterative clustering methods described by Lin (2005), or one tied to the articulatory-phonetic connections learned in babbling, as in Plaut and Kello (1999).

irrelevant generally—just that the particular feature set chosen is not responsible for the model's success. These issues will be explored further in Chapter 6.

3.4 Conclusion

This chapter has reviewed an influential strand of connectionist models, from Elman (1990) to Christiansen et al. (2005). It has noted some of the chief characteristics of these models, and commented on two of them, namely recurrence and the use of inputs based on phonological features. More importantly, it has discussed the inadequacies of transcription-based symbolic input (even when converted into features), chief among these being the loss of sub-word and subsegmental variation. Several attempts at recovering, or rather approximating, this variation (e.g. CA97) were reviewed, and found to be less plausible and generally inferior to a model that bases its input more directly on acoustic data, using the transcription only as an evaluation help rather than as the source of input data itself.

The next chapter more closely examines the two principle cues discussed in this chapter (Harris' cue, as first implemented by Elman (1990) and CSCL97, and Aslin's utterance-boundary cue) and combined in CA96. Specifically, it addresses how these cues may be applied directly as operationalized using simple statistics measured over pairs of segments, rather than implemented in an ANN framework as described in this chapter. This direct application allows aspects of learning to be considered separately from the cues themselves, and allows specific statistical variants to be compared and contrasted. Chapter 4 explicitly tests the cross-linguistic validity of these cues, by comparing previous work in English with novel simulations over a corpus of Modern Greek.

CHAPTER 4

STATISTICAL SEGMENTAL CUES FOR MODERN GREEK WORD SEGMENTATION

4.1 Introduction[1]

Chapter 2 provided an overview of possible acoustic cues available to infants engaged in the word segmentation task, including suprasegmental cues such as stress, subsegmental cues such as coarticulation, and cues depending on the distribution of the segments themselves (with relation both to one another and to utterance boundaries). Chapter 3 examined a strand of connectionist models using these latter (segmental) cues, both alone and (in the case of the Christiansen model) combined with stress. Subsegmental (or rather subphonemic) effects were either approximated (as in CSCL97) via phonological rewrite rules or handled via phonetic rather than phonemic transcriptions (as in CA97). Still finer acoustic cues depending on subphonetic detail could not be studied, due to a lack of suitable data.

Hence, within the connectionist strand of computational models, the emphasis has been on segmental, distributional cues—quite possibly because segmental, distributional data is the easiest type of data to obtain and process. This

[1]This chapter is a revised form of the author's (2004, unpublished) second precandidacy paper. Portions of the research discussed in this chapter were presented at HTL-NAACL'04, SIGPHON'04, and MCLC'04.

tendency is even more pronounced in the other types of computational WST models examined—the multimodal models (especially CELL) being a notable exception. But since these do not use standard corpora, and are focused on issues of word-to-meaning mapping (within a limited domain) rather than word segmentation per se, their results are difficult to compare with the other models examined here.

This chapter also restricts itself strictly to segmental, distributional cues in order to gain a fuller appreciation for the power and potential of these cues considered alone and unaided by other (e.g. sub- or suprasegmental) evidence. Upon this foundation rests a proper understanding of segmental cues in combination with other evidence. Opinions differ regarding the full potential of any one type of cue in general, as well as the best way of modeling the use of segmental cues specifically. On the one hand, experimentalists (and implicitly CAS98) favor the integration of multiple cues, even though the study of these interactions is often difficult. Aslin et al. (1998) says it well:

> [D]istributional analysis of language input is unlikely to rely on a single type of information [...] such as frequency of co-occurrence or transitional probability between adjacent syllables. Rather, natural word segmentation relies on a variety of sources of information [...] even though none of these sources of information alone is reliable enough in natural speech input to solve the word-segmentation problem (pp. 321-2).

Thus these studies do not see themselves as investigating *the* answer to the WST, but only *an* answer—and a partial answer at that.

More recent studies argued for a more prominent role for single cues, combined with some method of incrementally bootstrapping from these cues. Brent (1999) (as mentioned in Section 1.2.2.3) argued for an incremental word-learning

mechanism (modeled as INCDROP) that initially only uses information from one-word utterances to build a vocabulary, then uses this vocabulary (even if wrong), along with a sophisticated use of minimum description length, to break apart other utterances into smaller-size chunks. This approach was shown to be very effective, perhaps the most effective to date in pure performance, given its input assumptions). Furthermore, Brent and Siskind (2001) argued that the words children actually show evidence of learning correlate strongly with those that appear in isolated one-word utterances. A more general version of this claim was corroborated in another study (Seidl and Johnson 2006), which found that words on either the start or end of an utterance are easier to segment out and learn.

Hockema (2006) argues for a greater role for segmental cues (with or without stress) for a completely different reason: simply that they are more powerful than anyone previously supposed. He claims that, in American English at least, if infants only knew the "probabilistic phonotactics" of their language as they relate to word boundaries, they could, with that information alone, segment out the majority of words. However, since Hockema does not provide a solid account of a mechanism by which these cues could be learned (without first solving the WST itself), his claims must be taken as a proposal for an abstract "upper bound" rather than a full-fledged model of word segmentation.

Unfortunately, much of this debate regarding the sufficiency of segmental cues alone has been carried out (as Hockema acknowledges) with reference to English data only. Very few computational models have been tested in non-English languages.[2] This chapter, as an excursus from the examination of connectionist

[2]Exceptions include Batchelder 2002 on Spanish and Japanese, Swingley (2005) on Dutch. Both of these use a "clustering" paradigm pioneered by Olivier (1968) rather than a "parsing" or word-boundary detection paradigm such as those employed by the connectionist strand of models and INCDROP.

models' input assumptions, looks in more theoretical terms at the potential power of distributional cues (or "probabilistic phonotactics") at the segmental level as they relate to Modern Greek, a language with somewhat different segmental properties than English. By so doing, one may hope to gain a deeper appreciation for what is specific to English and what is (even potentially) applicable to a wider range of languages (to say universal would be too great a leap).

Section 4.2 puts forth in more detail four variant proposals (extrapolation for utterance ends, transitional probability, mutual information, and word boundary probabilities) for modeling the use of distributional segmental cues in terms of the statistics and heuristics involved. Section 4.3 examines some of the questions left unanswered by each of these variant proposals. In Simulations 1-3 (Sections 4.5, 4.6, and 4.7), these models are applied to Modern Greek—a language previously untested computationally—by using a corpus of CDS described in Section 4.4.2. The results and implications of these simulations are discussed in Section 4.8.

4.2 Variant heuristics for statistical segmentation

4.2.1 Extrapolation from utterance boundaries

Unlike the other variants described here, Aslin et al. (1996) (AWLB96) focuses on the distribution of various phonemes with regards to the end of utterances and extrapolates from the likelihood of certain phonemes to be utterance final to their propensity to be word final. The crucial statistic here is the conditional probability $\Pr[ub|\mathbf{S}\_\_\_]$ (the context marker ___ borrowed from phonological rule terminology), or more precisely $\Pr[ub_t|\mathbf{S}_{t-n:t-1}]$, where ub represents the presence of an

utterance boundary at time t given some string **S** consisting of n segments, immediately preceding time t. This statistic is calculated as shown in Equation 4.1 for some **S** of length $n = 2$, where ubm symbolizes the utterance boundary marker.

$$\text{(4.1)}\quad \Pr[ub|\mathbf{S}\_\_\_] = \frac{\text{Freq}(s_{t-2} = x, s_{t-1} = y, s_t = \text{ubm})}{\text{Freq}(s_{t-2} = x, s_{t-1} = y)}$$

The statistic, which will be referred to as (conditional) utterance boundary probability (or Utt.BP), is naturally being used to approximate $\Pr[\#|\mathbf{S}\_\_\_]$ (the probability of a word boundary given the preceding string **S**) which is of course unavailable to the infant word learner.

As previously described in Section 3.1.3, AWLB96 model this heuristic with an MLP rather than simply collecting and using the above statistic directly. This enables them to represent each segment using phonological features, which should in principle allow for the network to make generalizations over different classes of sounds. AWLB96 claim that this feature is essential to the model and suggest that without it (or, at least, with an arbitrary featural representation), accurate learning of word boundaries will not take place. They also found, after testing context lengths of $n = 1$, 2 and 3, that a single segment of context is not sufficient for their corpus of English CDS but that 2 or 3 phones was sufficient.

4.2.2 Transitional probability

As mentioned in Section 2.3.3.2, Saffran et al. (1996b) proposed that infants are able to use transitional probabilities to find word boundaries in the absence of other cues. While Aslin et al. (1998) note that a variety of statistics (including conditional entropy, mutual information, backward transitional probability, and correlation) fulfill the same role of representing normalized frequency of co-occurrence, transitional probability has become the standard way of referring to this type of

predictability. This was originally tested at the level of the syllable so that the (forward) transitional probability of /si/ given /ke/ (measuring time in terms of syllables, not segments) would be calculated as in Equation 4.2.[3]

$$\Pr[/\mathrm{si}/|/\mathrm{ke}/] = \frac{\mathrm{Freq}(/\mathrm{kesi}/)}{\mathrm{Freq}(/\mathrm{ke}/)} \tag{4.2}$$

A more recent study (Newport et al. 2004) provides some evidence that transitional probability between segments may also be important—perhaps even more so than syllables. While the exact unit (or units) of granularity most relevant for word segmentation is unknown (and may differ from language to language), the study here will restrict itself to examining segments.

4.2.3 Mutual information

Although for the purposes of experimental design the various statistics listed in the previous section may be regarded as equivalent (as they are all correlated and may be difficult to tease apart in stimuli already controlled for frequency), computational models need to be more precise. Hence it is worth considering the potential differences between transitional probability and mutual information, if only to know which one better represents the true potential of distributional cues to solve (or help solve) the WST. Fano's (1961) mutual information (now often called *pointwise mutual information*—see Manning and Schutze 1999:178,182) is very similar to transitional probability, but it is normalized for both phones rather than just one. It is calculated as shown in Equation 4.3.

$$I(s_1, s_2) = \log_2 \frac{\Pr[s_1 s_2]}{\Pr[s_1]\Pr[s_2]} \tag{4.3}$$

[3]Backward TP would reverse the places of /ke/ and /si/ in the equation, yielding $\Pr[/\mathrm{ke}/|/\mathrm{si}/]$.

While the main point of Brent 1999 is to explore the role of minimum description length in a bootstrapping mechanism, Brent provides in passing two simple but useful baseline models: one based on transitional probability (TP), the other on mutual information (MI). Since these models (particularly the TP model) are very closely related to, and obviously inspired by, Saffran and her colleagues' work (albeit applied on segments rather than syllables), the TP version may usefully be regarded as an implementation of the same. Brent's algorithms incrementally keep track of unigram (1-segment) and bigram (2-segment) frequencies by updating unigram and bigram counts at every new segment encountered. Then they apply the appropriate statistic (MI or TP) and insert a boundary if this statistic dips below the corresponding statistic of both neighboring pairs. For an example that applies MI to a string $s_0 s_1 s_2 s_3 \ldots$, a boundary would be inserted between segments s_1 and s_2 if $I(s_1, s_2) < I(s_0, s_1)$ and $I(s_1, s_2) < I(s_2, s_3)$. Word boundaries are also inserted automatically at all utterance boundaries.

Brent (1999) compares the MI and TP heuristics so defined (evaluated in terms of word tokens correctly segmented–see Section 4.6 for exact criteria), reporting approximately 40% precision and 45% recall for transitional probability (TP) and 50% precision and 53% recall for mutual information (MI) on the first 1000 utterances of his corpus (with improvements given larger corpora). Indeed, the performance on word tokens of these two simple heuristics is surpassed only by Brent's main model (MBDP-1, his implementation of INCDROP), which seems to have about 73% precision and 67% recall for the same range.[4] From these figures, it is clear that both are surprisingly effective, but that MI is superior to TP.

[4]The specific percentages are not reported in the text but have been read off Figures 3 and 4. Brent does not report precision or recall for word boundaries; those percentages would undoubtedly be higher.

4.2.4 Word boundaries and segment bigrams

Finally, Hockema (2006) performs a direct examination of conditional word boundary probability (CWBP) $P_{wb}(x, y) = \Pr[\#|s_t = x, s_{t+1} = y]$, the probability of a word boundary being found between two adjacent segments x and y. This statistic is of course highly unlikely to correspond directly to an actual cue used by infants, as it is not directly observable from the input infants are assumed to receive. Since the goal of the WST is to find word boundaries given some set of observable phenomena (e.g. a sequence of observed segments), it is circular to use statistics involving word boundaries directly. In machine learning terms, to assume even partial knowledge of the word boundaries themselves is to situate the task into a class called *supervised* learning problems, in which some training set of *labeled* data is given to the learner, whose task it is to generalize this to a distinct (test) set of unlabeled data. However, the WST is not assumed to be a case of supervised learning, since parents do not explicitly—but only indirectly—signal most word boundaries, as discussed in Chapter 2. All the other heuristics proposed in the literature assume only data with no labels for word boundaries (except what the heuristic itself can bootstrap), and hence are unsupervised learning mechanisms.[5]

Nevertheless, since it may be argued that MI, TP, and all of the cues mentioned by Aslin et al. (1998) are in a sense (observable) approximations of this ideal statistic $P_{wb}(x, y)$, it may serve as an "upper bound" for the information about word boundaries that is extractable *in principle* simply from the identities of the preceding and following segments. And, for American English at least, this upper bound is astonishingly high: when the threshold for P_{wb} is set (sensibly, and in fact

[5]Although the connectionist models discussed in Chapter 3 are in fact supervised for their "immediate task(s)"—predicting the next phone's identity, stress, and/or an upcoming utterance boundary—this supervision only uses cues that may reasonably be assumed to be observable (with some degree of accuracy. See Sections 5.2.2 and 6.2 for further discussion).

optimally) at 0.5, it finds boundaries with a precision and recall of 86.5% and 76%, respectively and word tokens with 66.6% precision and 59.6% recall.[6] Naturally, adding stress as a cue improves these numbers still further.

4.3 Open questions

4.3.1 Extrapolation from utterance boundaries

In reexamining all these variants for Modern Greek, it is useful to consider what the open questions behind these cues are and what types of cross-linguistic variation to look for in comparing these results with English. Additionally, there are some questions left unanswered in the variants themselves. For example, in their previous examinations of the first variant cue—extrapolation from utterance boundaries—AWLB96 claim that a featural representation is necessary but never test the heuristic directly on phones. Working outside of a connectionist framework, such a test becomes simple—simpler, in fact, than using features. Of lesser interest is the exact "window size" of context needed (which could in principle differ from language to language); however, since heuristics like this may be expected to improve in performance with increased amounts of context, this is hard to measure without some concrete threshold of performance in mind.

4.3.2 Transitional probability and mutual information

As for the two next variants, transitional probability and mutual information, several questions naturally arise. One—the optimal level of granularity—is touched on only partially. The simulations here test whether TP is a viable statistic at the

[6]This is using Hockema's "worst case" statistics in Footnote 4 of his 2006 paper. The main text cites even higher numbers for word tokens (70% precision, 61.9% recall) based on the assumption that all words not found in his dictionary be treated as utterance boundaries.

segmental level. Since syllabification in Modern Greek, though simpler than that for English, cannot be assumed to be completely unambiguous for children (or even adults: see e.g. Rytting 2005 for some difficulties in syllabification), the syllable level will not be tested here.

Another question—the relative performance of TP and MI—is examined in greater detail. Brent (1999) shows that for English corpora, MI surpasses TP in performance. Hockema (2006) suggests this is due to the higher correlation between MI and the conditional word boundary probability (CWBP) upper bound than between TP and CWBP. What is not known is whether this is a fact about English or is more generally applicable across languages.

Finally, while Harris, Saffran, and others speak of a "dip" in segmental predictability (however measured), and Brent defines this explicitly in terms of comparisons with neighboring phone pairs, it has not been demonstrated that such a reference to neighbors is strictly necessary. For example, it could be that the dip is merely in reference to some global threshold.

The global comparison, taken on its own, seems a rather simplistic and inflexible heuristic: for any pair of phonemes xy, either a word boundary is always hypothesized between x and y, or it never is. Clearly, there are many cases where x and y sometimes straddle a word boundary and sometimes do not. The heuristic also takes no account of lengths of possible words. However, the local comparison may take length into account too much, which resultantly disallows words of certain lengths. In order to see that, we must examine Brent's 1999 suggested implementation of Saffran et al. (1996b) more closely.

In the local comparison, given some string $\ldots s_0 s_1 s_2 s_3 \ldots$, in order for a word boundary to be inserted between s_1 and s_2, the predictability measure for $s_1 s_2$ must be lower than both that of $s_0 s_1$ and of $s_2 s_3$. It follows that neither $s_0 s_1$

nor s_2s_3 can have word boundaries between them, since they cannot simultaneously have a lower predictability measure than s_1s_2. This means that, within an utterance, word boundaries must have at least two segments between them, so this heuristic will not correctly segment utterance-internal one-phoneme words.[7] This difficulty may have only a small effect the the model's overall results, as only a few one-phoneme word types exist in either English or Greek (or other languages). Moreover, these words are often function words and so are less likely to learned early on by children. However, were the granularity of this model to be extended to apply to syllables rather than to segments, this limitation would be serious indeed for languages such as English, in whic a majority of its commonly used words are monosyllables.

Brent (1999) points out another length-related limitation—namely, the relative difficulty that the "local comparison" heuristic has in segmenting and learning longer words. The bigram MI scores may be most strongly influenced by—and thus as an aggregate largely encode—the most frequent, shorter words. Longer words cannot be memorized in this representation (although common ends of words such as prefixes and suffixes might be).

In order to test for this, Brent proposes that precision for word types (or "lexicon precision") be measured as well as precision and recall for word tokens. While the word-token metric emphasizes the correct segmentation of frequent words, the word-type metric does not share this bias. Brent defines this metric as follows: "After each block [of 500 utterances], each word type that the algorithm produced was labeled a true positive if that word type had occurred anywhere in the portion of the corpus processed so far; otherwise it is labeled a false positive."

[7] At the edges of utterances, this restriction will not apply, since word boundaries are automatically inserted at utterance boundaries, while still allowing the possibility of a boundary insertion at the next position.

Measured this way, MI yields a word type (or lexicon) precision of about 27%; transitional probability yields a precision of approximately 24% for the first 1000 utterances.[8] He does not measure word type (lexicon) recall.

This same limitation in finding longer, less frequent types may apply to comparisons against a global threshold as well. This is also in need of testing. It seems that both global and local comparisons, used on their own as sole or decisive heuristics, may have serious limitations. It is not clear *a priori* which limitation is most serious; hence both comparisons are tested here.

4.3.3 Word boundaries and segment bigrams

As for the last variant, it is beyond the scope of this thesis to speculate on how this statistic might be applied in practice, except through approximations such as those proposed above and elsewhere. For example, Brent and Cartwright (1996) proposed the logical extrapolation of this heuristic from utterance boundaries via a discretized, non-probabilistic version of AWLB96 itself. Another possibility (discussed by Hockema but explicitly excluded from the Brent and Cartwright 1996 model) would be for this cue to be updated as words are identified and segmented off by means of some other heuristic. In any case, the performance of such a bootstrap would need to be explicitly tested by simulation to be properly evaluated. Suffice it to say that it is unlikely to be useful at the earliest stages of acquisition due to the need for a prior bootstrap. Conversely, it is not likely to be the sole cue in later stages of acquisition, or to consistently overshadow other cues, such as suprasegmental cues (e.g. the trochaic bias—cf. Cutler 1994), for these latter cues (once acquired) continue to show strong effects even into adulthood (Cutler and Butterfield 1992; Mattys and Samuel 2000).

[8] Percentages are estimated from Brent (1999), Figure 5.

However, it is of some interest to consider CWBP as an upper bound for the potential contribution of distributional cues found in the bigrams surrounding word edges. While Hockema has shown this to be quite high for American English, the value of this upper bound is not known for other languages. A language with a smaller CWBP upper bound would then, in principle, have to rely either on a wider span of context (e.g. the syllable) or on other, nonsegmental cues (e.g. a greater relative role for stress, if present in the language, or other prosodic features)—unless such approximations as TP, MI, or extrapolation from boundaries come closer to that upper bound than for American English.

4.4 Testing the four variants

4.4.1 Differences between English and Modern Greek

Modern Greek differs from English in having only five vowels, generally simpler syllable structures, and a substantial amount of inflectional morphology—particularly at the ends of words. It also contains not only preposed function words (e.g. determiners) but postposed ones as well, such as the possessive pronoun, which cannot appear utterance-initially. While it is not anticipated that Modern Greek will be substantially more challenging to segment than English, it is sufficiently different in its properties to serve as a useful test of current assumptions about the nature of word segmentation.

Phonemic inventory Greek has five basic vowels, /a, e, i, o, u/, each of which may be stressed or unstressed. Unlike English, there is no centralized or 'reduced' vowel, although unstressed vowels may undergo slight shrinkage of the vowel

space (see e.g. Fourakis 1986; Fourakis et al. 1999). However, in certain contexts many speakers devoice unstressed high vowels /i/ and /u/ (and occasionally even mid vowels). While these devoiced vowels still may leave a few acoustic traces of their presence, these traces may not always be enough to recover the intended vowel, leading to the perception that these vowels have been deleted altogether. The implications of this are discussed further below.

The set of Greek consonants include (at least) the voiceless stops /p, t, k/, the fricatives /f, x, s, θ, v, ɣ, z, ð/ and the sonorants /m, n, l, ɾ/. Some hold that the voiced stops [b, d, g], and the affricates [ts, dz] are separate phonemes; others that they are realizations of the sequences /Np, Nt, Nk, ts, Nts/ respectively, where /N/ is an abstract nasal neutral for place. Similarly, the series of palatal consonants [c, ɟ, ç, j/j, ɲ, ʎ] are usually analyzed as palatalized versions of /k, g (or Nk), x, ɣ, n, l/ before certain contexts involving front vowels. Although these issues of phonemicization may not seem relevant at first, they do play a role in determining how to transcribe a corpus. For example, the transcription system used in the Stephany (1995) corpus (see Section 4.4.2 below) transcribes [c], [ʎ], and [ɲ] as two-segment strings: /kj/, /lj/, and /nj/. One consequence of this is noted in the evaluation section (4.8.1).

Syllable structure The maximum allowable syllable structure in Greek is technically the same, but generally simpler than in English. Whereas English allows syllables up to *CCCVCCC* (e.g. 'strengths'), Greek typically allows no more than four consonants in a row, which correspond to a maximum syllable type of *CCCVC*; furthermore, these instances are rare and are generally from borrowed or Ancient Greek sources, e.g. /af.stri.a/ 'Austria' and /ef.spla.xnos/ 'merciful' (See Kappa

2002). Codas of up to three consonants are possible in recently borrowed, non-nativized words (e.g. *CVCCC* /taŋks/ 'tanks'), as are *CC* codas in "learnèd" or "puristic" versions of Ancient Greek words (e.g. /zefs/ 'Zeus'), but the nativized core vocabulary allows only a single coda consonant. At the end of words, only /n/ or /s/ or a vowel are allowed in the native core vocabulary.

CDS sometimes is looser on that restraint than might be supposed, allowing onomatopoeic words with a variety of consonants, such as /m/, and /ts/. However, at the risk of some loss of generality, the work here will disregard the onomatopoeic words (following Brent and Cartwright 1996; Christiansen et al. 1998) and focus on the core native vocabulary, for which the words that do not end in either /n/, /s/, or a vowel are rare indeed. This does not completely eliminate word-final consonants other than /n/ and /s/ from the corpus used here. Some prepositions with devoiced or elided vowels were transcribed without those vowels present, such as [ap] from /ap(o)/ 'from' and [kj] ([c] in standard IPA) from /ke/ 'and'. Although written in standard orthography (and transcribed in the corpus) as separate words, these function words could be analyzed as *proclitics* or prosodically "defective" words, realized prosodically as part of the following word.[9] That analysis will not be adopted here, as convenient as it might be; rather, the spacing used in the original corpus transcription will be followed.

4.4.2 The Stephany corpus

The Stepany (Stephany 1995) corpus is a database of conversations between Greek-learning children and their caretakers included as part of the CHILDES database

[9]This is not to say that vowels cannot be devoiced or elided from content words as well—just that these were not found in the Stephany corpus, and are predicted to be rather rare for infant-directed Greek speech, at least in Standard Athenian Greek. Northern dialects of Greek show much greater rates of vowel devoicing, and insofar as the final vowels are devoiced and perceived as deleted, the segmental cues will be different than the Standard Athenian dialect presented here.

(MacWhinney 2000). The corpus contains transcripts from recordings of four different children with their caretakers; the age ranges of these children (given as *years;months*) were 1;9, 1;9-2;9, 1;11-2;9, and 2;3-2;9 at the time of the recordings. Naturally, conversations with younger children would have been preferred, but no other Greek transcriptions are available on CHILDES. However, the length of utterances in the Stephany corpus is still roughly comparable with those in the Korman corpus used by CAS98. The average number of words per utterance (2.9) is almost identical to that in the Korman corpus, although the utterances and words are slightly longer in terms of phonemes (12.8 and 4.4 phonemes respectively, compared to 9.0 and 3.0 in Korman). The increased word length is hypothesized to be a function of the language as opposed to the age of the children. The caretakers' utterances are broadly (or phonemically) transcribed, with no significant deviations from standard pronunciation, other than the elided vowels mentioned above. The children's utterances are not used in this study.

This corpus was utilized in stages. Rytting (2004a,b) reports preliminary tests of two single-cue models (corresponding to Simulations 1 and 2) performed on a smaller subcorpus involving only the youngest child. This preliminary subcorpus was split into a training corpus containing 367 utterance tokens with a total of 1066 word tokens (319 types) and a test corpus consisting of 373 utterance tokens with a total of 980 words (306 types). These simulations were subsequently replicated as Simulations 1 and 2, and Simulation 3 performed, using data from speech directed to the other three children in the Stephany corpus. The results from the three simulations over the larger corpus are reported here. This larger corpus was split $80\% - 10\% - 10\%$ into training, development, and testing subcorpora, as shown in Table 4.1.

Corpus	Utterances	Word Tokens	Word Types
Training	8,825	26,159	2,162
Development	1,102	3,266	735
Test	1,102	3,342	750
Total	11,029	32,767	2,368

Table 4.1: Size of the training, development, and test corpora for the Stephany corpus in terms of utterances, word tokens, and word types

In all three experiments, as in other studies, only adult input was used for training and testing. In addition, nonsegmental information such as punctuation, disfluencies, and parenthetical references to real-world objects, was removed. Spaces were taken to represent word boundaries without comment or correction; however, it is worth noting that the transcribers sometimes departed from standard orthographic practice when transcribing certain types of word-clitic combinations. For example, with imperative verbs with an enclitic object marker, the object marker is often included as part of the verb, particularly if the verb's final /e/ was dropped, or if a final /s/ was voiced. For example, /vale+to/ and /pes+mu/ are transcribed in the corpus [valto] and [pezmu], respectively. In addition, some (apparently) unusually long final vowels were transcribed with a colon, but no space, as in [ayapi:mu] 'my love'. It is unclear whether the transcribers intended these colons to be interpreted as marking word-boundaries or not. Finally, certain very common phrases where canonical stresses are elided are often marked as one word. For example /'ti#'ine#af'to/ 'what is this?' is often transcribed as [tinafto], or [ti:nafto] if the double /i/ was kept long. Since the second /'i/ is elided, quite likely the stress was lost as well. It may be that the last stress on /af'to/ was also

left unrealized by the speaker, but since the Stephany corpus does not mark stress, it is hard to know for certain.

The text also contains a significant number of unrealized vowels resulting from apocope rather than degemination, such as [ap] for /apo/ 'from', or [in] or even [n] for /ine/ 'is'. Such variation was not regularized but rather treated as part of the learning task. In both the training and the testing corpora, disfluencies, missing words, or other irregularities were removed; the word boundaries were kept as given by the annotators—even when this disagreed with standard orthographic word breaks.

4.4.3 Evaluation

The model variants were each evaluated on three metrics: word boundaries, word tokens, and word types. Note that the first metric, simple boundary placement, considers both utterance-internal word boundaries and the final boundary at the end of the utterance.[10]

Rather than using the "hits to false positives" ratio found in AWLB96, results are reported in terms of *precision* and *recall* (referred to as *accuracy* and *completeness* in CAS98) for boundaries, word tokens, and word types. *Precision* is defined as the number of true positives over the sum of true and false positives. *Recall* is the number of true positives over the sum of true positives and false negatives (or missed items). These are shown in Equations 4.4 and 4.5.

$$\text{(4.4)}\quad \text{Precision} = \frac{N_{tp}}{N_{tp} + N_{fp}}$$

[10] A more conservative boundary measure, using only utterance-internal word boundaries, was used in the previous version of this chapter presented as the second precandidacy paper. The conversion between these two is trivial: subtract the number of utterances in the test corpus from N_{tp}.

(4.5) $$\text{Recall} = \frac{N_{tp}}{N_{tp} + N_{fn}}$$

The following example illustrates the use of these terms. Suppose that the model, upon receiving as input the following test sentence (given here in orthography rather than the phonological representation for the sake of convenience and clarity), posits word boundaries at the positions marked by hashes.

(1) (#) the # f # at # cat # s # aton # the # cat # 's mat #

In this example, the finite-state transducer (FST) posits nine boundaries (including the boundary at the end of the utterance, but not the one at the beginning, which has already been counted for the previous utterance), six of which are correct. Therefore, its boundary precision is $6/9 = 0.67$. Since there are eight correct word boundaries to be found, its boundary recall is $6/8 = 0.75$.

However, the number of boundaries correctly found is of less interest than the number of words (tokens and types) correctly segmented. In order for a word token to count as correctly segmented, three conditions must apply:

1. The word's beginning must be correctly identified.
2. The word's end must be correctly identified.
3. There must be no false-positive boundaries posited in between these two.

In the example sentence, the network has found nine hypothesized words, of which three are correct (both instances of 'the' nd one instance of 'cat'). Therefore, its word token precision is $3/9 = 0.33$. Since there are eight words, the word token recall is $3/8 = 0.375$.

It is also of interest what types of words are found by the model. In this case, the model has correctly identified two word types: *the* and *cat*. It has falsely

learned *f*, *s*, *aton*, and *smat* as possible English words. It has also learned "spurious" tokens of *at* and *cat*, which are English words, but are not the ones intended in this context. Since true version of 'cat' was already learned in this example sentence, the spurious token of 'cat' is discarded for the purpose of calculating type precision. However (assuming a correct instance of 'at' was not learned in a previous utterance), 'at' is counted against the model, since it cannot be paired with the right context here. The type precision is $2/(2+5) = 0.29$, and the type recall is also $2/(2+5) = 0.29$, since five distinct word-form types (*fat*, *sat*, *on*, *cat's* and *mat*) were not correctly segmented and learned.

4.5 Simulation 1: Utterance boundary cues in Modern Greek

4.5.1 Purpose and Methodology

Simulation 1, while not a strict replication of AWLB96, is also designed to test the effectiveness of segmental information before utterance boundaries as a cue to plausible endings of words. Naturally, it resembles their experiment in many respects. However, Simulation 1 diverges from AWML96 at several points—not with an eye toward improving upon their results, but rather on keeping the implementation as simple as possible.

First, instead of learning the transitional probabilities indirectly with connectionist networks, the algorithm employed here was calculated directly from the training corpus. This allows one to abstract away from issues of optimal training length and so forth, and to see the information encoded in the cue itself. Secondly, these transitional probabilities are calculated over the actual phone identities (rather than feature bundles), which not only simplifies the calculations, but also tests whether features are a necessary component. Finally, context lengths of

$n = 1$ and 2 are tested, but contexts of three segments are not. This is done first of all to avoid data sparseness: it is quite likely that the statistics for three phones would be less than reliable given the size of the corpus. Since AWLB96 found two phones to be sufficient for English, it follows that a context of two segments should serve as an adequate comparison to the English case. Moreover, better-than-chance results on the one-phone level could suggest that, for Modern Greek at least, minimum information necessary for traction on the task may be less than what AWLB96 found for English.

The model described here differs from incremental models such as the baselines in Brent (1999) in that it precompiles statistics for the candidate word-final phonemes offline throughout the entire corpus. These transitional probabilities used are thus static and more akin to the training-testing paradigm in e.g. AWLB96 and other connectionist approaches. This difference can be justified as reflecting the fact that infants are exposed to a substantial amount of linguistic material at a very early age—from if not before birth, in fact—and thus may already have a sense of the relevant statistical tendencies before it is applied to finding word segmentations. However, a strong separation between training and test corpora is not anticipated to affect the general pattern of results; it merely simplifies evaluation.

As described in Section 4.2.1, the statistic used in Simulation 1 (Pr[*ub*|**S**___] or Utt.BP), may be seen as an observable approximation for the unobservable statistic Pr[#|**S**___]. This second statistic, while not directly observable by infants, may nevertheless serve (like Hockema's CWBP) as a useful upper bound by which to measure the relative efficacy of extrapolation from utterance boundaries.[11]

[11]When only one segment of context is used, the rankings resulting from using these two statistics may be easily examined and contrasted. The utterance-break approximation given in the second variant yields the ranking $e > s > o > u > i > a > m > n$, with /e/ most likely (and /n/ least likely) to end an utterance.The "true" or upper-bound ranking for Pr[#|**S**___] is $o > i > e > s > a > u > n > j > m > p$. The difference between the two rankings reflects

Since $\Pr[ub|\mathbf{S}\_\_\_]$ is obviously a gross underestimate of $\Pr[\#|\mathbf{S}\_\_\_]$, a threshold for positing word boundaries must be chosen. It is not immediately obvious whether precision or recall is a better measure of performance on the task. Clearly, the correct identification of as many of the words instances as possible (measured by word token recall) is important to learning words and (in time) mapping them accurately to their meanings. However, keeping in memory a large number of incorrectly segmented word-like units (false positives) quite likely increases the cognitive demands of vocabulary learning. As a result, a strategy that generates many wrong guesses for word forms (as measured by low precision) would be counterproductive. A commonly used metric for combining precision and recall measures is known as the F-score. When both are equally important, a balanced F-score (notated F_1 or simply F) is used, which is the harmonic mean of precision and recall, as shown in Equation 4.6.

$$F = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.6}$$

The same question of relative importance also applies to favoring word type (lexicon) or word token (instance) precision and recall. On the one hand, in order for infants to acquire a large receptive vocabulary, they must correctly segment a wide variety of different wordforms. On the other hand, it is also important to segment out enough instances of the more frequent words to learn their meaning (or range of meanings) properly. Hence, both are needed. For all simulations reported in this chapter, thresholds were determined empirically on the development data

the frequency of proclitic masculine and feminine articles /o/ and /i/, which are never utterance-final.

to maximize (with equal weight) word token and word type (lexicon) balanced F-scores, as shown in Equation 4.7.

$$F_{w,l} = \frac{2 * F_{word} * F_{lexicon}}{F_{word} + F_{lexicon}} \tag{4.7}$$

In addition to an upper bound, it is useful to have some baselines to help situate the results of the heuristic, in order to know how much the statistical cues under examination are improving performance over a model without this information. Accordingly, three baselines are provided. Two of these may actually be viewed as special cases (or extreme points along the operating curve) of the heuristic itself. One, labeled *Min[imum]* # or "Utterance as Word", posits *no* word boundaries at all, except those at utterance boundaries. The other, *Max[imum]* #, places boundaries after every instance of the (traditionally) "legal" word-final phonemes (vowels, /n/, and /s/—not counting j, m, and p, which arise in the corpus through apocope or the like). Finally, a "length-based" baseline similar to those used in AWLB96, CAS98, and elsewhere is used, which learns from the training corpus the distribution of word lengths (in segments) but gathers no information relating to the identity of the segments. It then applies these word lengths randomly to each utterance of test corpus, positing boundaries according to the distribution of lengths learned. (It is not claimed that infants themselves tabulate this information in any way; nevertheless, the baseline turns out to be somewhat more challenging than a simple "random walk" as used in AWLB96.)

4.5.2 Results and discussion

4.5.2.1 Results

The results for these four conditions are shown below in Table 4.2. In both cases, the precision scores pattern as expected. The upper-bound statistic (representing a supervised case, where statistics on the word boundaries are available for the training data) proved the most accurate on the test data. However, for the 1-phone context, the differences between the upper-bound statistic and the utterance-based approximation for lexical (word type) precision and recall were not statistically significant. Thus, while the heuristic of extrapolating word-final cues from utterances does not find as many instances of words as the supervised heuristic would, it does find as many word types.

More importantly, the utterance-based probability approximation performs significantly above baselines for all measures except lexical recall ($p < 0.001$ for all comparisons). It also outperforms the Max# baseline on word precision and lexicon recall, and the Min# (or "Utterance-as-Word") baseline in word recall ($p < 0.001$ for all comparisons), although it is somewhat worse at word and lexicon precision ($p < 0.05$, $p < 0.01$ respectively). Of course, the Min# condition is trivially better at boundary recall and worse at boundary precision (the obverse being true for Max#) than any other condition.

Given two phones of context, the utterance-boundary heuristic performs significantly better than all baselines on all measures (save the trivial boundary measures mentioned above). However, it cannot compete with the supervised 2-phone upper bound, which is quite accurate indeed. It may be surmised that, once children begin learning the common word endings in Modern Greek (particularly the more common suffixes), word segmentation becomes much easier.

Cue Type	Window	Boundary		Word		Lexicon	
		Prec.	Recall	Prec.	Recall	Prec.	Recall
Utt.BP (Pr_{ub})	1 ph.	0.562	0.671	0.215	0.257	0.184	0.236
Upp.Bnd. (Pr_{wb})	1 ph.	0.591	0.646	0.258	0.282	0.168	0.229
Utt.BP (Pr_{ub})	2 ph.	0.591	0.857	0.281	0.407	0.284	0.255
Upp.Bnd. (Pr_{wb})	2 ph.	0.737	0.826	0.471	0.528	0.340	0.403
Baselines							
length-based	—	0.437	0.512	0.125	0.146	0.134	0.200
Min.# (Utt→wd)	—	1.000	0.330	0.248	0.082	0.133	0.159
Max.# (*aeiouns*)	1 ph.	0.429	0.997	0.109	0.253	0.181	0.043

Table 4.2: Results for Simulation 1: Percent precision and recall for boundaries, word tokens, and word types on a heuristic extrapolating from utterance-final to word-final segments, compared with "upper bounds" using actual (supervised) word boundary information, and three baselines: one using distributions of word lengths, one segmenting only at utterance boundaries, and one segmenting at every canonically word-final phoneme.

4.5.2.2 Comparisons with Aslin, Woodward, LaMendola, and Bever (1996)

The results of Simulation 1 may be taken as a rough approximation of how well word boundaries may be predicted from unigram and bigram segmental information alone in the subset of Modern Greek heard by young children. These findings also provide a baseline for measuring and comparing the relative contributions of other cues such as those discussed in the remainder of this chapter. Moreover, it may prove interesting to compare these figures with those provided for English. A few comparisons with the AWLB96 data in Table 3.1 may be useful in this regard, although they should be interpreted cautiously given the differences in training and testing corpora (as well as in the threshold for boundary insertion).[12]

[12]This comparison also assumes that AWLB96 included utterance boundaries in their count for word boundary hits (which is not clear from their description). However, the results from a comparison excluding utterance boundaries from consideration (found in Rytting's (2004) second

AWLB96's results for the one-phone condition show nearly equal hits and false alarms—a precision of about 51%, as well as a recall of only 35%, for a balanced F-score of 41%—which they do not consider sufficient evidence of learning, although it is above their random baseline. The 1-phone Utt.BP condition in Table 4.2 shows better precision, and considerably better recall for word boundaries. In fact, the balanced F-score for the 1-phone result (61%) is comparable to that of AWLB96's results for their 2-phone context. This suggests that (in principle) a single phone of utterance-final context in Modern Greek is as useful as two in English; it may not suffice as a reliable cue on its own, but is enough to enable the learning of some words. Two phones of context, of course, are even better—comparable (with a balanced F of 70%) with three phones in English. Although full comparisons are difficult without having precision and recall results for word token and type for the AWLB96 study, it seems very likely that this cue is more useful in Modern Greek than in English—an unsurprising finding when the restrictive nature of Greek word-boundary phonotactics (and conversely, the comparative poverty of English suffixal inflections) is considered.

Since a direct comparison between this phone-level approach and a stricter replication of AWLB96 (i.e. one taking phonological features into account) was not performed, it is unclear to what degree phonological features play a role in this paradigm of Modern Greek word segmentation. However, the reasonable performance of this system, even without the explicit encoding of phonological features, suggests that they are not the *sine qua non* claimed in AWLB96.

pregenerals paper) confirms that, even with the most conservative assumptions, the model here performs no worse than the AWLB96 results.

4.6 Simulation 2: Segment predictability cues in Modern Greek

4.6.1 Purpose and methodology

Just as the Simulation 1 tested the general effectiveness of the cue proposed in AWLB96—namely, extrapolating from utterance-final segmental context to find word boundaries—to word segmentation in Modern Greek, Simulation 2 tests the next two variants (TP and MI) on the Modern Greek data. Brent's observation that mutual information works better than transitional probability for his English corpus data naturally evokes the question of whether Brent's result is specific to English, or if mutual information is a more robust measure of "segmental predictability" (or, more to the point, of conditional word boundary probability) than transitional probability in Modern Greek as well (and perhaps across languages generally).

While in its general approach the study reported here replicates the mutual-information and transitional-probability models in Brent (1999), it differs slightly in the details of their use. Whereas Brent dynamically updated his measures over a single corpus and thus blurred the line between training and testing data, the model presented here precompiles statistics for each distinct bigram-type offline over a separate training corpus before testing.

Like Brent (1999), but unlike Saffran et al. (1996b), all variants of the model in Simulation 2 focus on pairs of segments and not on pairs of syllables.

4.6.2 Three sub-simulations

Simulation 2a presents the "local comparison" found in Brent (1999) and described above in Section 4.2.3. That is, boundaries are inserted at points where the predictability measure within a segment-pair is lower than that of both "neighbors" (i.e., the overlapping segment-pairs on either side). No attempt is made to deal with the one-segment word problem discussed above. Two variant bigram-based cues are tested: TP and MI. For comparison purposes, a upper bound statistic of Hockema's conditional word boundary probability (computed with a global threshold of 0.5, without reference to neighboring pairs) is also included. The three lower bounds in Simulation 1 are also used as reference points.

Simulation 2b seeks to test the effect of comparison to neighboring pairs in application of the three variant cues in Simulation 2a. Instead of looking for "dips" in segmental predictability (or "spikes" in CWBP) with respect to the local neighboring context, a global threshold is used (optimized as before on balanced word token and type F-scores, using a development corpus for the optimization).

Simulation 2c re-examines the effect of neighboring pairs, but this time controls for the effect of the threshold optimization. It replicates Simulation 2a, adding an optimized threshold to present the over-segmentation found in Simulation 2a. In this simulation, the difference between a segment pair's relevant statistic and those of its neighbors must be higher than a certain threshold (optimized as before) in order for a word boundary to be posited.

4.6.3 Results

Simulation 2a: Local comparisons Table 4.3 shows that, while TP outperforms MI on boundary recall, MI performs better on boundary and word precision ($p <$

Cue Type	Context	Boundary		Word		Lexicon	
		Prec.	Recall	Prec.	Recall	Prec.	Recall
Simulation 2a							
TP ($\Pr[y\|x]$)	local*	0.540	0.851	0.259	0.407	0.183	0.160
MI ($I(x,y)$)	local*	0.580	0.763	0.301	0.396	0.171	0.221
CWBP ($P_{wb}(x,y)$)	global*	0.810	0.752	0.513	0.476	0.335	0.465
Simulation 2b							
TP ($\Pr[y\|x]$)	global	0.739	0.527	0.302	0.215	0.177	0.297
MI ($I(x,y)$)	global	0.858	0.488	0.368	0.210	0.196	0.292
CWBP ($P_{wb}(x,y)$)	global	0.776	0.798	0.495	0.510	0.344	0.433
Simulation 2c							
TP ($\Pr[y\|x]$)	local	0.655	0.686	0.307	0.322	0.183	0.263
MI ($I(x,y)$)	local	0.768	0.594	0.391	0.302	0.199	0.320
CWBP ($P_{wb}(x,y)$)	local	0.799	0.751	0.498	0.468	0.326	0.448

Table 4.3: Results for Simulation 2: Percent precision and recall for boundaries, word tokens, and word types on forward transitional probability and mutual information, compared with upper bounds using supervised conditional word boundary probabilities. Asterisked Conditions Were Performed With Unoptimized Thresholds.

0.001 for all comparisons), as well as on lexical recall ($p < 0.01$). Word recall and lexicon precision are approximately equal. Naturally, neither of these measures perform as well as the supervised upper-bound (CWBP). However, MI outperforms all three baselines (shown on Table 4.2) on all measures at the $p < 0.001$ level, except for lexical precision (for the Maximum and Minimum Boundary baselines) and lexical recall (for the length-based baseline).[13] TP performs similarly, though not quite as well: unlike MI, it fails to outperform the "Min#" baseline for word precision and lexical recall.

[13]This evaluation also excludes those baseline measures that are trivially perfect or nearly perfect: namely, boundary precision for "Min#" and boundary recall for "Max#".

Simulation 2b: Global thresholds Compared with the original heuristic (or "local comparison") suggested by Brent, using a global threshold significantly increases word precision and lexicon recall ($p < 0.01$ for all comparisons), though this at the expense of word (token) recall ($p < 0.001$). This is true for both MI and TP. Essentially, this particular optimization focuses on finding a wide variety of words (including less-common ones)without worrying about missing (at least some instances of) some more common words. However, the essential findings of Simulation 2a still hold: MI still outperforms TP (though only in word and boundary precision and not in lexical recall), and both still outperform the baselines.

Simulation 2c: Local comparsions with thresholds When the "local comparison" heuristic is also allowed to use a threshold optimized for word and lexicon F-scores, word recall improves significantly (compared to Simulation 2b), though it is still not as high as in Simulation 2a. However, this increase in word recall comes without significant penalty to either word precision or lexicon precision or recall. This is, then, the most optimistic (or generous) condition for viewing the performance of the MI and TP segmental cues. With this threshold, MI does marginally better than TP at lexicon recall ($p < 0.05$) and more clearly outperforms it in boundary and word precision, as before ($p < 0.001$). Both measures still lag far behind the upper-bound measure.

However, they are roughly comparable to the Utt.BP heuristic examined in Simulation 1 (using 2 phones of context). MI outperforms the 2-phone Utt.BP heuristic in boundary and word precision ($p < 0.001$) and lexicon recall ($p < 0.01$); it is significantly worse in the other three measures ($p < 0.001$). Indeed, a better comparison may be with the unoptimized threshold in Simulation 2a, where both MI an TP have roughly equal word precision and recall to this heuristic. In

this case, the Utt.BP heuristic outperforms MI in boundary recall, TP in boundary precision and lexical recall, and both in lexicon precision ($p < 0.001$ for all comparisons).

4.6.4 Discussion

The two distributional, segmental cues examined so far (extrapolation from utterance endings in Simulation 1 and various measures of "segmental predictability" in Simulation 2) are shown to be viable cues to word segmentation for Modern Greek. Both perform significantly better than baselines cues. Indeed, each performs about as well as the other. While Simulation 2 confirms Brent's findings that the MI variant outperforms the TP variant in some measures, this advantage was mitigated somewhat through use of optimized thresholds, suggesting that the exact statistic used to operationalize segmental predictability is not critical. Nevertheless, to use TP exclusively without examining MI may lead one to underestimate the efficacy of segmental predictability.

A comparison of the supervised, upper-bound cues—i.e. word boundary probability given two segments of prior context in Simulation 1 ($\Pr[\#|xy\_\_\_]$) or one segment of context before and after, as in Hockema's CWBP ($\Pr[\#|x\_y]$) in Simulation 2—suggests that these two cues would be of roughly equal effectiveness for Modern Greek. The former yields better boundary and word recall; the latter, better boundary and word precision ($p < 0.001$) and marginally better lexicon recall ($p < 0.05$). Lexicon precision is not significantly different.

It is worth noting also here that the results for this latter cue (Hockema's CWBP), while impressive when compared to unsupervised heuristics, are considerably lower for Modern Greek than for English. While Hockema (2006) reports boundary precision and recall at 86.5% and 76% and word precision and recall at

66.6% and 59.6% for an American English corpus, this same cue finds less than half the words (either token or type) in Greek, although the boundary recall is roughly the same. Precision (of boundaries, word tokens, or types) is also markedly lower for Greek than for English. This suggests that any advantage of a "balanced" CWBP over other supervised cues involving bigrams and labeled word boundaries (such as the upper bound in Simulation 1) does not necessarily generalize beyond English. For Modern Greek the difference, while significant for some measures, is relatively slight. While both CWBP and Pr[#|**S**___] would certainly be useful were the children able to obtain the necessary input, neither is the "silver bullet" that CWBP appears to be for English.

This same caveat applies for the two observable approximations of CWBP examined in Simulation 2 (MI and TP). Even when optimized thesholds are used, these two cues never reach the levels of performance that Brent (1999) reports for English. This suggests that by themselves, segmental ditributional cues are less valuable for word segmentation in Modern Greek than in English—not only in principle (as measured by the upper bounds CWBP and Pr[#|**S**___]), but also in practice (as shown by the MI and TP heuristics).

4.7 Simulation 3: Combining segmental cues in Modern Greek

4.7.1 Purpose and methodology

In Simulations 1 and 2 two general distributional cues were examined separately, and found to be roughly equal in efficacy for Modern Greek. Since they use different windows of input relative to the proposed word boundary, they should in principle tend to make different types of mistakes. Hence, a combination of the

Cue Type	Context	Boundary		Word		Lexicon	
		Prec.	Recall	Prec.	Recall	Prec.	Recall
Simulation 3a							
TP (Pr$[y\|x]$)	glob. + 1 ph.	0.522	0.641	0.346	0.311	0.215	0.327
MI ($I(x,y)$)	glob. + 1 ph.	0.683	0.697	0.350	0.358	0.224	0.309
TP (Pr$[y\|x]$)	loc. + 1 ph.	0.680	0.704	0.340	0.352	0.214	0.293
MI ($I(x,y)$)	loc. + 1 ph.	0.736	0.727	0.437	0.432	0.259	0.356
Simulation 3b							
TP (Pr$[y\|x]$)	glob. + 2 ph.	0.719	0.733	0.378	0.385	0.263	0.369
MI ($I(x,y)$)	glob. + 2 ph.	0.720	0.721	0.383	0.384	0.259	0.341
TP (Pr$[y\|x]$)	loc. + 2 ph.	0.677	0.724	0.348	0.372	0.222	0.295
MI ($I(x,y)$)	loc. + 2 ph.	0.734	0.739	0.441	0.444	0.264	0.360
$P_{wb}(x,y)$	loc. + 2 ph.	0.855	0.838	0.633	0.620	0.432	0.557

Table 4.4: Results for Simulation 3: Percent precision and recall for boundaries, word tokens, and word types on forward transitional probability and mutual information, combined with utterance boundary probabilities given one and two segments of context, and compared to conditional word boundary probability (CWBP) or $P_{wb}(x,y)$ with two segments of prior context

two cues should outperform either cue separately. Simulation 3 combines the Utterance Probability cue used in Simulation 1 with the cues used in Simulation 2 (MI and TP). Simulation 3a uses one phone of context for the first cue; Simulation 3b uses two.

As with the optimized-threshold conditions in Simulations 1 and 2, the thresholds used in Simulation 3 were optimized to maximize word and lexicon balanced F-scores. The relative weighting of the Utt.BP and Segment Predictability cues (MI, TP) were also adjusted to maximize this same metric.

4.7.2 Results and discussion

Simulation 3a: One-phone context The results of Simulation 3 are shown in Table 4.4. When these results are compared with those in Simulations 1 and 2, it

may be seen that the two cues in combination do perform significantly better than either cue in isolation. Combining the local-MI cue with the Utt.BP cue (using one phone of context) improves performance significantly over MI alone on every measure except lexicon recall ($p < 0.01$) and over Utt.BP alone on every measure ($p < 0.001$). TP performs similarly, although significant improvement is not seen for lexicon precision in this case. The local-MI cue benefits the most from the combination of cues, outperforming local-TP on every measure ($p < 0.05$).

Simulation 3b: Two-phone context However, adding a second phone of context helps considerably less. When the segment-predictability cues are combined with the two-phone Utt.BP cue, only the global MI and TP cues benefit. No further improvement is shown in either local-comparison MI or local-comparison TP relative to their performance when combined with the one-phone Utt.BP cue. Indeed, the figures are nearly identical.

The local-MI Utt-Prob combination is still superior to the 2-phone Utt.BP cue (when considered alone as in Simulation 1) on most—but not all—measures. This cannot be due to a ceiling effect since the combination cues still fall below both of the (separate) supervised upper bounds on most cues. When the upper bounds themselves are combined, the gap widens even more.

At first glance, it may seem counterintuitive that adding a second phone of context didn't improve the local-comparison MI and TP cues. However, when the morphology of Greek is taken into account, this result is no longer mysterious. It must be remembered that most Greek words end in some sort of inflectional suffix indicating relevant grammatical properties (gender, number, and case for nouns and adjectives; tense, number, and person for verbs). These suffixes are often two or more phones long. Since they are typical endings not only for words,

but for utterances as well, they naturally are encoded in the two-phone Utt.BP cue. However, since morphemes also consist of segmental bigrams (or ordered pairs of phones), these bigrams end up having higher-than-average MI and TP scores. The local-comparison heuristic looks not only for low MI (or TP) at the bigram around the proposed boundary, but also for high MI/TP scores at the bigram directly before. Hence, information about common Greek suffixes is also encoded by the local-comparison heuristic. When this heuristic is combined with the two-phones Utt.BP cue, this information is redundantly encoded in both cues, and no improvement is seen. The global-comparison heuristic, however, did not already use this information and hence does benefit somewhat from its addition.

4.8 General discussion

4.8.1 Evaluating the four variants

The simulations run in this chapter verify that segmental distributional information is of some utility for Modern Greek as it is for English—but not necessarily to the same extent or in the same ways. On the one hand, extrapolation from utterance boundaries, a cue explored by AWLB96 (and an essential component of CAS98), was found to be just as useful in Modern Greek as in English, and potentially even more helpful. Simulation 1 showed that levels of performance similar to those reported in AWLB96 can be achieved with shorter contexts. One segment of context in Modern Greek performed as well as two in English, and two in Greek as well as three in English. Although the size of the Greek dataset used did not allow three segments of context to be tested, it is possible that even better performance could be achieved thereby.

On the other hand, transitional probability and mutual information—two cues proposed by Saffran et al. (1996b) and Brent (1999)—did not perform as well in Modern Greek as reputed to do for English. The best performance for these cues in Simulation 2 (30.7% word token precision, 32.2% recall for TP; 39.1% word token precision, 30.2% recall for MI) is considerably below the figures reported in Brent (1999) (40% P, 45% R for TP; 50% P, 53% R for MI). Only when combined with the Utt.BP cue (as in Simulation 3) did they approach these figures. While part of this may be due to the differences in the optimization used and the properties of the individual corpora, this discrepancy seems too large to be explained by that alone. It seems more likely that this is a difference in the properties of the languages themselves.

For example, Greek syllable structures (briefly discussed in Section 4.4.1 above) weakens the effectiveness of both TP and MI when used alone to segment Greek. There are a few consonants in Greek (e.g. /p, f, k, x, r/), which, though rare in word-final position, nevertheless participate in a number of different consonant clusters word-medially. Since these consonants had a relatively large number of possible successors, some of the less frequent clusters (e.g. [ps], [tr]) had incorrect boundaries placed between their members.

In other cases segmentation errors in the TP model were precipitated by the phonemic transcription system used. For example, the palatalized lateral [ʎ] was transcribed in the Stephany corpus as the sequence [lj]. Whether children actually hear this sound as a sequence of two is an open question, but the great frequency of [l] in other contexts lowered the transitional probabilities for this "cluster"—with predictable results. This suggests the choice of units—or choice-points at which the system is permitted to consider placing a word boundary—may be crucial. However, in both of these cases, the use of knowledge about possible (or the most

probable) word endings, as extrapolated from utterance boundaries, ameliorated these problems and led to improved results when the cues were combined in Simulation 3.

4.8.2 Implications

This cross-linguistic difference in the performance of the "segmental predictability" cues is mirrored by a similar cross-linguistic gap in the theoretical performance of upper bounds based on conditional word boundary probabilities given the identity of the surrounding segments. Whereas Hockema (2006) demonstrated that nearly 60% or more of American English word tokens could be successfully segmented by this statistic alone if knowledge of these probabilities is assumed, the corresponding figures for the Stephany corpus show that barely half the word tokens, and less than half of the word types, are correctly segmented by this statistic alone. Only when it is combined with the other supervised, "upper bound" statistic $\Pr[\#|\mathbf{S}\_\_\_]$ (as shown in the last row of Table 4.4) does it reach the levels reported for English.

While a thorough examination of the differences between English and Modern Greek phonotactics leading to these differences in performance is beyond the scope of this dissertation, it may be briefly noted that a cursory examination of the distribution of the CWBPs for Greek bigrams explains this difference in performance. The distribution of CWBPs for American English is very strongly skewed to the edges of the probability space, with over half of the extant bigrams having a P_{wb} of either 0.02 or less, or 0.98 or greater—that is, either nearly always straddling a word boundary (about 33%), or hardly ever doing so (about 20%). The same distribution in Modern Greek, while still bimodal, is skewed in the other direction, with half of the extant bigrams with $P_{wb} \leq 0.02$, but only 13% with $P_{wb} \geq 0.98$.

When this distribution is weighted by each bigram's frequency, the distribution remains bimodal (with modes near 0 and 1) in English (see Hockema 2006, Figure 2b) but ceases to be so in Greek: the mode near 1 disappears. Hence, while the CWBP, or observable approximations thereof, may help one learn which bigrams are *not* likely to straddle word boundaries in Greek, cues focusing on the ends of words (such as those extrapolated from utterance boundaries) seem better suited to finding where the word boundaries *are*.

A weakness of all these models and heuristics is an over-reliance on transcribed data, combined with an implicit assumption that infants are able to identify, without error, the identity of each segment as it is transcribed. Naturally, there are several routes one could take to ameliorate this weakness. One would be to reduce the number of distinctions that need to be made, by using "broad classes" such as those suggested in e.g. Shipman and Zue (1982); Carlson et al. (1985). Such an approach is briefly explored in Briscoe (1989), though never extensively tested on a large-scale corpus. Another is to extend the time-scale of granularity from segments to syllables. Swingley (2005), noting research suggesting babies are in fact better able to track the repetition of whole syllables than of segments (cf. e.g. Eimas 1999; Jusczyk et al. 1995), takes this second route. Naturally, certain problems of syllabification arise if one adopts this strategy, though these are probably not insurmountable.

The first section of the following chapter (5.1) explores a third possibility, which keeps the assumption of segments as the basic unit and also avoids backing off completely to broad classes. This third possibility is to generalize the heuristics here discussed to include segments with probabilistic identities as opposed to fixed, symbolic identities.

4.9 Conclusion

In this chapter, several variants of the segmental distributional cues have been examined, namely: extrapolation of word-final phone distributions from utterance-final phone distributions (as suggested in Aslin et al. 1996) and two variants of Zelig Harris' 1954; 1955 segmental predictability heuristic (mutual information and transitional probability). These have been compared with two supervised or "upper-bound" statistics: conditional word boundary probability given either (a) the preceding segmental context only, or (b) the bigram straddling the potential boundary. Each of these variant statistics for these cues has been directly calculated over a training corpus, and the resulting table directly applied to a test corpus (rather than learned indirectly through ANNs or other machine learning techniques), in order to examine the relative value of the cues themselves. Finally, several combinations of these were also examined, to see how these cues might complement each other.

Overall, the technique of extrapolating potential word boundaries from utterance boundaries was found to be more promising for Modern Greek than previous studies have found it for English. Conversely, mutual information and transitional probability were found to be somewhat less effective for Modern Greek than for English, although still better than the extrapolation from utterance boundaries. The supervised upper-bound cues for these two approaches were found to perform about equally well to each other, but not so well as the (balanced) conditional word boundary probability for English as measured by Hockema (2006). Finally, a combination of the cues was found to improve results significantly over either cue alone. However, this improvement was not as much as might be expected *a priori*, due to redundancies of coverage among the various cues.

CHAPTER 5

PRESERVING SUBSEGMENTAL VARIATION IN A CONNECTIONIST MODEL

This chapter argues that one advantage of this class of model—namely, the ease with which continuous, multi-dimensional, and probabilistic input (as opposed to discrete, symbolic input) may be represented —has not been fully utilized.

Section 5.1 discusses the problems with transcription-based inputs that assume perfect certainty of the segments' identities (5.1.1), and some previous attempts to extend models to address these problems (5.1.2). These attempts are argued to be inadequate. Three desired characteristics of a more adequate model are proposed in Section 5.1.3.

Section 5.2 outlines how subsegmental variation present in the acoustic signal might be preserved in a way that is still compatible with the types of inputs connectionist models such as CAS98 expect and proposes a new model for input representation that uses phone probability vectors rather than symbolic transcriptions. Section 5.2.1 defines these vectors and Section 5.2.2 shows how they may be applied to the statistical heuristics used in Chapter 4. Section 5.2.3 demonstrates how the phone probability vectors may be transformed into probabilistic phonological feature vectors for a closer replication of CAS98 and CCC05. Finally, Section 5.2.4 explains how they may be obtained from the signal by means of a variant of automatic speech recognition (ASR) known as automatic phone classification (APC).

5.1 Towards a model of natural input variation

5.1.1 The problem

One major weakness shared by the connectionist models reviewed in Chapter 3 and the direct statistical heuristics examined in Chapter 4 (along with the majority of computational models) is the overreliance on transcript-derived, symbolic representations of the input the child is receiving rather than a more appropriate focus on the audio signal itself. In order to understand why, let us revisit the "Mondegreen" example found in this work's title and the introduction. As shown in Example 2, the miscomprehension of the lines of the poem famously reported by Wright (1954) depends not on mere missegmentation of a string of segments, but on several of those segments being pronounced (or at least understood) non-canonically.

(2) a. Intended utterance: *. . . and laid him on the green.*

b. Canonical and recognized phonetic strings (aligned):

(#) - æ n d # l e͡ɪ d # h ɪ - m # a n # ð ə # g ɹ i n #

(#) - æ n d # l e͡ɪ d - - i # m - a n # d ə - g ɹ i n #

c. Perceived utterance: '. . . and Lady Mondegreen.'

In this case, all the deviations from a dictionary pronunciation of these words can be explained by the sorts of variation in pronunciation often found in speakers. The deletion of /h/ from *him* and the hardening of /ð/ to [d] in *the* (along with the deletion of /v/ in *of* in the preceding line) are all quite believable casual pronunciations of those words and hence would very likely be fully accounted for in a phonetically transcribed corpus. Likewise, the change from /ɪ/ in *him* to make

the /i/ in *Lady*, while a rather unlikely slip of the tongue (except in hyperarticulated speech), is still explainable. First of all, /le͡ɪdɪ/ could have been the canonical and expected pronunciation for *lady* in Ms. Wright's dialect, although /le͡ɪdi/ is more typical now.[1] But even if not, the distinction between /i/ and /ɪ/ is slight enough (especially in unstressed syllables) that the same acoustic signal could be interpreted as either /i/ or /ɪ/ even by trained transcribers; 180 such disagreements were found in the Buckeye corpus (Pitt et al. 2005, Table 2).

However, the point that misperceptions—or at least uncertainty—at the segmental level can contribute to word missegmentations is not in doubt, as it may be may be observed in any number of other attested mondegreens: /n/ is perceived as [m] in the mishearing of *goodness and mercy* as 'Good Mrs. Murphy'; numerous segmental mishearings are required to transform *a single block of steel* into 'a single glockenspiel' and convert *life is but a dream* into 'like a butter stream'. And yet such mishearings happen not infrequently, even among adult native speakers.[2] This is not to say that misidentification of the segments directly causes word missegmentation; it is also possible (in adults at least) for the expectation of a particular word to influence how the signal is interpreted (and hence what phones are perceived). The point rather is that the potential for uncertainty and ambiguity at the segmental level, which arises from subsegmental variation in the sound signal, complicates the entire interpretation process at all levels—in ways that previous models do not address. For example, the subsegmental variation endemic to ordinary speech often leads to sounds that are ambiguous between [n] and [m] or

[1] Kenyon (1950, qtd. in Long and Trudgill 2005) notes that the tensing of /ɪ/ to /i/ word-finally occurred in the United States sometime in the early twentieth century; this tensing occurred even more recently in southern British English (cf. Harrington et al. 2000).

[2] See Jon Carroll, "Zen and the Art Of Mondegreens" Friday, September 22, 1995, available at `http://www.sfgate.com/cgi-bin/article.cgi?file=/chronicle/archive/1995/09/22/DD61909.DTL`.

[i] and [ɪ], since the acoustic cues differentiating them are very slight, and some overlap between them is possible.

Naturally, in everyday speech, such mistakes are relatively rare. However, when non-ideal hearing conditions mask what distinguishing cues are present in the acoustic signal, the resulting ambiguity is greater and the effect of misinterpretations of the signal may be seen more clearly. For example, Cutler and Butterfield (1992), examining adults' mishearing of very soft stimuli) demonstrate that missegmentation at the word level is often accompanied by mishearings at the phone level. However, some cues are more robust in poor hearing conditions than others. For example, prosodic cues are often still interpretable when segmental cues are lost. Cutler and Butterfield found that their participants were often guessing at the phones (or the words), but could still hear the stress pattern accurately. A model of word segmentation that is protected from making the sorts of perceptual errors or misinterpretations that people make—or from the ambiguities that arise in the signal due to subsegmental variation—is in some danger of overestimating the performance of a heuristic that depends on "error-free" input and inversely risks underestimating by comparison those cues that are more robustly salient in non-ideal conditions.

5.1.2 Previous work

Recognizing this problem, several researchers have dealt with this limitation of their models by devising various approximations to the actual (audio) input that children receive and testing their models on this approximate input. For example, CA97 (discussed above in Section 3.2.3) improves upon CAS98 in this regard by using phonetic transcriptions from a corpus of spontaneous speech. This captures on the phonetic or segmental level what was actually said (to the transcribers'

best approximation). This in itself, however, does not completely demonstrate the model's ability to deal with great variability in input. The total number of distinct input vectors encountered by the model is still equal to the number of distinct phonetic symbols used by the transcribers. In order to add further variation in the inputs, a certain subset of the binary features were randomly flipped with specific probabilities. (Only those features deemed to be "peripheral"—in the sense that they did not of themselves encode the distinction between the segment in question and some other phone in the transcriber's set—were allowed to be flipped.) This increases the difficulty of the learning problem (in principle) by increasing the number of distinct input vectors that the model receives as input.

However, this manner of adding variation is still quite artificial, particularly in the assumption that each peripheral feature is equally likely to be "misheard" or misinterpreted in the input representation. Another problem is that all of these changes occur independently. Such a method of modeling the variability in the speech signal is quite unlike the type of variation that actually occurs. Furthermore, it proves to be quite a mild test of the model as very few of the input features at any given time are actually changed. It also rules out those mishearings which do result in the perception of a different phone from what the transcriber wrote down, such as the preceived [m] for /n/ in the 'Good Mrs. Murphy' confusion discussed above. Finally, it makes such confusions as /f/-/θ/ and /r/-/w/ (cf. Chapter 3, note 11) less likely (given that they require two features to "flip" randomly and independently) than warranted by the phone confusion tendencies reported by Graham and House (1971).

Another way of modeling the noise inherent in the speech signal (as opposed to the idealization represented by textually derived input) is to use the output of an automatic speech recognition (ASR) system. Traditional ASR systems

are not well-suited for use as such a model for the input since they already have a large knowledge base about the pronunciations and relative frequencies of many thousands of words encoded in a lexicon and language model. Since the stage of the word segmentation problem under question is prior to the acquisition of all but a few of the most common words, access to such resources is unrealistic. Rather, a more appropriate system is an automatic phone recognition (APR) system or phone transcriber, which, given only some acoustic models for the phones in the language (and, optionally, some information about the relative frequency of these phones), generates the most likely sequence of phones for a particular utterance.

Such an experiment was reported by de Marcken (1996), who developed a phone transcriber based on HTK (Young et al. 2002). This automatic transcriber was trained on the 3,696 utterances from the TIMIT corpus (Garofolo et al. 1993) and tested on the WSJ1 corpus (Charniak et al. 2000). While the nature of De Marcken's method does not allow for easily comparable performance metrics, qualitatively it shows interesting results—far from perfect, but suggestive that some lexical structure can be automatically learned even from the raw audio output of adult-directed speech. (Even some of its mistakes are interesting–the model is thrown off by such word collocations as 'last year' and 'want you'—precisely because of the palatalization at the boundary point.) Of course, with the availability of child-directed audio corpora, more meaningful comparisons are possible. De Marcken's model, or a close variant of it, is in fact tested on CDS data in Simulation 5 (Section 6.2). However, it still lacks an essential element of human speech that we want to capture—subsegmental variation. The importance of preserving this variation in the input is discussed in the next section. Sections 5.1.3.1 and 5.2.4 describe in more detail how subsegmental variation may be preserved and represented automatically using APR and variant systems.

5.1.3 Desiderata for a model of input variation

One possible approach would be to take a stream of raw audio input, and attempt to segment that directly (cf. e.g. Gold and Scassellati 2006). However, comparing the performance of such a system to the current systems that do assume symbolic, segmental input quickly becomes non-trivial and hard to quantify: how many milliseconds off from the "true" word boundary counts as a missegmentation? And where is the true word boundary, anyway—is it to be determined by a human transcriber, by a consensus among several transcribers, or by automatic transcription? It is perhaps because of such difficulties that many segmentation systems based more directly on the signal have only done very limited evaluation.

For these reasons, certain assumptions are made in these models to make evaluation of results and comparison with other models more practical. Within the constraints imposed by evaluation on larger corpora, there are three essential properties of a model of input variation that can and should be modeled:

1. Variation that occurs below the level of the word (and particularly below the level of the segment) should be preserved.
2. Areas of clear and unclear speech should be distinguished.
3. Reliance on (adult, native-speaker) transcribers should be minimized.

Each of these three desiderata will be discussed in turn.

5.1.3.1 Preserving subsegmental variation

Models that use only canonical or dictionary-based representations of word pronunciation end up representing each instance of a given word identically and equivalently. This makes it easy to find words simply by looking for recurrent

patterns. Indeed, under this paradigm, the words that occur most frequently and in the greatest variety of contexts will be the easiest to find.

Natural language is not so simple. In natural language, the most frequent words—function words like *and*, *the*, and *is*—are produced with astonishing variation (Greenberg 1996). These are also words which children produce relatively late in their language development (though this likely influenced by other factors as well, such as the difficulty of mapping these words to meaning and usage norms).

A good phonetic transcription, such as the Carterette and Jones corpus used in AC97, avoids one part of this problem by preserving a portion of the variation to be found in the pronunciation of spontaneous speech—but only those types of variation that are large and obvious enough to change the value of a phone in the minds and ears of the transcribers. Transcriptions by their nature still treat all instances of phones as identical, removing variation below the level of the segment. A narrower transcription, with a correspondingly larger set of transcription symbols or *phoneset*, can ameliorate the problem; however, this only goes so far and does so at the cost of a much slower and error-prone transcription process.

Subsegmental variation, while difficult to model well, is nevertheless important to model as accurately as reasonably possible, because it is not simply noise in the system. Subsegmental variation is not random; it follows certain patterns or tendencies. Some of these patterns (for example, the partial devoicing of word-final /z/ in English as well as differing degrees of coarticulation across versus within word boundaries) may help signal word boundaries as a separate cue. (See Section 2.3.4 for more examples.) On the other hand, it also makes the finding of recurrent patterns considerably more difficult, given that the recurrent patterns can no longer be counted on to be identical or even to fall within some

relatively small number of possible variant pronunciations. Since it is not obvious *a priori* whether the facilitating or the complicating factors of subsegmental information will prove the stronger and more influential, one cannot know ahead of time whether excluding them will lead to an overestimation or an underestimation of the heuristic's true performance potential—or both for different sets of words.

CSCL97 and CA97 recognize that subsegmental variation is not completely arbitrary; therefore, their models of this variation are not completely random. Rather, they apply a mixture of rules or constraints and random processes. However, as the particular habits of individual speakers and speech communities (with probabilistic "rules" or tendencies of varying strength) are exceedingly hard to capture, it is more straightforward and in the end more accurate to use the speech itself (when available) as directly as possible. Automatic methods, such as APR, make this possible by taking the raw output and, through a series of transformations, producing a probabilistic input which may then be made comparable to the input used by other models in a variety of ways (discussed below).

For example, the APR system may simply take the single most likely label for a segment and assign that as the phone's identity. This representation method may be referred to as a "hard decision" and is equivalent to using a phone-level transcription. Since the representation is symbolic, it may be used by systems that deal with symbols; this is precisely how de Marcken used his phone transcriber to provide input for his symbolic word segmentation system. Another example of such a system is seen in Simulation 5 (Section 6.2, below).

Alternatively, the APR system can be made to return "soft" or continuous measures (e.g. posterior probabilities) of how likely it is for the segment to be an [i] or [ɪ] or any other phone. By returning all the probabilities as a vector, the phone

recognizer can return a "soft decision" (so called because it is not final but retains in consideration multiple alternatives that could be reweighted given further evidence).[3] In full-scale ASR systems, the use of soft decisions at the phone and word levels allows for the combination of higher-level linguistic knowledge, such as syntax or semantics (or even just knowledge of word frequencies) to play a role in determining the most likely sequence of words. The exact differences between standard ASR and phone-based variants will be discussed further in Section 5.2.4. For the present use (word segmentation models), a probability vector is one convenient way to preserve subsegmental variation in the input representation.

5.1.3.2 Distinguishing clear and unclear speech

The preceding section noted in passing that some words (usually more frequent words) are more likely to be produced with greater variation than others. Many of these same words (particularly those that do not receive lexical stress) may also be produced with less volume and less clarity than other, more prominent words. While some individuals may enunciate more clearly in their CDS than in ADS, this may not be true for all speakers and all cultures.

While subsegmental variability and segmental clarity are closely related, the distinction is worth making, at least in principle. Segmental variability refers here to the range of possible realizations of a word or segment made by a speaker. Segmental clarity refers to the likelihood of the listener correctly perceiving what the speaker said. One might also distinguish another related concept, *segmental confidence*, which refers to the degree of surety the listener has in the phone he

[3]Technically, a soft-decision APR system does not return a vector, but a *lattice* that also contains information concerning probabilistic boundary points—for the phone boundaries and their locations, as well as the phone identities, are also probabilistic in an APR system. A soft-decision automatic phone classification (APC) system (a variant of APR, discussed in Section 5.2.4 below, that assumes "hard" phone boundaries) does return a phone probability vector for each phone.

thinks he heard relative to the other phones it might have been. In practice, it may not be possible to distinguish rigidly between subsegmental variation and segmental clarity or confidence, although the former applies most properly to variation across the whole corpus, whereas the latter can be applied to a particular segment.

One useful side effect of the soft-decision input representation suggested at the end of the preceding section (that is, a vector of probabilities rather than a discrete symbol) is that it distinguishes implicitly between clear-cut instances of phones (e.g. phones that were obviously /f/ or obviously some other thing) and the intermediate or difficult-to-determine cases (whether due to true indeterminacy—e.g. f_1 and f_2 values halfway between /ɪ/ and /ɛ/—or because of slurring, soft or muffled speech, background noise, and so on). The Carterette corpus cannot do this: the transcriber must record a single symbol.[4]

A further corollary of this (explored further in Chapter 7) is that a "segmental confidence score" for each segment can be easily extracted from each probability vector. This value could become a valuable cue in its own right, if we suppose (hypothesize) that children may pay greater attention to stretches of clearly articulated (or easily understandable) speech, such as that found in stressed syllables (cf. e.g. de Jong 1995), than to unclear speech. Transcribed corpora (as noted above) typically do not allow for this type of information (unless one makes assumptions by extrapolating from other phenomena supposed to be associated with segmental confidence or clarity—stress, part of speech, position in the word or utterance, and so forth). Naturally, a phone recognizer may not be perfect at this: spurious measures of indeterminacy (low confidence) may result from a mismatch of the mothers' CDS with the ADS training material used to form the acoustic models.

[4]In theory, the transcription could record inter-transcriber disagreements, confidence scores, etc. but in practice few if any corpora actually do this.

However, this problem may be mitigated to some degree by adapting the automatic phone recognizer to the mothers' speech.

5.1.3.3 Minimizing human involvement

When working with large corpora, it is a principle of sheer practicality to minimize the amount of labor- or decision-intensive human effort whenever possible. There are a number of reasons for this. One is the simple matter of cost in terms of person-years of effort: phonetic transcription is slow-going work, and producing an adequate transcription of any appreciable size can take years of preparation (even with a small army of trained transcribers). For example, the Buckeye corpus, which contains spontaneous adult-directed speech, reports 80.3% overall agreement ($kappa = 0.797$) with unanimous agreement on 62% of the segments. For vowels, this drops to 69% agreement ($kappa = 0.66$) and 49% unanimity (Pitt et al. 2005). This is for a corpus that took over five years to annotate. Word-level transcription, while still time-consuming, is much easier and does not require extensive prior training in phonetics.

Accuracy is another inherent problem in transcription projects. Phonetic transcription is tedious, fatiguing work, and even expert transcribers occasionally make mistakes—a separate issue from the differences of opinion represented by the disagreement statistics above. Naturally, ASR systems also make mistakes—many of them—but their mistakes are consistent and (in principle) predictable.

However, another issue (counterintuitively) is that adult, native-speaker transcribers know too much. Naturally, transcribers are trained to identify phones as objectively as possible, but any native speaker cannot help but be influenced by the fact that they know the words, and hence hear words (and not just phones) when listening and transcribing this speech. This bias is often a good thing, but it

injects into the transcription task (i.e. phone-identity judgments) knowledge of the language that the infant does not yet have (and cannot yet be expected to have).

Traditional ASR systems also use knowledge of words and grammar, encoded as a vocabulary (or pronunciation dictionary) and a language model, to improve their accuracy at guessing the identity of particular phones. However, with machines, this knowledge can be easily attenuated or turned off completely. Automatic *phone* recognizers differ from automatic *speech* recognizers not only in their tasks, but also in the way higher-level linguistic knowledge is used.

5.2 The proposed model

5.2.1 Use of phone probability vectors

While there are a number of models that would fulfill the desiderata listed above, perhaps the most straightforward approach (in terms of keeping it comparable to earlier models that assume phones as discrete symbols) is to assume some set of phones, not as symbols or labels as such (in the sense that each segment can have only one label), but as dimensions in a probability space. For such a model, we will assume a corpus of audio recordings of child-directed speech. Since we will want to compare our results to the discrete, symbolic input in the "traditional" models discussed above, we also assume access to a transcription for each utterance in the corpus.

Then we assume that each utterance in our corpus is made up of some known number of segments. For convenience, we set this number equal to the number of segments or phones used in the symbolic (phone-based) transcription, if one is available, or to the number of segments one would expect to find given the canonical pronunciation of each word in the word-level transcription. We will call

this number T. As T is not the same for every utterance, the value for a specific utterance u may be notated T_u.)

It is assumed for the time being that the infant listener has already acquired an adult-like phonemic inventory, such that he or she treats each of the symbols in the chosen transcription alphabet as separate categories or symbols over which statistics might be collected as in the models above. We will call this inventory or alphabet of symbols A. It is also assumed that the boundaries between each segment (as well as the number of segments in each utterance) are correctly identified. This allows us to divide up the utterance into T sections, s_1 to s_T. However, the assumption that each instance of each segment is correctly identified is relaxed and replaced with a probabilistic assumption: that every stretch of sound between two adjacent segment boundaries is represented by a probability distribution $\Pr[Q|X]$ over the alphabet of phones used in transcription, where Q signifies the phone, and X the acoustic signal between the two given phone boundaries. For example, if a particular segment at position t (that is, in section s_t) were correctly and unequivocally recognized as an [a], then the probability distribution Q at segment position t could be represented by the vector $Q_t = [1, 0, 0, 0, \ldots]$, signifying $\Pr[q_t = [a]|X_t] = 1; \Pr[q_t = [b]|X_t] = 0$, and so forth. This is the special case in which the symbolic models discussed above have operated. However, a more typical case might be the situation when the phonetic identity of the segment given the acoustics (and no other cues) is somewhat uncertain: for instance, the probability vectors for two adjacent segment positions (for example, the two segments around the boundary in /bi#gʊd/ *be good*) may be something more like that shown in Example 3.

(3) a. $\Pr[Q_t|X_t]$ = [[i]: 0.8, [ɪ]: 0.1, [e]: 0.05, [ɛ]: 0.05]

b. $\Pr[Q_{t+1}|X_{t+1}]$ = [[g]: 0.6, [k] 0.2; [d]: 0.1, [b]: 0.1]

5.2.2 Adapting segmental heuristics to probabilistic input

The use of phone probability vectors is compatible in principle with any model of word segmentation that uses statistics over phones to evaluate the probability of a word boundary at discrete "choice-points" such as phone boundaries. For illustration, this section describes how the statistical heuristics in Chapter 4 would be adapted under this new approach. Using the phone probability vectors described above, any statistic $S(x, y)$ of the type examined in the preceding chapter can still be calculated by using weighted averages of the statistic in question. Assume for the moment that the infant listener has, through prior experience, learned some statistical model M for statistic S, which lists the particular values for $S(x, y)$ for all $x, y \in A$, the alphabet of symbols or categories the listener is assumed to be using. When applying this statistic S_M to determine the probability of a word boundary in the face of uncertainty as to the underlying phone identities, the value of S_M that should be used is the expected value for S_M given the phone probability distributions involved, as shown in Equation 5.1.

$$\mathbb{E}(S_M) = \sum_{x \in A} \sum_{y \in A} \Pr[Q_t = x, Q_{t+1} = y | X_{t:t+1}] S_M(x, y) \tag{5.1}$$

Assuming conditional independence of Q_t, Q_{t+1} given $X_{t:t+1}$, this equation can be simplified to Equation 5.2:

$$\mathbb{E}(S_M) = \sum_{x \in A} \sum_{y \in A} \Pr[Q_t = x | X_t] \Pr[Q_{t+1} = y | X_{t+1}] S_M(x, y) \tag{5.2}$$

Just as discussed before in Chapter 4, $\mathbb{E}\left(S_M[t]\right)$ for some segment position t can then be compared either to some threshold θ (in the global comparison case) or to its temporal neighbors $\mathbb{E}\left(S_M[t-1]\right)$ and $\mathbb{E}\left(S_M[t+1]\right)$ (in the local comparison case).

These equations provide one method for generalizing the cues discussed in Chapter 4 into a probabilistic framework given some corpus-derived statistical model M. But how does one learn M, given access only to these same probabilistic inputs? The answer is that the phone and phone-bigram frequencies on which the models rest can also be constructed from probability-weighted input. Instead of using $\text{Freq}(x)$ for phone (unigram) frequencies, the following equivalent may be used for a corpus of length N:

$$\text{Weighted Freq}(x) = \sum_{i}^{1:N} \Pr[Q_i = x | X_i] \tag{5.3}$$

As long as a combined probability mass of 1 is added at every time count, the weighted frequencies of all the symbols in A will still total to the length of the corpus, and the probabilities generated by dividing these frequencies by the length of the corpus will likewise total to 1—just as the canonically calculated probabilities used above do. With the independence assumption, the same will be true of the bigram frequencies and probabilities, as given in Equation 5.4:

$$\text{Weighted Freq}(xy) = \sum_{i}^{1:N-1} \Pr[Q_i = x | X_i] \Pr[Q_{i+1} = y | X_{i+1}] \tag{5.4}$$

This approach can only be applied on a corpus with probabilistic segmental data available. Canonically transcribed corpora such as the Korman or Stephany corpora do not have data from which phone probability vectors may be generated, unless some ad-hoc method for mapping word transcriptions to a probabilistic set of phone sequences were derived. In principle, corpora with multiple transcribers

could also use the inter-transcriber disagreement to generate estimates of phone probabilities. However, it is not clear that either of these solutions would be much preferable to the ad-hoc approximations of noise used by CSCL97 and CA97. With corpora that have the original audio recordings available in sufficiently high quality, phone probability vectors can be obtained, as described in Section 5.2.4 below.

5.2.3 Extending the model to featural representations

The preceding section has shown how phone probability vectors can be applied directly to the types of statistical heuristics given in Chapter 4. This section outlines an analogous extension for ANN-based WST approaches, such as the AWLB96 and Christiansen models. Since these models use phonological features rather than phone identities in their input representations, some additional transformations are required. To begin, The whole sequence of phone probability vectors for some utterance u, $\mathbf{Q}_{1:T_u}$—or even for the whole corpus—can be treated as a matrix. In this latter case, utterance boundaries can be represented with special utterance boundary markers (UBMs), as done before in the AWLB96 and Christiansen models. Under the assumption that utterance boundaries are in fact long enough to be recognized unequivocally by infants, the vector representation for each UBM at the end of an utterance is simply $Q_{T+1} = [0, 0, 0, \ldots, 1]$, where the last dimension of the vector is a special one reserved for utterance boundaries (as signalled by silence or non-speech). Hence the number of dimensions needed for each vector is $|A| + 1$—one dimension for each phone in A, and one for the final (ubm) dimension.

In order to convert these phone probability vectors into the featural input-vectors used in the connectionist models discussed above, we may arrange the set of phone-to-feature-vector mappings used for these models (displayed in the

Appendices of CAS98 and CCC05) as matrices with one row for each phone and one column for every phonological feature used in the ANN's input. We then add one row for the special UBM symbol, and one column for a special UBM input feature. Call the resulting matrix $\mathbf{F}$. Similarly, we may arrange the sequence $Q_{1:T}$ as a matrix with T rows for the time-points and $|A| + 1$ columns for the phone probabilities. We then can calculate a probabilistic (continuous-valued) input for the SRN through matrix multiplication, as shown in Equation 5.5:

(5.5) $\mathbf{G}_{1:T+1} = \mathbf{Q}_{1:T+1}\mathbf{F}$

For each segment position t, the corresponding row vector G_t contains the mean expected feature activations for each feature (between 0 and 1) at that position given the probability space Q_t. Such vectors are still well-formed for the input of an SRN, although they are less commonly used in this domain than are binary vectors. In the case of the localist ("one-hot") representation, the row vectors of the matrix $\mathbf{Q}_{1:T+1}$ may be used as they are. (Thought of another way, the phone-to-feature mapping that the localist representation assumes is simply the identity matrix $\mathbf{I}$ of size $(|A| + 1) \times (|A| + 1)$.)

5.2.4 Obtaining the phone probability vectors

Up until this point, we have not discussed where the phone probability vectors Q come from. Again, a straightforward choice (in terms of practicality in meeting the desiderata discussed above) is the use of mainstream ASR technology as it has been developed over the last three decades. Section 5.1.3.1 introduced APR systems and briefly discussed how they could be used to produce automatic transcriptions of speech consisting either of sequences of discrete symbols (through hard-decision APR, as in de Marcken 1996) or of probabilistic input (through soft-decision APR,

the method used e.g. in Roy's CELL model). Since it is the probabilistic input that preserves the subsegmental variation which is of interest, soft-decision APR is more relevant. However, as noted in footnote 3, soft-decision APR does not return a sequence of vectors, but rather a lattice that encodes not only phone identities, but also phone boundary points probabilistically. This has the effect of making the very choice-points for the word segmentation task probabilistic rather than discrete. This not only increases the difficulty of the parsing task, but also complicates its evaluation and the interpretation of the results. However, the phone probability vectors needed for this task can be supplied by a close variant of APR known as automatic phone classification (APC) (cf. Halberstadt and Glass 1997). As with other ASR variants, a variety of techniques can be used, including HMMs (see below).

APC, like APR, differs from standard ASR in the basic unit of recognition and the types of linguistic knowledge that it utilizes. In traditional ASR, the basic unit is the word, and the task is to identify the sequence of words most likely to have given rise to the utterance (or sound signal) just heard. In addition to a phoneset (or set of phones used for transcribing the language in question) and an *acoustic model* describing the (ranges of) audio input associated with each of those phones, an ASR system has a vocabulary of words and their usual pronunciations (sometimes with multiple listings for a single word) and some sort of grammar or *lexical model* to describe how those words are likely to be sequenced together.[5]

In APR and APC, the basic unit is not the word but the phone. Hence, no pronunciation dictionaries or lexical models are used. This is appropriate to the word segmentation task because we cannot assume that babies have a vocabulary

[5]For example, lexical models could encode a grammar of syntactic or semantic rules to explain how words fit together in a sense, or they can simply (as is usually done) consist of a list of frequencies for single words, pairs of words (or *bigrams*), and sometimes word triples or *trigrams*.

yet; the goal is to find the words in order to build one. APC differs from APR in that the former assumes that the number of phones in an utterance and their boundary points are known. Hence, for each phone position and its associated temporal slice of audio signal, discovering the phone's identity may be construed as a classification task. APR does not make this assumption, but allows for phone insertions and deletions, just as typical (word-based) ASR allows for word insertions and deletions. This is a less constrained task—and perhaps a more realistic approximation of the task before the infant—but it generally yields lower performance. In addition, in a soft-decision APR system, since it has probabilistic rather than pre-determined phone boundaries, the choice-points for word boundaries are probabilistic rather than a discrete set, which makes evaluation and comparison with transcription-based WST models much less straightforward.

In both APR and APC, just as in standard ASR, the audio file corresponding to a particular utterance is transformed into a sequence of acoustic feature vectors called mel frequency cepstral coefficients (MFCCs) (cf. e.g. Davis and Mermelstein 1980), by means of a variant of the Fourier transform over mel-scale frequency bands. (For a thorough discussion of these terms and the general process of applying HMMs to ASR, see Young et al. 2002.) Note that both the time-scale and the feature vectors used here are quite different from those in the final vectors being generated for the SRN input. The vectors used for generating the MFCCs cover time-slices of 10 ms, whereas the segmental time-steps used in connectionist approaches are of variable length since they correspond to the length of time taken by one phone. This duration may be assumed for the purposes here to be at least 30 ms, and often is much longer. To avoid confusion, the discussion will henceforth refer to the connectionist time-steps as *segment positions.* The acoustic

features used are likewise completely different from any of the phonological feature systems assumed in the connectionist frameworks, being derived (after the Fourier transform) from levels of acoustic energy in particular frequency bands.

It is thus necessary to convert the audio signal (in its MFCC representation) into the sequence of phone probability vectors needed for the proposed model's input. This process is conducted in two stages: one to find the phone boundaries, and one to generate the phone probability vectors within each boundary. If the audio corpus happens to come with pre-marked phone boundaries, these may be used. In the more typical case (as is the case with the corpus used in Simulation 6, below), only the utterance boundaries are marked, and only a word-level transcription is supplied. In this case, the first stage relies on the *forced alignment* of the audio file for each utterance with a phone transcription derived from the pronunciations in the pronunciation dictionary—in this case, CMUdict (CMU 1993–2002). The process involved for APR, APC, and forced alignment within the HMM paradigm may be explained by use of an analogy.[6]

Step 1: Finding the phone boundaries Suppose you have a friend who visits particular cities while traveling across the world. She emails you every day, but never tells you which city she is in on a particular day. However, she does tell you, in excruciating detail, what the weather conditions (e.g. temperature, humidity, pollution, and pollen count) are where she is. In the general APR task, you try to guess which cities your friend visited from the weather reports alone (aided by an almanac that tells you what the whether is likely to be like at that time of year in each city). In the APC task, you know how many times she moved from city

[6]For a thorough introduction to HMMs as applied to speech recognition, see Rabiner (1989). Many helpful tutorials for HMMs can also be found online, e.g. `http://www.ee.surrey.ac.uk/Personal/P.Jackson/tutorial/`

to city, but not what the cities are. In forced alignment, you have the itinerary of cities listed in order, but without the dates and durations of each visit. However, you can find the most probable dates for each city by using the almanac and the daily weather reports.

The cities in this analogy are the phones—or rather, the *states* in the HMM.[7] The daily weather reports correspond to the sound in the audio file or, more precisely, the 10-ms acoustic feature vectors. In place of the almanac, an HMM uses an acoustic model, which describes the probability of a particular MFCC-vector being part of a particular phone. These probabilities are known as *emission probabilities*. The probabilities of moving from one state to another are known as *transition probabilities.*

The best (most probable) path through the all of the states, given the observed sequence of acoustic (MFCC) feature-vectors, may be calcuated by using Equation 5.6, where X^* is the best path, $\mathbf{O}$ is the sequence of observations (MFCCs), M is the (acoustic) model, a_{ij} is a transition probability from state i to state j, and $b_i(o_t)$ is the probability of observing some observation o_t when the model is at state i. In the case of forced alignment, we are interested in finding the boundary points—that is, each segment position t given X^* when $x_t \neq x_{t+1}$ or, more precisely, when x_t and x_{t+1} belong to different phones.

$$X^* = \arg\max_X \Pr[\mathbf{O}, X|M] = \arg\max_X \left\{ a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(\mathbf{o}_t) a_{x_t x_{t+1}} \right\} \tag{5.6}$$

[7]For the purposes of exposition, it is convenient to think of each phone as having a unique state associated with it. Actually, in the typical tri-state model, each phone in this model has three separate states associated with it: one for the beginning of the phone, one for the middle, and one for the end. When context of neighboring phones is taken into account (a *triphone model*), extra states are added for each combination of three phones. However, the systems used in preparing input for Simulations 5 and 6 (described further in Chapter 6) use only a monophone model.

Step 2: Calculating the phone probability vectors Once the (most probable) phone boundaries are found, the automatic phone classification task begins in earnest. The goal of this second step is to find out how well each phone fits the audio data in a given segment—or in other words, how likely it is to have been the intended phone communicated by the given slice of sound (set of obervations **O**). This may be calculated using Bayes' theorem:

$$\Pr[Q|\mathbf{O}] = \frac{\Pr[\mathbf{O}|Q]\Pr[Q]}{\Pr[\mathbf{O}]} \tag{5.7}$$

Since the denominator ($\Pr[\mathbf{O}]$) is the same in every case, it may be factored out of the equation. Also, this model assumes equal prior probabilities $\Pr[Q]$ for every phone. Hence, the likelihood $\Pr[\mathbf{O}|Q]$ and the posterior probability $\Pr[Q|\mathbf{O}]$ end up being essentially equivalent.

This second step was implemented by using a previously developed ASR system based on HTK (Young et al. 2002). Each time interval between two phone boundaries was treated as a separate utterance. The system was then contrained to treat each single-segment "utterance" as a single segment—no phone transitions were allowed. The total probability for a particular phone (given the observations in this single-segment utterance) is (in principle) equal to the sum of the probabilities for every path through that phone's states. In practice this is abbreviated by using the probability of the most probable path only, appropriately normalized so that the total probability for all phones in the phoneset sum to 1. More details concerning the implementation are given in Section 6.3.1.

It is worth emphasizing again that this model for calculating the phone probability vectors for each segment includes no linguistic knowledge besides the phoneset, segment boundaries, and the acoustic models. It has no knowledge of words or word frequency or other language models, nor any knowledge even of

probabilistic phonotactics or phone frequency. Each phone was assumed to be equally probable prior to observing the acoustic cues in a given segment. In fact, the acoustic models were not even adapted to the mothers in the corpus, but used as they were. Such a model leads to relatively poor performance when compared with state-of-the-art models. However, it is worth remembering that the purpose of this model is not to give the best possible phone classification results, but rather to preserve subsegmental variation in the signal and to model the sorts of uncertainty that an infant listener might experience. In this case, a model that overestimates this uncertainty (or underestimates the actual abilities of an infant) is not a bad thing: if the model nevertheless performs better than chance on the word segmentation task with noisier input and less certainty than an infant actually has to deal with, then the model has proved its effectiveness even more surely than if it had succeeded with less noisy data—unless the noise somehow makes the task easier, which is quite unlikely.

5.2.5 Non-essential aspects of the proposed model

The model proposed in this thesis, for the sake of convenience in evaluation and comparison, makes two assumptions. First, the proposed model assumes that the infant has some model of what phones are possible in its language—what categories or "bins" the sounds it encounters might be placed into—and where each individual segment starts and stops (along with how many segments are found in each utterance). Second, without assuming a close correspondence between human phone confusions and phone-level errors made by ASR systems, the proposed model does assume that the output of an automatic phone classifier preserves at least the general shape of subsegmental variation found in the signal as an infant might hear it.

A previously learned phonemic inventory As for the first of these two assumptions, it is unclear to what extent infants between 6-12 months have coherent categories of phones or phonemes by which individual segments may be identified or labeled—or whether this category set resembles that of adult native speakers. In the model here proposed, these categories are not used as labels, but they are still used as dimensions by which particular segments may be located in probability space, so the choice of an appropriate phoneset still may affect the model's performance and predictions. The model presented here makes no claims about the most appropriate phoneset. Since it is not known what phoneset is most appropriate for children, it uses a phoneset appropriate for adults, as do most of the previously discussed word segmentation models. Such an assumption does not seem terribly far-fetched, given infants' remarkable abilities of distinguishing sounds even a few days after birth (cf. Section 2.2.2). Regardless of the phoneset used, moreover, the use of continuous (real-valued) input features preserves the model's ability to represent a nearly unbounded degree of variation in the input.

Phone boundaries The model proposed above also assumes that the infant is able to identify phone boundaries that correspond at least roughly to those used in the phone-level corpus transcriptions (or pronunciation dictionaries). This insures that every transition between two phones is considered as a possible word boundary and excludes the possibility of a word boundaries being posited in the middle of a phone. Again, this assumption is not an essential claim of the model but is rather a convenience deemed necessary for evaluation and comparison for the reasons discussed above in Section 5.2.4: specifically, APR is less accurate than APC and greatly complicates comparisons with transcription-based models. However, a simulation using APR is reported in Section 6.2. The results in that simulation are

roughly comparable to those of APC (assuming phone boundaries). This suggest that, insofar as comparison with other models is possible, the prior assumption of phone boundaries does not change the overall pattern of results.

5.3 Conclusion

Section 5.2 proposes a model for replacing traditional symbolic input (or phonological feature-vector input derived from symbolic transcriptions) with real-valued phone probability vectors that preserve the subsegmental variation in the signal itself. Section 5.2.1 describes the properties of these vectors, and Section 5.2.4 delineates how they may be obtained or approximated automatically from the acoustic data. The implications of this new type of input representation are explored in theoretical terms in Section 5.2.2.

In the following chapter, the proposed input representation model will be tested empirically. Specifically, in Simulation 6 (Section 6.3) the new input representation is embedded in a version of the CAS98 SRN model and compared with an subsegmentally invariant input representation of the same corpus.

CHAPTER 6

REPLICATING AND EXTENDING THE CHRISTIANSEN *PHON-UBM* MODEL

Chapter 5 outlines a new model for input representations that can be applied to a corpus consisting of high-quality audio recordings, a word-level transcription for each child-directed utterance, and time-stamps marking the start and end of those utterances. It is argued that this input representation provides a more plausible approximation of the types of auditory cues infants are likely to have available than transcription-based inputs such as those reviewed in Chapters 3 and 4, while still maintaining sufficient similarity with those methods to allow for straightforward comparisons with their results.

The purpose of this chapter is to perform that comparison, by replicating a crucial portion of the simulation found in CAS98 involving the "phon-ubm" condition—the simulation in CAS98 which combined phonological (or segmental) distributional information with that provided by utterance boundaries, but did not include the lexical stress cue. Since the corpora used in previous studies do not meet the criteria listed above, a new corpus that does meet those criteria was identified and prepared for use in this model. This new corpus, the Brent corpus (Brent and Siskind 2001), is described in further detail in Section 6.2.1.

The bulk of the comparison with the original Christiansen model is performed in three simulations. Simulation 4 tests and confirms a basic (though not heretofore explicitly tested) claim of CAS98, and shows comparable results for the

"phon-ubm" variant of the Christiansen model on the Korman corpus. Simulation 5 tests the performance of the "phon-ubm" model on a subset of the Brent corpus under two conditions: one replicating the idealized assumptions of CAS98 with regards to (lack of) variation below the word level, and one adding some degree of variation, in a way distinct from that suggested in CA97, using "hard decision" APR. Simulation 6 extends the Christiansen model by replacing its near-binary input and target activation levels with continuous-valued activations (using "soft-decision" APC), allowing subsegmental variation to be modeled. Simulations 5 and 6 reveal that the "phon-ubm" model degrades in performance when presented with highly variable data, but that it still outperforms a length-based baseline on some (though not all) of the relevant performance measures.

6.1 Simulation 4: Verifying the utility of catalysts

One of the basic claims of AC96 (on which foundation CAS98 rests) is that training the network simultaneously on multiple cues constrains the hypothesis space considered by the networks and hence leads to quicker convergence to better solutions. While speed of convergence is not a primary consideration here, final performance on the task is.

If catalyst output units are beneficial in learning word segmentations for human language input, then networks using relevant catalyst units (such as those proposed in CAS98) should perform better than those trained and tested without such units – even with the same set of inputs. While CAS98 compares various combinations of cues combined with the utterance boundary marker (UBM) unit, it never investigates the performance of the UBM unit on its own.

Networks of this type have been investigated by Aslin et al. (1996). However, Aslin's study used MLPs with a fixed time window of 1, 2, or 3 preceding

segments, and were trained and tested on a different corpus, making comparisons to CAS98 less straightforward.

In order to verify the utility of the 36 catalyst output units corresponding to phone identity, two simulations are reported here. Simulation 4a uses Elman nets without these units, to verify that they perform above baseline. Simulation 4b uses SRNs with these units, to verify that they perform better than those in Simulation 4a.

6.1.1 Method

Input corpus Following CAS98, Simulations 1a and 1b used the Korman (1984) corpus, freely available as part of the CHILDES (MacWhinney 2000) collection of child-directed language corpora.

The CHILDES version of the Korman corpus was found to be somewhat larger than the figures reported in the original CAS98 study, containing 43,385 word tokens distributed over 13,350 utterances. (CAS98 reports 37,529 word tokens over 11,376 utterances.) However, the number of word types was found to be the same (1888), suggesting that the discrepancy was due to repetitive material, and is not expected to affect performance significantly. As with the original study, onomatopoeic forms and less frequent forms not found in the MRC Psycholinguistic Database dictionary were excluded, along with the utterances containing them. (Word-forms appearing more than three times in the Korman corpus were added to the dictionary, so that the utterances containing them could be included.) The resulting corpus contained 11,318 utterances, 35,248 word tokens and 875 word types (compared with Christiansen's 9108 utterances 27,467 word tokens, and 830 word types). This corpus was split $80\% - 10\% - 10\%$ into training, development, and test corpora. Table 6.1 shows the size of the corpora used.

Corpus	Utterances	Word Tokens	Word Types
Christiansen et al. (1998)			
training	8181	24,648	814
test	927	2819	379
total	9108	27,467	830
Simulation 4 Replication			
training	9056	28,230	860
dev.	1131	3508	483
test	1131	3510	497
total	11,318	35,248	875

Table 6.1: Size of the training and test corpora in terms of utterances, word tokens, and word types (cf. Christiansen et al. (1998), Tables 1 & 2)

Input representations In Section 3.3.2 some discussion was made concerning the two different phonological feature representations used for input in the Christiansen model: the 11-feature system used in CAS98 and the updated 17-feature system in CCC05. As these papers never explicitly compare these two input representations in terms of performance or plausibility, it is unclear why the 17-feature input is assumed to be preferable, or how much performance advantage it gives over the 11-feature phonological input features. Indeed, it is not clear that either of these representations is superior (in performance or in plausibility) to a completely localist "one-hot" representation, where the input matches the output in having a separate unit for each phone. An additional motivating goal for this and the following simulations is to investigate the relative performance of the various input representations.

Model and task In order to examine the effect of the catalyst output nodes on the Christiansen model, several simplified versions of the model were re-implemented

using the Conx toolkit, and trained and tested with data from the Korman corpus (omitting stress cues). Three main variations were used in the input representation: one using the 11-dimension phonological-feature vectors original to CAS98, one using a 17-dimension system used in CCC05, and one using the same localist (36-dimension "one-hot") representation used for the output.

Simulation 4a: Elman nets using one output unit For Simulation 4a, both the phonological features (or segmental identity, for the localist representation) and the absence or presence of an utterance boundary are provided as input features, but only the latter cue is used for training. These correspond to a recurrent version of Aslin et al. (1996) or to CAS98's "phon-ubm" training condition in the input, and "ubm only" in the output. The number of units in the hidden layer was adjusted in order to keep the number of parameters roughly constant among the three variants.

Simulation 4b: Elman networks using catalyst outputs For Simulation 4b, the phonological or segmental features (along with an extra node for the utterance boundary marker) are used as before in Simulation 4a for the input layer, but localist, segmental "catalyst nodes" are added to the output layer. These three variants thus correspond to CAS98's "phon-ubm" training condition (for the 11-feature condition), or the equivalent for the other models. Again, the number of units in the hidden layer was adjusted order to keep the number of parameters roughly constant among the three variants, and with the three variants in Simulation 1a. The exact numbers of parameters with changeable weights (not counting "bias" weights) are shown in Table 6.2.

Each of the six variants were trained on one pass through the training set, using the same settings as CAS98 (learning rate of 0.1, momentum of 0.95, and

Input Type	SRN	Total No. of Trainable Parameters
Simulation 4a: "phon-ubm" input, "ubm" output		
11-feature (CAS98)	12-96-1	10,464
17-feature (CCC05)	18-92-1	10,212
36-phone (localist)	37-84-1	10,248
Simulation 4b: "phon-ubm" input and output		
11-feature (CAS98)	12-80-37	10,320
17-feature (CCC05)	18-77-37	10,164
36-phone (localist)	37-70-37	10,080

Table 6.2: Set-up for Simulation 4, including figures for input, hidden, and output nodes, and total number of parameters

initial weight randomization ranging from -0.25 to 0.25), then tested (with the network weights "frozen") on the development set. In order to account for the natural variability in the networks, nine separate runs of training and testing were performed for each of the six variants, each differing only in the randomized starting weights.

6.1.2 Results

Since the network is supposed to generalize from utterance boundaries (as marked in the transcript) to all word boundaries, the activation of the output unit corresponding to the utterance boundary marker (UBM) is used to determine the model's level of belief in a word boundary before a given segment. Following Aslin et al. (1996) and CAS98, the threshold used for determining a posited word boundary is the average activation for the UBM output node over all positions.[1]

[1] AWLB96 and CAS98 make no claims either for optimality or for any special cognitive plausibility for this threshold. It seems to have been selected arbitrarily. It is used here to allow direct comparison with those previous models. An alternative approach would be to optimize for the balanced type/token F measure, as done in Chapter 4. (As runs using different initial weights may be expected have different optimal thresholds, each run would have to be optimized separately.) As

Utterance boundaries are automatically considered to be gotten right, regardless of the activation at that point; similarly, a spurious above-activation immediately following a boundary (before the first segment of the next utterance) is disregarded.

Results are reported for precision and recall (accuracy and completeness in CAS98's terminology) for boundaries, word tokens, and word types, as described in Section 4.8.1. Following CAS98, significance in precision and recall between two conditions is measured comparing $\langle N_{tp}, N_{fp} \rangle$ and $\langle N_{tp}, N_{fn} \rangle$ respectively, for the two conditions in a 2×2 Chi-square test. Unlike CAS98, who only reports one run, all simulations reported here take the mean precision or recall ($\langle \bar{N}_{tp}, \bar{N}_{fp} \rangle$ or $\langle \bar{N}_{tp}, \bar{N}_{fn} \rangle$) over all nine runs, unless otherwise noted.

Simulation 4a Table 6.3 shows that even without the catalyst nodes, each of the three variants of the SRN exceeds the length-based random baseline in mean precision and recall for both words and boundaries ($p < 0.0001$ for all comparisons). For lexicon (or "word type") precision and recall, the model's performance is not so clearly above baseline. While all variants have significantly better mean lexicon precision than baseline ($p < 0.0001$), their mean lexicon recall is somewhat below baseline—significantly so for the 12-96-1 SRN ($\chi^2 = 5.21, p < 0.05$) but not for the other two (18-92-1: $\chi^2 = 1.79, p = 0.1812$; 37-84-1: $\chi^2 = 1.6, p = 0.2057$).

Comparisons among the three variants show that the localist 36-phone representation performs better than the 11-feature variant, but worse than the 17-feature variant, for both mean word precision and mean word recall ($p < 0.01$ for all comparisons). However, they do not significantly differ in lexicon precision or recall.

this would require reserving portions of the corpora used (already small for Simulations 5 and 6) as development sets, this type of optimization was not performed, in order to preserve more material for the training and test sets.

Simulation 4b In order to see the effect of the 36 phone-based catalyst output nodes, the three SRN variants without these catalyst nodes are compared with three variants with those catalyst nodes. The first of these latter three variants (the 12-80-37 SRN) corresponds to the "phon-ubm" condition in CAS98; the other two differ from it only in input representation (with the number of hidden nodes adjusted accordingly).

For the 11-feature and the 36-phone localist input representations, the catalyst units improve performance for both boundary and word precision and recall ($p < 0.01$ for all comparisons), but no improvement is seen for the 17-feature input representation. No significant differences are seen for lexicon recall or precision for any of the input representations, with the exception of the 36-phone's improvement on precision ($\chi^2 = 5.4563$, $p = 0.0195$).

Interestingly, when the three variant input representation are compared, the ordering between them fails to hold: when the "catalyst" units are added, the localist 36-phone representation outperforms the 17-feature representation in boundary and word precision and recall ($p < 0.05$ for all comparisons), which is no better than the 11-feature representation. (Again, differences in lexical precision and recall are not significant.)

6.1.3 Discussion

The model for combination of multiple cues proposed in AC96 and CAS98 claims that it is not enough simply to have the cues available in the input; to gain maximal benefit from them, the network must be trained to learn to predict them (in a sense, to pay active attention to them). Indeed, Harris' "peaks of segmental unpredictability" cue is incorporated into the CAS98 model only insofar as the task

Network	Boundary		Word		Lexicon	
Simulation 4a: "phon-ubm" input, "ubm" output						
12-96-1	56.0	72.5	25.1	32.3	15.0	25.2
18-92-1	59.7	79.1	32.1	42.6	18.2	27.9
37-84-1	57.7	80.1	27.7	38.5	17.0	28.1
Simulation 4b: "phon-ubm" input and output						
12-80-37	58.8	79.0	30.1	40.3	18.0	29.1
18-77-37	59.8	80.7	31.3	42.1	18.8	29.4
37-70-37	63.8	82.9	36.4	47.2	21.9	32.4
Baselines						
Utt-as-Word	100	32.20	26.50	8.6	6.0	8.1
Length-based	48.5	54.6	18.9	21.3	9.2	31.9

Table 6.3: Percent precision and recall for the three nets trained with the utterance boundary cue with and without catalyst nodes, for an algorithm that treats utterances as words, and for a pseudorandom algorithm that predicts lexical boundaries given the mean word length

of actively predicting the next phone interacts with that of predicting the the utterance boundaries. Without this extra task, the model is simply a recurrent version of Aslin et al. (1996).

Simulation 4 confirms that these catalyst nodes do indeed help—but not as universally as might be supposed. It is unclear why they fail to improve performance in the 17-feature case, though it may be that the difficulty of learning the extra task interferes with its possible benefit. The somewhat more straightforward input mapping of the localist input may make the prediction task easier for the net to learn, and hence better mastered and more beneficial. Indeed, the 37-70-37 SRN shows lower overall SSE rates (combined for all output nodes) on the training and testing sets than either the 12-80-37 or 18-77-37 SRNs, suggesting that it is learning the segmental prediction task better than it would with distributed (feature-based) input.

Comparisons with the results reported in CAS98 for the "ubm-phon" condition reveal that these replicated models consistently oversegment relative to the original ubm-phon network in CAS98. For the networks with the 11- and 17-feature distributed input, this yields significantly worse performance in terms of both boundary and word precision, but significantly better performance in boundary recall.[2] For the network with the 36-phone localist representation, the deficit in boundary and word precision is not significant, but boundary and word recall are both improved ($p < 0.001$ for all comparisons).

Therefore, although the strictest replication of CAS98's phon-ubm condition (using the same input representations that they did) failed to match their results on all measures, the ability of the localist 36-phone model to match or even surpass the performance may be taken as evidence that the general approach is nevertheless replicable.

6.2 Simulation 5: *Phon-ubm* with recognized input[3]

Questions of distributed versus localist input representations aside, there is reason (as argued in Section 5.1.1) to doubt the realism of the input used in CAS98, based as it is on canonical dictionary pronciations of the words in a word-level

[2]Given that these deficits in word precision are primarily caused by oversegmentation relative to the original CAS98 SRNs, it is very likely that these deficits could be eliminated by as simple a matter as adjusting the threshold. As there is nothing of obvious theoretical significance about the assignment of the threshold value to the mean activation level, such an adjustment would have little if any theoretical import. However, as no principled alternative method of assignment for the threshold presents itself, and finding the optimal threshold would require additional validation data, this matter is left as an open issue of marginal importance.

[3]This section reports on experiments originally reported in (Rytting 2006a), with an updated and expanded analysis of the results.

transcription of the corpus. Although CA97 addresses one of the problems of variation (namely, non-canonical speech at the phone level) by using linguists' phone-level transcriptions of conversational speech, the study uses an arguably unrealistic model of sub-segmental variation, by assuming (a) that production and perception of segments is binary at the featural level: e.g. that either a phone is $[+voice]$ or $[-voice]$, with no gradations in betweeen, (b) hence no variation is possible in the "core features" of a segment, as changing these would render the linguists' transcription incorrect, and (c) change amongst the "peripheral" features is both binary (all or nothing bit-flipping, rather than gradiant variation) and independent of other featural changes.

In Sections 5.1 and 5.2, it is argued that automatic phone recognition and classification techniques, while clearly far from perfect at mimicking human performance, are still likely to be more realistic approximations than the method in CA97. Hence, testing the Christiansen model of WST on input taken from an automatically transcribed corpus of CDS is a more realistic test of the model.

One of the drawbacks of the Korman corpus is the lack of usable audio recordings of the speech from which automatic transcriptions could be derived. Although audio recordings are available through TalkBank (MacWhinney 2000), they are too faint and unintelligible (even for human listeners) to be of any use for ASR. The audio recordings for the Brent corpus (Brent and Siskind 2001), discussed in the next section, are much more suitable for automatic transcription.

6.2.1 Materials

The Brent corpus provides fourteen 90-minute sessions for each of eight American English-speaking mothers living in Baltimore who participated in the study, spaced at roughly two week intervals at 8-14 months of the baby's age. For each of

the mothers, the middle 75 minutes of the earliest sessions (typically three or four) were transcribed at the word level.

The mothers' voices were recorded using portable DAT recorders and lapel-mounted microphones placed on the mother. The recordings took place in the families' home, in order to capture typical utterances of everyday life; in order to focus on speech directed at the child, the mothers were asked to avoid phone conversations with other adults. The mothers' utterances were defined as stretches of speech demarcated on either side by at least 300 ms of silence; time stamps for utterance boundaries are marked in the word-level transcriptions. (See Brent and Siskind (2001) for more details.)

This simulation (as reported in Rytting 2006a) used three of the eight mothers' voices found in the Brent corpus. In order to keep the input focused on speech directed to infants of as young an age as possible, only the first few recordings from these mothers were used. The very first recording for each dyad is excluded, since the mother may have been more self-conscious due to the novelty of the microphone; recordings 2-5 from mothers "c1", "f1", and "f2" were used. The infants ranged between 9 and 10 months of age at the time of these recordings. These twelve sessions contain a total of 8285 utterances.

As with the Korman corpus, no phone-based transcriptions were available, so these had to be created. Two versions were created: a canonical reference transcription created in the same manner as in CAS98, by replacing each word with the word's canonical pronunciation, and an automatically (and imperfectly) "recognized" transcription input created by a phone recognizer as described below.

6.2.2 Method

6.2.2.1 Generating the input for the model

In order to create the input for this simulation, the SONIC Speech Recognizer (Pellom 2001; Pellom and Hacioglu 2003) was used as the basis for a phone recognition system. A version of the CMU dictionary adapted to the SONIC phoneset was used as a base dictionary. Words in the transcripts not found in this dictionary were added to it, with the pronunciation checked by consulting the original sound files as necessary. To form the canonical input, each instance of a given word was transcribed using the first pronunciation of each dictionary entry. These pronunciations were then mapped from the 55-phone "TIMITbet" phoneset that SONIC uses to the 36-phone MRC phoneset used in Christiansen et al. (1998), in order to facilitate comparisons with that study.

To create the "recognized" input, each of the sound files was segmented and resampled with a 8 kHz sampling rate. SONIC's default feature extraction system was used along with an off-the-shelf acoustic model for female speech provided with SONIC. No adaptation or re-training of the acoustic model was done, except for SONIC's on-line adaptation. A triphone language model was used, taken from 90% of the utterances used in the study; however, relatively little weight was given to the model.[4] Obviously, no dictionary was given to the recognizer, as that would defeat the purpose of the pre-lexical word segmentation task (which of course is to find the word boundaries so as to facilitate the acquisition of a lexicon). In place of a dictionary, SONIC was given the same 55-to-36 phone mapping that was used for

[4]Although children may be assumed to be developing an implicit knowledge of transition frequencies between phones—indeed, Chapter 4 is based on this assumption—it is unclear how much this knowledge should be relied on for determining the phone identities themselves. The use of the triphone language model was minimized in order to keep the input as true to the raw acoustics as feasible.

Mother	Utterances	Word Tokens	Word Types
Canonical Pron.			
c1	2087	7476	882
f1	2324	8414	922
f2	2882	8125	733
total	7293	24,015	1576
Noisy Pron.			
c1	2087	6731	3519
f1	2324	7384	3632
f2	2882	7526	3144
total	7293	21,641	8234

Table 6.4: Size of the Brent corpus subset in terms of utterances, word tokens, and word types

the canonical phones, with each of the 36 phones treated as a word entry. Settings were adjusted to produce roughly as many phone insertions as phone deletions.

Although the Brent corpus recordings are more suitable than those of the Korman corpus, they are still susceptible to the quality lapses inherent in naturalistic, out-of-the-soundbooth recording situations. Instances of background interference (e.g., crying children) and changes in the mother's volume are not infrequent, causing significant misrecognition in segments. In fact, certain utterances failed to be recognized by SONIC (which gave either null output or recognized a single non-continuous, non-sonorant phone like [t]). Such utterances were excluded from both the canonical and recognized input sets. Of the 8285 utterances, 992 were excluded, leaving 7293 utterances.

Exact figures for the utterances used for each of the three mothers is found in Table 6.4.

Even with these problematic utterances removed, the resulting recognized transcript was extremely noisy: the correctness measure (calculated as in Equation 6.1) was 43% and the accuracy (Equation 6.2) was 23%. As a result of this noise, each of the 1576 word types in the canonical input had on average five different "recognized" realizations. Naturally, such noisy input makes for a very demanding test of any word segmentation model's abilities. In the word types actually found in the test set, however, this proliferation of variant forms is less severe: the 589 canonical types yield 1322 distinct pronunciations.

(6.1) $$correctness = \frac{hits}{hits + deletions + substitutions}$$

(6.2) $$accuracy = \frac{hits - insertions}{hits + deletions + substitutions}$$

6.2.2.2 Preparing the input for the model

Both the recognized and canonical input sets were divided 90%-10% into training and test sets, with 6564 utterances in the training set and 729 utterances in the test set. The canonical sets had 77952 and 8903 segments in the training and test sets, respectively; the recognized sets 77303 and 8866 segments. All utterance-internal pauses were deleted from both the canonical and recognized input strings, and a pause symbol was inserted if missing at the end of each utterance (symbolizing the 300ms pause at the end of each utterance, which should be clearly audible to the child in the context of the running conversation). In actual practice, the utterance-internal pauses are expected to be an additional helpful cue; removing them makes the results a more conservative estimate of the corpus' segmentability.

In evaluating performance of the word segmentation model for the case of recognized input, one may wonder where word boundaries ought to be placed in

the gold standard. For this study, the recognized transcriptions were previously aligned with the corresponding canonical transcriptions for those same utterances using the HTK (Young et al. 2002) tool `HResults`, in order to evaluate the performance of the recognizer. These same alignments were used to assign the gold-standard word boundaries for the recognized transcriptions. In each recognized transcription, a word boundary was placed before each segment aligned with a word-initial segment in the canonical transcription of that utterance. Resulting one-segment "words" at the beginning and ends of utterances were combined with the following or preceding word, as most of these were spurious and corresponded to extra material on either side of the utterance.

In the following example, the '#' symbol indicates word boundaries, and the hyphen indicates an insertion or deletion in the alignment.

(4) a. Canonical utterance: "You're a wild man today, huh?"

b. Canonical and recognized phonetic strings (automatically aligned):

(#) - j ʊ r # ə # w a ɪ - l d - # m æ n # t ə d - - e ɪ - # h - - ʌ #

(#) n a ʊ - # - # w a ɪ a m d ɪ # ŋ æ n # - - d s ʌ m ɪ ŋ # h a ʊ n #

c. Segmented "recognized" transcription: naʊ # waɪamdɪ # ŋæn # dsʌmɪŋ # haʊn #

6.2.3 Training and testing the model

Training and testing were done as with Simulation 4 above. The same three types of input representation (11-feature and 17-feature distributed representation, and 36-phone localist) were used. The "ubm-only" category was not tested; the two levels of corpus transcription ("canonical" and "recognized") were both evaluated using SRNs with 37 output units, corresponding to the "phon-ubm" condition of

Simulation 4. The simulation reported here followed the adjustment of the number of hidden and context units described in Simulation 4 in order to keep the number of parameters relatively constant, whereas those reported in Rytting (2006a) used 80 hidden units and 80 context units for each of the SRNs. Hence, the results for the 17-feature and 36-phone input representations will be slightly lower here.

6.2.4 Results

6.2.4.1 Simulation 5a: The canonical transcription

As in Simulation 4, the 37-output-unit networks corresponding to the "phon-ubm" condition clearly perform better than the baselines. For all three input variants, the SRN trained and tested on the canonical transcription significantly outperforms the length-based baseline on boundary and word precision and recall, and also on lexical (word type) precision ($p < 0.001$ for all comparisons). Differences in lexical recall are not significant.

The networks also outperform the single-word baseline in boundary, word and lexicon recall ($p < 0.0001$ for all comparisons) and for lexicon precision ($p < 0.05$ for the 11-feature variant, $p < 0.01$ for the 36-phone localist variant, and $p < 0.001$ for the 17-feature variant).

As with the "ubm-only" condition in Simulation 4a, the 17-feature input representation yield the best results. The 17-feature representation outperformed the 11-feature representation on boundary and word precision and recall, and the 36-phone localist representation on boundary and word recall – though not for precision ($p < 0.01$ for all significant comparisons). No significant differences in lexical precision or recall were found among the three input representations. Results are shown in Table 6.5.

Network	Input	Boundary		Word		Lexicon	
ubm-phon SRNs with "canonical" input							
12.80.37	Canonical	0.555	0.675	0.237	0.287	0.177	0.274
18.77.37	Canonical	0.580	0.752	0.272	0.353	0.209	0.289
37.70.37	Canonical	0.597	0.682	0.264	0.301	0.191	0.303
ubm-phon SRNs with "recognized" input							
12.80.37	Recognized	0.481	0.645	0.159	0.213	0.239	0.170
18.77.37	Recognized	0.472	0.650	0.155	0.214	0.235	0.161
37.70.37	Recognized	0.507	0.634	0.177	0.221	0.234	0.171
Baselines							
Length-based	Canonical	0.460	0.513	0.160	0.179	0.104	0.277
Utt-as-Word	Canonical	1.000	0.292	0.266	0.078	0.127	0.119
Length-based	Recognized	0.441	0.508	0.117	0.135	0.136	0.161
Utt-as-Word	Recognized	1.000	0.324	0.299	0.097	0.267	0.140

Table 6.5: Percent precision and recall for the three nets trained and tested with a canonical, dictionary-based transcription and an automatically phone-recognized transcription of a three-mother subset of the Brent corpus, compared with two baselines for each transcription

6.2.4.2 Simulation 5b: The recognized transcription

The networks trained and tested on the automatically recognized transcription do not fare as well as those using the canonical transcription. When each variant SRN (by input representation) is compared with its canonical counterpart, it performs worse on all measures except lexical precision ($p < 0.001$ on all comparisons except boundary recall for the 11-feature variant, where $p < 0.05$). While they appear to do better on lexical precision, this is an artifact in how lexical precision and recall are measured for the recognized condition, as will be discussed below.

In contrast with the canonical condition, there are no consistent or significant differences among the three input variants in the noisy condition, except that

the 36-phone representation outperforms the 17-feature representation for boundary and word recall ($p < 0.01$, $p < 0.05$, respectively). Thus, any advantage of a featural representation is not retained when highly variable data of this type are used.

Comparing the networks in the recognized condition yields mixed results. Compared to the canonical condition's length-based baseline, the networks do better than expected on boundary and word recall, but not on precision, except boundary precision in the case of the localist 36-phone variant ($p < 0.01$ on all comparisons). Against the single-word baseline, they fare somewhat better, but do not clearly outperform them. However, it could be argued that comparison against the canonical baseline is inappropriate, since the large numbers of insertions and deletions in the phone recognizer's output (combined with the combination of the stranded, "spurious" single phones at utterance boundaries with the adjoining word during corpus preparation) changed the distribution of word and utterance length, and hence the relative effectiveness of the baseline strategies. In addition, the values for lexical precision and recall differ considerably and cannot be compared with the canonical case, since the definition of a word type is different: for the recognized transcription, every distinct (recognized) pronunciation of a word counts as a separate word type. As noted briefly above, this yields 1322 distinct types rather than 589.

In order to control for this potential mismeasurement, a new set of baselines were calculated for both baseline strategies using the same methods as before, but on the recognized transcription of the corpus rather than the canonical one. The results are considerably different, and a new comparison reflects this. Analogous with the canonical case, all three variants of the "recognized" SRNs outperform the baseline based on the recognized corpus' word lengths on all measures except

lexical recall ($p < 0.05$ for all comparisons). However, the differences between the nets and the baseline are not as great as they are in the canonical case. With the single-word baseline, the networks again do better on recall (except lexical recall in the case of the 17-feature network), but not on lexical precision, unlike the "canonical" networks.

6.2.5 Discussion

The results above are difficult to interpret with surety. On the one hand, the SRNs trained and tested on recognized transcription outperform their length-based baseline, and even (in word and boundary recall) the canonical transcription's baseline. However, these differences are not as large as for the canonical condition. This is particularly true when the recognized condition's SRNs are compared to the two single-word baselines, where performance is only marginally better if at all (insofar as gains in word recall are more than balanced out by losses in word precision). Furthermore, inasmuch as the networks in the canonical and recognized conditions can be compared against each other (overlooking the differences in the properties of the resulting corpora), the canonical condition's networks are clearly superior.

Still, the fact remains that Simulation 5 is not properly informative on its own. While it suggests that there is a degradation of performance when we use recognized input, it is not clear that the two corpora generated are really comparable in a straightforward way. In order to keep them comparable (and avoid the whole issue of separate baselines and un-intuitive definitions of word types), it is best to keep the number of segments between gold-standard boundaries constant, and remove the confounding factors that arise from insertions and deletions

of phones in the recognition. One way to accomplish this is to use not a phone recognizer, but a phone classifier.

6.3 Simulation 6: Modeling subsegmental variation

Although CA97 and Simulation 5 each give us some indication of the Christiansen model's performance in the face of certain types of variation in its input, neither is a particularly good model of the actual input inherent in speech. The CA97 experiment relaxed the "phone invariance" assumption (that each instance of each phone be represented the same) but varied it in a rather artificial way, preserving the notion of "all or nothing" (or binary) activation for each input unit, as well as assuming independence of errors in the set of "peripheral" features. Simulation 5, replaced the assumption of error independence with an input representation more closely tied to the audio signal: the output of an automatic phone recognizer. However, it still assumes perfect confidence in the phone chosen (whether that phone choice is correct or incorrect), and hence retains both all-or-nothing activation and a type of phone invariance. The phone recognizer's insertions and deletions also created differences in the underlying input corpora sufficient to significantly complicate comparison of results.

One way to approximate the raw acoustic input from speech that children might experience, while still preserving some of the assumptions necessary to make the input comprehensible to the Christiansen model and comparable with previous results on the same corpus is to use the output of an automatic phone classifier. In the phone classification task (cf. e.g. Halberstadt and Glass 1997), the number of phones in each utterance, and the location of the boundaries between them, is assumed to be known (or derivable from some other source). What is not

known is the appropriate label of each segment. Higher-order linguistic knowledge such as a language model or word identity are also not used, as it would be unrealistic to assume that such are available to a young infant. Phonotactic constraints (in terms of triphone modeling) are also excluded from the phone classifier's knowledge base, unlike the phone recognizer used in Simulation 5.

6.3.1 Materials[5]

While the method used for representing the input in Simulation 5b is arguably more realistic than the canonical word-based transcriptions used in CAS98 and Simulations 4 and 5a, it nevertheless represents only a small step toward preservng the subsegmental variation found in the signal. Like the preliminary experiment in de Marcken (1996), it still uses "hard decisions" over a finite phoneset, and hence only captures variation below the level of the word. Variation below the level of the phone is still lost.

Simulation 6 implements the model proposed and described in Section 5.2 to overcome this limitation. In this simulation, as in Simulation 5, the Brent corpus is used because the acoustic quality of the Korman corpus is not sufficient to allow for ASR techniques to produce reasonable results for phone classification. Simulation 5's subset of speakers (*c1*, *f1*, and *f2*) was augmented with audio data from speaker *q1*. Additional care was taken in controlling for the background interference and other occasional lapses in recording quality incident to naturalistic recordings.

[5]Technical assistance by Eric Fosler-Lussier and Soundararajan Srinivasan gratefully acknowledged in designing the automatic speech recognizer used for Simulation 6.

Simulation 6 uses HMM-based acoustic models trained on the TIMIT corpus (Garofolo et al. 1993) of read speech. These acoustic models were used to phonetically align the Brent corpus, performing a forced-alignment on the canonical pronunciations found in the dictionary used for Simulation 5. The resulting phonetic boundaries were utilized to segment each utterance into individual phones. The average frame likelihood of the 20 best monophones was then calculated for each segment; these likelihoods were converted into posteriors by normalization.[6] After a conversion from the 55-phone TIMITbet to the 36-phone "MRCbet" (described in more detail below), these posteriors were then utilized as inputs to the neural network, with each posterior serving as an input activation (in the localist 36-phone case). Weighted averages of the 20 best monophones' feature representations (after phoneset conversion) were taken to form the activations for the distributed-input SRNs.

While in typical ASR tasks it is unusual to first train the models on the same material that will be evaluated, it should be noted that what we are trying to derive is an approximation of the phonetic confusability in the acoustics. Thus, if the models are trained on one phone but during testing they prefer another, this is a clear indication of acoustic confusability, and we can have more confidence that misrecognitions are not due to training/test mismatch.

In order to further increase the confidence in these phonetic materials, utterances that did not have good performance in phone classification across the entire utterance were discarded.

The performance of the phone classification across an utterance was calculated using a measure called *approximate accuracy* defined as the number of phones

[6] As mentioned in Section 5.2.4, since it is not clear whether one should posit that infants have developed a prior model over phones $\Pr[Q]$, it is assumed for this experiment that $\Pr[Q]$ was uniform (and thus $\Pr[Q|X] = \alpha \Pr[X|Q]$ for some constant α).

correctly detected within the top two guesses for each phone. Using this definition rather than exact accuracy allows for more of the desired variation while ensuring that the correct phone was a good candidate, suggesting that the automatic phonetic alignment process was valid.

Two subsets of the Brent corpus were created: one that has utterances of approximate accuracy of at least 33.3% and that had more than one phone classified correctly (hereafter called "Brent33"), and a second, higher-confidence corpus that had an approximate accuracy of at least 60% per utterance (hereafter called "Brent60").[7] Each of these two subsets were further divided $90\% - 10\%$ into training and test corpora, as shown in Table 6.6, below.

Corpus	Utterances	Word Tokens	Word Types
The 60%-Accuracy Subset (Brent60)			
training	2861	6443	782
test	316	740	258
total	3177	7183	819
The 33%-Accuracy Subset (Brent33)			
training	7030	22193	1493
test	781	2486	552
total	7811	24679	2592

Table 6.6: Size of the training and test corpora for the two Brent corpus subsets in terms of utterances, word tokens, and word types

[7]The 60% cutoff point was chosen to represent a subset of acoustic roughly comparable to that found in low-noise speech corpora such as TIMIT. It is hypothesized that the types of variation found in Brent60 correspond to a large degree to those resulting from normal variation speech such as reduced or casual pronunciations of words, allophony, and dialect differences. In contrast, it is hypothesized that Brent33 contains a higher proportion of actual ASR errors resulting from background noise, poor microphone placement, and the like.

6.3.2 Method

The method of training was the same as for Simulations 4 and 5, with the exception that the input and target vectors for the "recognized" condition used were no longer binary or near-binary, but continuous (rounded to four decimal places) in the range $[0, 1]$. As with Simulation 5, a conversion from the 55-phone TIMITbet to the 36-phone MRC system (including the mapping of TIMITbet's diphthongs and affricates to two-segment sequences) was necessary in order to keep the phoneset the same as that used in CAS98. However, the details of how this conversion was implemented differ considerably between Simulations 5 and 6.

One difference between TIMITbet and "MRCbet" which prevented the use of a straightforward conversion scheme was TIMITbet's use of separate symbols for stop closures and stop bursts. Since Simulation 5's input and target data were essentially treated as sequences of symbols anyway, combining the stop closure-burst sequences into a single symbol did not change the nature of the input in any substantive way. However, for Simulation 6, where every segmental vector directly represents a segment-sized temporal subsequence of the audio signal, capable in theory of communicating co-articulation and other sub-phonemic cues which might well be important to the correct identification of word boundaries, removing or combining these in the phone conversion process would result in the loss of interesting and potentially critical information. To solve this problem, two new symbols were added to "MRCbet": `C` and `c`, symbolizing voiced and voiceless stop closures, respectively. In the SRNs' input and target representations, `C` and `c` were treated as equal mixtures of {[b, d, g]} and {[p, t, k]}, respectively.[8]

[8]One problem left unsolved here arises from the tendency of voiceless stop closures to be confused with silence. In this model, since silence was used as a signal for utterance boundaries, the posteriors for silence were assumed to be 1.0 at all utterance boundaries, but ignored within an utterance, in order to remove additional information or cues from pauses from the "recognized"

To prevent the additional segmental unit of stop closures from adding artificial difficulties in the word segmentation task relative to the corpora used in other studies and in Simulations 4 and 5 (which do not include stop closures), an additional step was taken in the evaluation process to prevent spuriously posited word boundaries between a stop closure and its burst from being counted. Under this modified evaluation procedure, a word boundary before a stop is considered correct if it is posited *either* before the stop closure or immediately afterward (before its burst or another following consonant). Additionally, two consecutive false-positive word boundaries before and after a stop closure are counted as a single false-positive.

While this special handling of stop closures may seem ad-hoc and somewhat generous in terms of evaluation, it prevents the absurd situation of positing "words" consisting only of a stop closure (and no non-silent segmental information). Such a restriction is arguably of narrower application than (for example) a constraint forcing each posited word to have a vowel (cf. Brent and Cartwright 1996), a constraint that is even contradicted by words like *hmm*, which do in fact occur in the corpus relatively frequently. In order to keep the comparisons fair, this modified evaluation method was also applied to the "canonical" and baseline conditions in Simulation 6.

6.3.3 Results

6.3.3.1 Simulation 6a: The Brent60 subset

Results are reported first for the smaller, more restrictive subset of the Brent corpus, consisting of the utterances for which approximate accuracy (as defined above)

condition not available to the "canonical" condition. A more realistic handling of silence and pause information remains an open topic for future investigation.

was at least 60%. These results are shown in Table 6.7. This corpus subset (hereafter referred to as the "Brent60" corpus) has a much shorter mean utterance length (measuring on the test corpus) than the Korman corpus or the Brent33 subset (2.3 words per utterance, as opposed to 3.2 for Brent33 and 3.1 for Korman), and a much higher incidence of single-word utterances (46% for Brent60, versus 28% for Brent33, 27% for the Simulation 5 subset, and 26% for Korman). It follows that the single-word baseline will have substantially better recall on the Brent60 subset than on other corpora.

Simulation 6a.i: The canonical transcription The 17-feature and 36-phone SRNs using the canonical transcription perform above the length-based baselines for all measures except lexicon recall ($p < 0.01$ for all comparisons). The 11-feature SRN outperforms the length-based random baseline for boundary and word recall, and for lexical precision, but not for boundary or word precision. (It also performs significantly worse than the the 17-feature SRN on word precision and recall, and worse than the localist SRN in word precision. The 17-feature and 36-phone SRNs differ only in boundary recall, $p = 0.0255$).

As in other corpora, all are (trivially) below boundary precision, and above boundary recall, for the utterance-as-single-word baseline. However, due to the unusual number of one-word utterances in the Brent60 subset, the utterance-as-single-word baseline also outperforms all the SRNs on word precision (and at least matches them on lexicon precision) as well. Therefore, the performance of the SRNs is not so clearly superior to the baselines as it was for the Korman corpus.

Simulation 6a.ii: The recognized transcription The SRNs trained and tested on the recognized transcription on the whole perform fairly similarly to those using

Network	Input	Boundary		Word		Lexicon	
ubm-phon SRNs with "canonical" input							
12.80.37	Canonical	0.528	0.766	0.236	0.343	0.208	0.149
18.77.37	Canonical	0.578	0.835	0.300	0.432	0.251	0.170
37.70.37	Canonical	0.589	0.772	0.291	0.381	0.250	0.176
ubm-phon SRNs with "recognized" input							
12.80.37	Recognized	0.548	0.801	0.233	0.340	0.213	0.190
18.77.37	Recognized	0.543	0.808	0.237	0.352	0.202	0.179
37.70.37	Recognized	0.577	0.773	0.258	0.345	0.218	0.196
Baselines							
Length-based	base	0.515	0.613	0.204	0.243	0.133	0.155
Utt-as-Word	base	1.000	0.427	0.462	0.197	0.250	0.108

Table 6.7: Percent precision and recall for the three nets trained and tested with a canonical, dictionary-based transcription and an automatically phone-classified transcription of the "Brent60" corpus subset, compared with two baselines

the canonical transcription—much more similarly than the two conditions in Simulation 5. Only the 17-feature SRN differs significantly in performance between the two, and then only for word precision and recall. Although we see the same trend as in Simulation 5 (the 17-feature distributed representation performing best in the "canonical" condition, but the localist performing best in the "recognized" condition), the differences are not significant.

The SRNs using the recognized corpus also outperform the length-based baseline in boundary and word recall and for lexicon precision, though only the localist, 36-phone SRN outperforms the baseline in boundary and word precision ($p < 0.01$ for all comparisons).

Network	Input	Boundary		Word		Lexicon	
ubm-phon SRNs with "canonical" input							
12.80.37	Canonical	0.465	0.845	0.163	0.295	0.196	0.202
18.77.37	Canonical	0.481	0.852	0.172	0.303	0.206	0.209
37.70.37	Canonical	0.531	0.861	0.233	0.377	0.226	0.261
ubm-phon SRNs with "recognized" input							
12.80.37	Recognized	0.458	0.753	0.131	0.215	0.129	0.260
18.77.37	Recognized	0.465	0.736	0.137	0.216	0.126	0.263
37.70.37	Recognized	0.482	0.720	0.148	0.222	0.131	0.276
Baselines							
Length-based	base	0.464	0.533	0.150	0.173	0.095	0.276
Utt-as-Word	base	1.000	0.314	0.288	0.091	0.122	0.129

Table 6.8: Percent precision and recall for the three nets trained and tested with a canonical, dictionary-based transcription and an automatically phone-classified transcription of the "Brent33" corpus subset, compared with two baselines

6.3.3.2 Simulation 6b: The Brent33 subset

Because the relatively small size and short average utterance length of the Brent60 subset made it difficult to distinguish the performance of the SRNs in the "Canonical" and "Recognized" conditions from the baseline, it is necessary to examine a larger subset of the Brent corpus to obtain reliable figures. It is also useful to see how the Christiansen model (with the various input representations examined here) fare with a greater degree of subsegmental variation than that provided in the Brent60 subset. The Brent33 subset, more than double the size of the Brent60 subset, makes this closer look possible. Results for this corpus subset are shown in Table 6.8.

Simulation 6b.i: The canonical transcription As in Simulation 4 with the Korman corpus, the localist representation performs better than either of the distributed input representations. The 36-phone input outperforms both the 11- and the 17-feature inputs on boundary precision, word precision and recall ($p < 0.001$ on all comparisons) and on lexical recall ($p < 0.05$ on both comparisons). (The 11- and 17-feature inputs do not differ significantly from one another.)

The 36-phone input outperforms the length-based baseline on all measures except lexical recall ($p < 0.001$ on all comparisons), whereas the 17-feature is no better than baseline on boundary precision, and the 11-feature fails to clear baseline on boundary or word precision. Both of these are significantly worse than baseline on lexical recall ($p < 0.05$). The performance of the SRNs compared to the single-word baseline is worse on word precision (as in Brent60), but better on lexical precision, as well as boundary, word, and lexical recall ($p < 0.01$ on all comparisons).

Simulation 6b.ii: The recognized transcription As in Simulations 2b and 3a, there are few differences in the performance of the different input representations on the recognized corpus. As with the others, the localist 36-phone input performs slightly better, but this is only significant for boundary and word precision, which are slightly better for the 36-phone representation than for the 11-feature representation. This small improvement in precision is offset by boundary recall, which is somewhat worse ($p < 0.05$ for all comparisons).

However, there is a significant drop in performance relative to the "canonical" condition. Boundary recall, word precision and recall, and lexicon precision are all significantly worse for the SRNs trained and tested on the recognized corpus compared to the corresponding SRNs on the canonical corpus ($p < 0.001$ for

all conditions). This drop is sufficient to bring to bring boundary and word precision down to the level of the length-based baseline (though boundary and word recall are still significantly better than baseline, $p < 0.001$, as is lexicon precision, $p < 0.01$ for all three comparisons). Strikingly, lexicon recall is better for the "recognized" condition than for the "canonical" condition, though it still is no better than the length-based baseline.

6.3.4 Discussion

Simulation 6a shows that the Christiansen model, even without the stress cue, is robust to data with subphonemic variation when this variation is carefully controlled. This finding is consistent with previous tests of the Christiansen model in CA97 (Christiansen and Allen 1997). Insofar as the near-continuous vector output of a phone recognition classifier is a more accurate representation of human perception than the featural byte-swapping done in CA97, this study makes the point more strongly.

However, this point must be tempered by the observation that the corpus subset used to make this point is contains sufficiently simple language that even simple baselines with no use or knowledge of segmental distributional cues are able to do fairly well, at least on precision metrics. Simulation 6b, performed on a larger subset of the Brent corpus, including utterances that are considerably more difficult both for the phone classifier and for the Christiansen word segmenter (even using the "canonical" corpus), shows that there is a point where the variation does cause significant degradation to the model.

One hypothesis entertained during the construction of these simulations was that a featural input representation may prove beneficial in the face of highly variable input. Since it is to be expected that much of the subsegmental variation

in the automatic phone classifier's output will be among phones with similar featural representations, it stands to reason that the featural representation will absorb much of the messy variation, and bring more sense to the data. The direct localist representation, on the other hand, would not have this advantage, and might be expected to be less robust in handling subsegmental variation.

This hypothesis was only partially borne out. A truer statement would be that any apparent advantage of a particular input representation (17-feature for Simulation 5 and Brent60, 36-phone for Brent33) largely disappears in the "recognized" condition. (Naturally, this is untestable for the Korman corpus.) That being said, the localist 36-phone tends to perform the best in the "recognized" condition, though the differences are often too small to be significant. In any event, there is no clear advantage for distributed, featural input.

6.4 General discussion and conclusions

In conclusion, the simulations described in this chapter show that, despite some degradation of performance, the "phon-ubm" variant of the Christiansen model does perform above baseline on certain measures of performance (most notably boundary and word recall) even on data with a high degree of subword and subsegmental variation. Since the the data here were derived from automatic transcription methods using the actual acoustic data as a starting point, rather than randomized alterations of human- or dictionary-transcriptions, the simulations here are arguably a more realistic test of this than that provided by CA97. Inasmuch as generalizing over continuous-valued input vectors is a more difficult task than generalizing over near-binary vectors (even "noisy" ones), the simulations also provide a more rigorous test of the model than CA97. Finally, it must be reemphasized that the model tested is not the most powerful variant used in CAS98

(the one also used in CA97), but one deprived of the lexical stress cue, suggested by e.g., Cutler (1994); Johnson and Jusczyk (2001) to be a very potent cue for English word segmentation.

Specifically, Simulation 4 demonstrated that the catalyst units do afford a significant gain in performance over the SRNs without them, when using canonically transcribed corpora. Other simulations (not described here) suggest that this is also true using a recognized transcription of Brent60. Simulation 5 showed that, while the sub-word variation added by an automatic phone recognizer caused a significant drop in performance, the phon-ubm model still performed better on word recall (though not on precision) than the baselines not using phonological information. Finally, Simulation 6 demonstrated that this basic pattern of performance is maintained when sub-segmental variation is added to the input and target activations. Simulation 6a shows that, when the degradation of the input is relatively slight, the degradation of performance is also slight; when more variation is allowed, as in Simulation 6b, performance drops accordingly.

What this chapter does not attempt to answer is the role played by suprasegmental cues (e.g. cues associated with lexical stress) in the face of highly variable (and hence less reliable) segmental and subsegmental cues. It may well be that stress cues will "take up the slack" as it were, and improve performance even more than they do for the "canonical" condition, closing the gap between the models trained on recognized input and those traned on canonical input. It seems quite plausible that an *idealized*, binary (all-or-nothing) stress cue would have that effect.

However, just as the thrust of this dissertation was to investigate segmental cues deriveable from the acoustic signal more directly and automatically than those previously used, so any future work on this subject ought to be devoted to

finding a "smooth function reflecting the continuous nature of the acoustic parameters recognised as stress" (Christiansen et al., 1998:255). A programmatic sketch of how this could be done is offered in the following chapter (Section 7.4), along with preliminary but suggestive results for even a very simple model of segmental prominence derived solely from the phone-classification data prepared and used for Simulation 6.

CHAPTER 7

NEXT STEPS

7.1 Goals of the dissertation

The goals of this dissertation are twofold. The primary goal, examined in Chapters 3 and 6, is to examine the role played by representations of the input in models of word segmentation by pre-lexical infants, and to propose and test a representation of input that better approximates the actual input the infants in question are receiving (at least at the audio level). Some older works (e.g. Brent and Cartwright 1996) have spoken of applying their methods to audio data, but have had to make do with transcriptions. Other models (e.g. Roy and Pentland 2002) have developed multi-modal input corpora for training and testing, but have not provided comparisons to readily available corpora in CHILDES, nor contributed their corpora to CHILDES. Given the growing number of child-directed speech (CDS) corpora with available audio data in CHILDES and the increased sophistication and availability of automatic speech recognition (ASR) technology, it is time to "raise the bar" as it were on the level of realism expected from a model's input.

A secondary goal is to continue to expand the number of languages under which computational models of the WST are tested. This number is still relatively small. The author is aware of no other work on modern Greek WST specifically. Since modern Greek has greater levels of morphology and significantly different

phonotactic patterns from English, as well as a less predictable distributional pattern of lexical stress (not examined here), further study of this language is expected to yield interesting insights into which cues may be useful across languages and to what degree their relative utility follows certain patterns.

7.1.1 Preserving natural subsegmental variation

The primary goal of this dissertation has been to propose, test, and evaluate a more realistic method for preserving the subsegmental variation in the audio signal than those used in previous models, which either simply attempted to recreate or approximate that variation through rules or or rule-constrained random behavior (CSCL97 and CA97), or ignored it altogether.

This is important because, while it seems quite likely that infants are beginning to categorize speech sounds into phonemes during their first year, it is far from obvious—even doubtful—that they have completed this task so thoroughly as to make the subsegmental variation unimportant enough to gloss over. In fact, if it is the case (as e.g. Beckman 2003 suggests) that phonemic classification is not complete until some words are learned, then the assumptions made by a model that assumes symbolic, phonemic representation—or a representation deterministically derived from a phoneme set—come near to being circular in that the input these models use for learning to segment words does not become available until after the WST has already been solved. While showing that a heuristic is potentially circular does not automatically invalidate it, it *does* place a burden of proof on the proposer that the heuristic can be bootstrapped incrementally.

The nature of this proof may differ from model to model. In the case of INCDROP, the proof would be that the different instances of a given symbol-sequence (or string) segmented off as a (potential) word can be consistently recognized as

being equivalent (instances of "the same thing"). For Hockema's (2006) CWBP, it would be that some other observable approximation for the CWBP (be it TP or MI or some utterance-boundary based approximation) is able to converge to the CWBP as segmentation progresses and a vocabulary of potential word candidates is amassed. For Swingley's (2005) model, the burden is somewhat lighter: he only needs to show (besides solving syllabification issues for the language in question) that syllables are reliably distinguished one from another, something that experimental evidence suggests is easier for young infants than distinguishing segments.

However, the proposed model avoids this potential circularity in a different way: it uses a phonemic inventory only as a set of reference points or dimensions for a probability space within which each segment can be placed as it is heard. While the boundaries of each segment are still assumed for convenience of evaluation and comparison, neither the identity of any given segment, nor the phonemic equivalence of any two segments or strings, is assumed to be known. This must be discovered from the information available in the acoustic signal, and when this is difficult, the model must be robust to these difficulties.

It is possible in principle that some of the subsegmental variation preserved by the APC system used here is in fact helpful for the WST. It is beyond the scope of this research to distinguish the properties or patterns of subsegmental variation which help the model from those which add to its challenge; for our purposes they come as a package. However, once the contribution of distributional segmental cues *as available to the infant listener* is more accurately estimated by taking subsegmental variation into account, a truer picture of the relative contributions of other cues, such as the suprasegmental cues of that language, should be available. Insight into that interaction between cues (be it tradeoff, cooperation, or both) is what the Christiansen-style connectionist models promise. A major point of the

proposed model is to make this promise believable by making the input assumptions on which it rests more plausible.

7.1.2 Exploring variant statistical cues

A secondary goal of the dissertation is to examine the basic statistical cues underlying many of the proposed heuristics and computational models in order to see whether the relative effectiveness of these statistics remains constant cross-linguistically, or whether some of these findings are specific to English. This investigation emphasizes the statistical cues and not on the learning mechanism. The application of the cues was kept as simple as possible (straight table lookup from a training corpus combined with simple threshold comparisons) in order to focus on the statistical cues themselves. If certain types of cues are found to be particularly important in a wide variety of languages, then one may begin to investigate whether there is some universal or fundamentally necessary place for these cues in language acquisition.

That children *can* use statistical distributional cues such as TP to help them acquire any language is not in question. Indeed, the ability to spot certain statistical patterns appears to be a general learning ability that infants have, applicable not only to language (natural or artificial), but to other learning tasks (music, visual stimuli, etc.). Other species show evidence of statistical learning abilities as well (Hauser et al. 2001).

The question probed here is one of details, one that Aslin et al. (1998) expressly leave open. Granted that statistical cues are important, what is the best way of formulating them? Which variants of the statistical cues, or which heuristics for exploiting them, are most effective? Do they change from language to language, or are they independent of the statistical properties of the language's segmental

distribution or phonotactic system? While some of these questions may seem to be small, they add important details to the larger picture of language acquisition. In order to compare the relative importance of different cues (e.g. segmental vs. suprasegmental), one must not only know how salient and robust to variation or "noise" each cue is (which is the primary goal of this dissertation as discussed above) but also how best to use each cue. If, for example, the information from segmental distributional cues is represented in a model at the level of segments only, and it turns out that some other representation is more appropriate for that language (and also available to the infant learners), then the model will end up under-representing the importance of segmental cues.

While Chapter 4 restricts itself to cues involving a "window" of up to two segments, and even within this window does not purport to conduct an exhaustive survey of all possible distributional statistics, it nevertheless distinguishes between cues that make use of pauses or utterance boundaries and those that rely solely on variant measures of segment cooccurrence. As will be seen in the next section, the relative importance of these cues can vary.

7.2 Findings of the dissertation

7.2.1 Distributional cues in English

As discussed above, a number of models of word segmentation in infants (particularly those claiming some degree of universal, psycholinguistic plausibility) have incorporated heuristics drawn from one or both of the following two assumptions: first, that areas of low segmental predictability (that is, where it is more difficult to predict what the next phone will be) correlate with word boundaries; and second, that utterance boundaries often resemble other word boundaries in certain

ways (e.g. the types of phones that precede them). However, various researchers use a variety of methods to operationalize both segmental predictability and the predictability of an upcoming utterance boundary:

1. Operationalizations of low segmental predictability:
    a) Harris (1955): High possible successor count.
    b) Saffran, Aslin, and Newport (1996a): Low transitional probability (calculated on adjacent syllables).
    c) Elman (1990); Cairns et al. (1997), and Christiansen et al. (1998): Various measures of error in an SRN's output nodes.
    d) Brent (1999): Two baselines, each measured on the segment (not syllable), relative to surrounding segment pairs:.
        i. Low transitional probability
        ii. Low mutual information
2. Methods for using utterance boundary information:
    a) Aslin et al. (1996); Allen and Christiansen (1996), and Christiansen et al. (1998): higher-than-average activation of a trained utterance-boundary output node on an ANN.
    b) Brent and Cartwright (1996, Experiment 4): Utterance boundary information used implicitly (as extant utterance-initial and utterance-final consonant clusters are used to approximate possible word-initial and word-final clusters).
    c) Brent (1999): Utterance boundaries (and one-word utterances) provide the "boot" for a vocabulary-building bootstrap, assisted by MDL.

A full and detailed comparison of these various formulations for equivalent corpora and conditions in English has not been conducted. However, some comparisons can be drawn from the literature. The performance statistics reported for Brent's (1999) baselines show that MI performs better than TP at the segmental level for English. This result is consistent with Hockema (2006), which shows that MI correlates more strongly with the "upper bound" statistic CWBP than TP does.

Less attention has been paid to the utterance-boundary heuristic as such—only Aslin et al. (1996) (AWLB96) has studied it directly, not in combination with other cues. Their results are difficult to compare directly to Brent's (1999) baselines given different corpora and different metrics, but it seems highly likely that this cue's performance on its own is substantially below either TP or MI for English (although still better than chance). However, in combination with other cues it has been highly successful (cf. Brent and Cartwright 1996; Christiansen et al. 1998; Brent 1999).

7.2.2 Segmental distributional cues in Modern Greek

Chapter 4 examines these cues for modern Greek, so that their performance can be compared to one another, and also to their performance in English. Simulation 1 (Section 4.5) indicates that utterance endings are, even alone, quite helpful cues for word segmentation in Greek. Given a single phone of context, the conditional utterance boundary probability (Utt.BP) is statistically equivalent to its upper bound ($\Pr_{wp}(\#|x\_\_\_)$ given only the left-side context), as indicated in Table 4.2. Comparisons with AWLB96's results in Section 4.5.2.2 suggest that (glossing over corpus differences), one phone of context in Greek is as useful as two in English. Adding a second phone of context improves results significantly on all measures except

lexical recall. Indeed, two phones of context are as useful as three in English (by AWLB96's study).

Simulation 2 (Section 4.6) shows patterns similar to Brent's and Hockema's findings—MI still outperforms TP—although the use of optimized thresholding lessens the difference between them somewhat. However, neither MI nor TP does as well as in Brent's (1999) English corpus; nor does their equivalent upper bound (CWBP) fare as well in Greek as it does for English (cf. Table 4.3 with Hockema 2006). The differences in the effectiveness of these cues between English and modern Greek relate directly to the differences in phonotactics between the two languages. A few examples here will suffice. In English, the phonotactics of syllable boundaries and word boundaries are very similar (cf. CSCL97's results as discussed in Section 3.1.2) in terms of the "dips" in segmental predictability that they cause—hardly surprising considering the high proportion of one-syllable words in English! In contrast, Modern Greek word boundaries are for the most part subject to strict phonotactic contraints—more so than word-internal syllable boundaries. The rich variety of consonant clusters in Greek also makes segmental predictability heuristics less effective. Finally, Hockema's (2006) bimodal distribution for CWBP fails to hold for Greek when weighted for frequency (see Section 4.8.1).

For all these reasons, predictability measures like MI and TP are less effective for Greek than for English at the segmental level.[1] Conversely, Greek's (phonotactically restrictive but distributionally pervasive) set of suffixal inflections makes cues focusing on the end of the word relatively more successful. Comparisons between Simulations 1 and 2a suggest that Utt.BP with two segments of context performs quite favorably compared to segmental probability. Two-segment Utt.BP

[1]Transitional probability or mutual information at the syllable level, as Hockema (2006) suggests for Spanish or Japanese, may well be more appropriate for modern Greek. This is not tested here.

performs about the same as unthresholded TP in boundary and word recall and better in lexicon recall and all precision metrics ($p < 0.05$). It also outperforms MI in boundary recall and lexical precision ($p < 0.001$).[2]

Taken together, these results suggest that the balance between segmental predictability and cues based on ends of utterances are different in Greek than in English. Simulation 3 shows that only when combined do they approach the same level of performance that TP alone has for English. Unless word segmentation is simply harder in Greek than in English, other cues must help compensate.

7.2.3 The effects of input representation

In Section 5.2, a model was proposed for preserving subsegmental information from the acoustic signal in audio CDS corpora. In Section 5.2.2, a way to adapt this model to generalize the cues explored in Chapter 4 was proposed, although these not tested directly on Greek, owing to the lack of a suitable Greek coprus. Chapter 6 tests the new model of input representation on an English corpus using the Christiansen SRN (CAS98) learning model as a basis for departure. The new input representation could have in principle been adapted to and tested on some other model for word segmentation, but ANNs' allowance for arbitrary, real-valued featural input simplified the adaptation, and the potential of the CAS98 model for examining the interaction between multiple cues made it attractive. However, in this work only two cues (utterance boundaries and segmental predictability) were tested and combined explicitly. Lexical stress was left for future work, as explained in Section 7.4.

[2]The optimal operating points on Simulations 2 and 3 are sufficiently different to make comparisons with Utt.BP difficult.

7.2.3.1 Effects of the feature set

Before subsegmental variation itself is discussed, a few words will be said about a another aspect of the input representation: the type of vector or feature set used to represent the input. This was examined as a separate variable in Simulations 4, 5, and 6. Three feature sets were examined: one used in CAS98, one used in CCC00 and CCC05 (described in detail in the latter), and a one-hot or localist feature set used as a control. It was hypothesized that, since the 17-feature CCC05 feature set more closely matched typical feature sets in mainstream theoretical phonology, it might constitute a better representation of the underlying properties found in English phonotactics, or simply be a better representation of how the different types of segments sound. Either way, it was expected to outperform the 11-feature CAS98 representation. Similarly, if the particular representations encoded in CCC05 are relevant to English phonotactics, then CCC05 might also be expected to outperform the baseline representation in both the canonical and audio-based conditions. If the featural representation helps "smooth out" the variation found in natural speech, then we might expect an additional variant in the automatically recognized audio-based conditions.

In general, the feature set used in CCC05 performed better than the one in CAS98, consistent with expectations. However, CCC05 did not consistently outperform the localist representation. With the *ubm-only* SRNs in Simulation 4a, it significantly outperformed the localist model, but the reverse is true when catalyst nodes are introduced in Simulation 4b. Simulation 5 yields similarly mixed results: The CCC05 feature set outperformed the localist input representation with the canonical transcription data in Simulation 5a, but not for the APR-recognized transcription in Simulation 5b.

This suggests that the phonological features used in CAS98 (and possibly those from CCC05) may not be the most appropriate ones for representing the English input for this task. However, this result must be interpreted cautiously, since there could be factors of ANN learning involved that make localist representations inherently easier to learn. Moreover, since there are many phonological feature sets from which to choose, a number of other systems would have to be tested before any generalization over all feature sets could be made.

7.2.3.2 Effects of subsegmental variation

Simulation 6 explicitly examined the effects of subsegmental variation in the input. This variation significantly degraded the performance of the model, suggesting that any benefit of subsegmental cues to this model of word segmentation are overshadowed by the increased difficulty in learning generalizations in the face of such variation. This is not to say that infants, at least in later stages of development, do not benefit from patterns they find in subsegmental variation. On the contrary, Jusczyk et al. (1999a) suggests that by 10.5 months, they do (see Section 2.3.4.1). That this model with this input is not sufficiently sensitive to do so may be a weakness in the connectionist model itself, or simply reflect the fact that the APC-derived input is "noisier" than what infants at that age have available to them. That will be left an open question; it is of course known that sighted infants have more cues at their disposal, including the facial "visemes" of their caretakers' lips, jaw, and tongue, and that all infants naturally babble and are beginning to learn sound-motor mappings that may assist them (cf. e.g. Plaut and Kello 1999). Neither the original Christiansen model nor this proposed variant attempts to include these features.

However, given that the APC-derived input is almost certainly an underestimate of what infants have available to them, the fact that the model still performs better than the baselines with this input demonstrates the robustness of the model to the kind of variation found in natural acoustic input. This is consistent with the findings in CA97, but considerably more impressive, insofar that the variation found here was not only more realistic, but also more challenging. The implications of this finding will be explored in the next section.

Finally, it is worth noting that the differences between the three feature sets observed in the canonical-transcription input conditions are greatly diminished when subsegmental variation was introduced. This suggests either that each of these are equally well-suited (or unsuited) for representing variable input, or that it is not the feature sets that make the model robust to subsegmental variation, but properties of the ANN itself.

7.3 Implications of the findings

7.3.1 Implications for studies of language evolution

While it is certainly possible that a language may change in ways that make segmentation (and other parts of language acquisition) easier, it seems unwise at this stage to claim that languages as a general rule evolve to maximize the effectiveness of particular statistical heuristics (e.g. reliance on the bimodality of Hockema's (2006) CWBP or on Harris's (1955) dips in segmental predictability). Some languages may; other languages may change in ways that enhance other cues instead, such as the reliability of extrapolation from utterance-final segmental distributions to word-final phonotactics. While this latter cue is a relatively weak guide to WST in English, in Greek (and very likely in other inflection-rich languages as well) it is

more powerful. Indeed, if Greek has evolved to make any particular WST heuristic more effective, word-final phonotactics make a better candidate than segmental predictability cues. While proving such a change is well beyond the scope of this thesis, it may be noted in passing that modern Greek has restricted the number of consonants and consonant clusters allowable word finally, discarding final /r/, /ks/, and /ps/ and making /n/ considerably rarer. Indeed, some dialects restrict word-final /n/ even more, by dropping it on the genitive plural ending or adding an epenthetic vowel to final /n/ on verb endings.

A more serious issue is that an investigation of language evolution that ignores the sound signal (or abstracts away the variation therein) is missing a crucial aspect of language. While many higher-level aspects of language can be profitably modeled by symbolic means, it is obvious that the first and primary medium through which language is learned by the next generation is speech. While it is appreciated that speech is messy, and that mishearings do contribute to language change (cf. examples in Welby's 2003 introduction), it takes a model such as the one proposed here to pinpoint more precisely *where* such mishearings are likely to occur. A model of language change that takes into account changes caused by mishearings and missegmentations could be improved by incorporating insights from the proposed model.

7.3.2 Implications for special population modeling

The model of input proposed in this dissertation is intended to approximate the hearing of normal, pre-lexical infants who have begun to develop a phonemic inventory of their native language, but do not yet have higher-level linguistic capabilities to assist in segment recognition or classification. While the proposed model for normal children is still only approximate, in principle it could be extended and

specialized to provide models for various special populations, including cochlear implant recipients and other hard-of-hearing children, as well as those with specific language impairment. Connectionist modeling of language disorders is not new ground (see e.g. Conway et al. in press, for a review); however, the evidence here argues that even for normal children, the consequences of assuming invariant representations of input (as is typically done when the source of the input is transcription data) are not negligible, and that even approximate attempts to model subsegmental variation provide a needed corrective. Obviously, modeling the auditory stimuli that special-population individuals receive is far from trivial. Nevertheless, these results suggest the potential cost in a model's accuracy for not trying.

A related but often overlooked concern is the language learning performance of normal children in non-ideal learning environments. Recent studies have begun to measure the potential risks of background noise in language acquisition: for example, Newman and Jusczyk (1996); Newman (2005) examine the effect of multiple speakers in the background (often called "cocktail party" noise). Newman (2005) found that, although children less than one year old do have a limited ability to separate out different talkers in the auditory scene, they are nevertheless unable to segment and recognize their own names when the signal-to-noise ratio is less than 10 dB. Since speech segmentation is crucial to later language development (Newman et al. 2006), the ability to identify and correct kinds and levels of background noise that might delay speech segmentation and general language development could be very useful in designing and maintaining appropriate educational environments for young children—especially (but not only) for populations at risk for language impairment. Since ASR in noise is already a well-investigated

area of research (see Fosler-Lussier et al. 2005 for just one example), an extension of the model to this domain is comparatively straightforward.

7.3.3 Implications for automatic speech recognition

This research illustrates a potential use of ASR technology as a tool for speech science, separate from its already-established utility as a practical tool in industry. While it is beyond the scope of the model to estimate the accuracy of the APC's approximation to infant human speech abilities, the results from Simulation 6 suggest that the model is essentially successful in using the APC-derived inputs. A closely related example using ASR as a tool in modeling human speech processing is Scharenborg et al.'s (2005) SpeM project, which embeds ASR into a new implementation of Norris's (1994) SHORTLIST. Comparisons between this research and SpeM are not possible here; however, its very existence indicates the growing interest in ASR as a method of modeling human speech processing.

A more speculative potential application of the research in Chapter 4 specifically is the problem of unit selection in ASR. Mainstream ASR has traditionally relied on segment-based models of recognition, which then get concatenated into words via pronunciation dictionaries. A number of attempts have been made to use other types of units (syllables: e.g. Sethy et al. 2003; Hämäläinen et al. 2005; articulatory gestures: Deng and Sun 1994; Sun and Deng 2002; automatically learned units: e.g. Bacchiani 1999). One difficulty of using syllables is that for many languages, including English, syllabification is relatively hard, and mistakes in syllabification will lead to mistakes in word boundaries, sending cascading errors through the ASR system's hypothesis. Even more importantly, the number of possible syllables in English is enormous, quickly leading to intractable data-sparsity issues.

Chapter 4 considers a number of statistics which may be helpful for finding word boundaries. It was seen in Section 4.8.1 that, while many of the cues were less successful in Greek than in English at demonstrating where the boundaries likely to be, a very simple cue could pinpoint quite clearly where the word boundaries were *unlikely*—even disallowed for "core" native vocabulary (excluding loan-words, neologisms, and onomatopoeia). Such a cue may be beneficial to ASR in suggesting which segment-sequences may be profitably combined into single units with minimal (and measurable) risk of their straddling a word boundary. For example, in Greek, one cannot use syllables directly as units because /n/ and /s/ may be either word-initial or word-final. However, if one separates off /n/ and /s/ as designated special units,[3] one is then free to treat any string ending in a vowel (C^*V) as a single unit with little fear (once the limited number of exceptional words are controlled for) of missing a word boundary. These same statistics can be used for any language to identify clusters unlikely to cross word boundaries in that language. Whether or not the benefit for the acoustic model of such combinations outweighs the cost in data sparsity is a separate issue, which can be measured by cooccurrence statistics and hands-on comparison with such models over more traditional triphones.[4]

7.3.4 Implications for comparing cue strength

An interesting, but difficult, question in studying speech segmentation is knowing which cues contribute most to segmentation in a particular language, whether due

[3]We could term these units "mora-like" by way of a very loose analogy with Japanese /n/, but being clear that no theoretical implications are meant by such an analogy.

[4]Similar ideas have been proposed for English by Keith Johnson, examining very common segment clusters as potential units—though without reference to the likelihood of a word boundary within the clusters.

to high reliability, wide provenance, or high salience of the cue in the speech signal. In practice, it is highly unlikely that this question can be fully answered for a language such as English (or perhaps any language), where subsegmental, segmental, and suprasegmental phenomena are often intertwined—for example, the strong correlation between lexical stress and subsegmental vowel quality, or the segmental inventory of vowels allowed. Nevertheless, a model that can account even approximately for the relative strength of various cues when they do diverge or clash would be highly desirable. To be able to predict or explain changes in relative cue strength during different stages of development or in different environments (e.g. noise levels) is even more impressive.

This is difficult, and few studies are so ambitious as to attempt it. Most content themselves with being able to show the potential applicability of a single cue or heuristic. That being said, there is a curious mismatch between experimental studies like Johnson and Jusczyk (2001) and Thiessen and Saffran (2003) which show that stress is the major cue in English word segmentation by 9 months, and the preponderance of computational models which focus primarily if not exclusively on segmental distributional cues. While a few studies (e.g. Cairns et al. 1997; Swingley 2005) interpret their results as modeling the early stages of word segmentation—finding the bootstrap that allows infants to discover the trochaic bias in English—several models that explicitly incorporate stress into their representation, such as Curtin et al. (2005) and Hockema (2006), seem to treat lexical stress as little more than an extra feature added on to help distinguish certain segments or syllables.

Both of these latter studies find that when unstressed vowels or syllables are distinguished from segmentally identical stressed ones, distributional statistics such as TP work better. This is undoubtedly true, but the performance of these

statistics without lexical stress cues is already so good that the improvement from adding stress, though statistically significant, is rather small. The findings in Simulations 5 and 6 in Chapter 6 strongly suggest that this is because their results without lexical stress are too optimistic. The assumption of idealized, error-free input at the segmental level—for any model—overestimates the potential performance of segment-based statistical heuristics, as used by infants still learning the segmental inventory on which such statistics are based.

In this regard, Swingley's (2005) model is more cautious, and more realistic, but even it may be assuming too much of the infants' discrimination powers. A modification of a syllable-based model allowing for confusion of syllabic identity (especially for the unstressed syllables) would be more plausible still. Briscoe (1989) suggests a model of the input in which segments in stressed syllables may be distinguished, but segments in unstressed syllables are represented only as broad classes. Fosler-Lussier and Rytting (2005) and Fosler-Lussier et al. (2005) show that such a model is sensible for "adult" ASR, due to the great variation of pronunciation found in unstressed syllables. Still, unless such a model takes into account actual audio data where available, it risks making unwarranted assumptions of what is confusable in speech and what is not. Better still is a model of input based on the speech itself. To this end, the APC model used here could in principle be adapted to a syllable-based framework, by performing automatic classification on syllable-sized units, assuming some solution to the syllabification problem is found.

In order to obtain a more accurate sense of how segmental and suprasegmental cues compare in a particular language, it is necessary to develop accurate estimates of the reliability of each cue in the context of the variable stimulus the infant encounters (or, failing this, models that over- or under-estimate each cue's

strength to similar degrees). Naturally, to do this—and to know when one has succeeded—is far from trivial. Perhaps the best one can do is constrain the range of a particular cue's effectiveness through plausible upper and lower bounds. However, the first step to a more accurate model of suprasegmental cues is logically parallel to that for segmental cues: just as one can use automatic phone classification for the segments, one can conceive of automatic stress classifiers for the syllables. This is in keeping with CAS98's promissory note for a continuous measure for stress quoted in Section 6.4. The next section will sketch out some possible approaches for such a tool.

7.4 Towards an acoustic-based model for lexical stress

The choice of type of stress-detection mechanism to use in a model of infant speech processing depends on how much knowledge of suprasegmental cues is assumed. On the one hand, the research reviewed in Section 2.2.2 shows that babies can distinguish almost from birth the types of contrasts that indicate stress in English (Jusczyk and Thompson 1978) and other languages (e.g. Sansavini et al. 1997, for Italian). Recognizing the contrast is not the same as mapping particular suprasegmental phenomena to the prosodic categories of "stressed" and "unstressed" syllables that adult speakers of English (or Italian, or Greek) recognize. However, the role that stress (or more particularly sentential accent) plays in early language acquisition may not be primarily categorical, but simply serve to focus an infant's attention on particularly distinct or relevant parts of the signal. If so, then the mapping to prosodic categories may not be as important as the acoustic cues themselves. However, before we revisit that thought in Section 7.4.3, let us consider how stress could be learned as a category.

7.4.1 Supervised approaches[5]

If we are willing to assume that infants have learned by the target age of our model to associate certain acoustic cues with certain prosodic categories in their native language, then we can instantiate that assumption by using the results of a supervised learning method to model that acquired knowledge. This could be argued to be roughly parallel with the proposed model's use of acoustic models for the phones, which were taken from a pre-trained ASR system rather than using unsupervised clustering.

An implementation of a simple stress detector for English using a feed-forward MLP on the audio signal is described in Fosler-Lussier and Rytting (2005). The TIMIT corpus was first syllabified using the NIST Tsylb2 syllabifier (Fisher 1996), then re-annotated frame by frame with lexical stress markings. All frames of each stressed syllable (including onset and coda consonants, not just the nucleus) were marked as stressed. A multi-layer perceptron with 100 hidden units was trained to predict Pr[Stress|Acoustics] with a nine-frame context window. No additional phonetic information besides the binary label stressed/unstressed was used in training. Frame-by-frame results on the TIMIT test set were 75% accurate (chance: 52%), and when MLP output was greater than 0.9, a precision of 89% was obtained (recall: 20%). While far from perfect, this result strongly suggests that even very simple methods can predict lexical stress fairly reasonably. This demonstrates that English stress is at least partially learnable from the audio signal alone, if labeled training data is provided.

Dubě da and Votrubec (2006) describe a similar, though somewhat more elaborate, stress detector for Czech. Since Czech content words are always stressed

[5]Some of the material in this section is taken from Fosler-Lussier and Rytting (2005).

on the first syllable, this lexical stress detector doubles as a word segmenter for that language. The Czech stress detector used a 15-30-1 MLP using f_0, duration, and intensity, each normalized for the particular segment(s) in the syllable, to predict stress over a three-syllable window (five parameters for each syllable: one each for duration and intensity; three for the f_0 contour). The best-performing net, which obtained 80% accuracy compared with a human rater of accent realization, used a balanced context window (i.e. one syllable context on each side of the target) and all three parameters; however, f_0 was necessary and sufficient to obtain better-than-chance performance. Since stress is not contrastive in Czech, and vowel-length is, making duration cues less useful for detecting stress, these results are fairly impressive. One would predict that for a language like Greek where stress is contrastive, similar methods would be even more successful.[6]

7.4.2 Unsupervised approaches

However, in keeping with a general preference for assuming as little supervised or pre-acquired knowledge as possible, it is arguably preferable to use an unsupervised model of stress, particularly if the current model for segmental cues is replaced by a segmental inventory derived through unsupervised clustering (see e.g. Lin 2005, for an example of such a model). One way to model suprasegmental cues without supervised learning is simply to feed known acoustic correlates of lexical stress (e.g. f_0, duration, intensity) directly to the ANN's input. The assumption here is that infants are able to hear differences in these measures, and somehow know they are relevant to word segmentation, but have not necessarily learned to cluster them into any particular prosodic categories (as the supervised method assumes).

[6]For a full description of the acoustic correlates of stress in Greek, see Arvaniti (2000).

In addition to f_0 and duration, another relevant acoustic correlate in English is *spectral tilt* (also called *spectral balance* or *spectral emphasis*). Sluijter et al. (1997) define it as the relative balance of intensity between the lower and higher parts of the spectrum, using 500 Hz as a cut-off point. Increased intensity in the higher frequency bands correlates with impressions of greater loudness or "force" in speech (Glave and Rietveld 1975, qtd. in Sluijter et al. 1997). Sluijter et al. find that spectral tilt correlates better than across-the-spectrum measures of intensity (which are relatively poor cues for lexical stress) with lexical stress in Dutch.[7] Campbell and Beckman (1997) find that, in English as well, spectral tilt is affected by sentential accent and, to a lesser extent, by lexical stress on unaccented words. Thiessen and Saffran (2004) find that infants also can use spectral tilt as a cue to lexical stress in English.

The use of spectral tilt in ASR is not entirely new. Wu et al. (1997) describe the use of intensity in particular frequency bands, a measure very similar to spectral tilt, to find syllable nuclei. Since Wu et al. were specifically interested in the onsets of syllable nuclei, they used a temporal filter equivalent to finding the first derivative of the frequency-banded intensity, in order to find regions of increasing intensity in the relevant bands.[8] While they used this to detect onsets generally, given the empirical findings above, it should in principle (perhaps with some modification) work to find onsets of stressed nuclei specifically.

The exact operationalization and combination of these acoustic cues to lexical stress is left as an open question, and deferred to future work. However, the next section will briefly sketch an alternative approach to detecting acoustically

[7]Sluijter and colleagues reserve f_0 as a cue not for lexical stress directly, but for sentence accent; however, since sentence accent rarely falls on a non-stressed syllable, it still serves as a cue for lexical stress.

[8]In order to filter out the regions of decreasing intensity, they map negative values to zero, a technique known in engineering terms as *half-rectification*.

prominent stretches of speech, which requires no additional information from the speech signal than what is already provided from the automatic phone classifier already employed in generating our model's input.

7.4.3 Segmental confidence

Lexical stress has been identified as a highly useful cue for finding the starts of words in English because such a large proportion of English content words (especially nouns) begin with stressed syllables. However, this is a fact about English that infants must learn through experience. Moreover, there is another aspect of stress that helps infants with learning language: stressed syllables are pronounced more distinctly than unstressed syllables, and hence the phonemic content of those syllables is arguably easier to extract.

Rather than detecting stress directly as one of a number of cues for finding word boundaries, an alternate strategy is to think of the word segmentation task as simply finding short stretches or chunks of speech that are worth paying attention to. Finding the start of these salient stretches may facilitate word learning, especially if other cues indicate they are likely to match up with "adult" word boundaries.

Section 5.1.3.2, briefly exploring this line of reasoning, alluded to a measure of "segmental confidence" or distinctiveness, representing the degree of confidence that the APC system has in the most probable label for a particular segment. This could be operationalized in a number of ways. One straightforward way is simply to take the maximum activation (probability) value output by the phone classifier for that segment, regardless of the identity associated with that

value.[9] A brief, preliminary experiment over the Brent60 corpus using a close variant of this measure yields significant improvements on boundary and word precision, although boundary recall suffers. In fact, its performance is roughly equivalent to (and for the CAS98 feature system, better than) that of the corresponding "canonical" conditions. This suggests that even a very crude approximation of the prominence associated with lexical stress, using no additional resources, can nevertheless ameliorate much of the degradation seen in Simulation 6's "recognized" contexts. However, since these results are preliminary, no further speculation on their implications will be given here, and comparisons with this cue and more mainstream correlates of stress will be left to future work.

7.5 Future work

7.5.1 Cross-linguistic extensions

In addition to expanding the model to incorporate stress as outlined in the previous section, there are a number of directions in which this work could be extended. An obvious candidate is further cross-linguistic exploration. Without audio recordings for the Stephany corpus, little more can be done with it that is relevant to the primary goals of this research program. However, there are now audio or audio-visual corpora in the CHILDES/TalkBank collection in a number of languages, including Romance (French, Italian, Portuguese, and Spanish), Germanic (German and Danish), and Slavic (Polish), as well as Japanese and Thai. For

[9] The value for segmental confidence at time t is given by Equation 7.1:

(7.1) SegConf for time t $(SC_t) = \max_{q \in A} \Pr[Q_t = q | X_t]$

this research specifically, an ideal starting point on these would be the Italian, Portuguese, and Spanish corpora, which are expected to be roughly similar to Greek in terms of their interactions between suprasegmental and segmental cues.

There is also another multilingual audio corpus of child-directed speech (the Paidologos corpus) reported to be in process of integration into the CHILDES database. This corpus includes recordings of speech in Cantonese, English, Greek, and Japanese.[10] The Greek portion contains 10 hours (1 hour each from 9 mothers and one grandmother) of child-directed speech from native Greek speakers living in Thessaloniki and Larissa. In addition, the same caretakers were recorded in adult-directed conversation. For both the CDS and ADS, labeled transcriptions at the utterance and word levels have been made, and segmental sandhi and other non-canonical pronunciations marked (Asimina Syrika, p.c.).[11]

7.5.2 Improved analysis and evaluation

In the results given here, six measures of performance are used: precision and recall for boundaries, word tokens and word types. While these are sufficient for a comparison with many previous models (e.g. Brent 1999, AWLB96, and CAS98), they are far from telling the whole story about what is really being acquired by these models. One of the most obvious things to investigate next is not just *how many*, but *which* words are being learned. CSCL97 breaks these down into *content* or "open-class" words versus *function* or "closed-class" words. Content words

[10] Inquiries concerning further details and availability should be made to Dr. Jan Edwards at the University of Wisconsin-Madison.

[11] It is worth noting, however, that some of the transcription choices at the word level reflect the particular research interests of the original corpus creators, and may need some adjustment for use in other projects.

carry more concrete meaning to them (e.g. nouns, verbs, adjectives) and are usually learned by infants relatively early (especially nouns). In English they are also more likely to have strong initial syllables (again, especially nouns). Function words (determiners, conjunctions, prepositions, and other "connecting" words) are usually learned later, and are typically unstressed in speech. Cues such as the "trochaic bias" in English (cf. Section 2.3.2) and the "segmental confidence" heuristic mentioned above (Sections 5.1.3.2 and 7.4.3) are both focused on learning content words rather than function words. Swingley (2005) also uses this distinction, but goes further to break down his results into major parts of speech (nouns, verbs, etc.) and to discuss what types of two-word collocations are wrongly learned—and the logic behind some of these errors.

Another possible distinction that could be made is that between "fatal" and "non-fatal" errors. Even with morphologically plain languages like English, there are still several parts of a word that may be broken off (or wrongly added) without affecting its meaning greatly. For example, if the final *-'s* in *cat's* is wrongly segmented (as in Example 1 in Section 4.4.3 above), the possessive relation is lost, but the reference to the four-legged creature (more likely to be relevant to the infant) is preserved. Similarly, if the final *-ie* (/i/) is wrongly segmented off of the word *birdie* or *froggie*, little meaning is lost. These would be examples of "non-fatal" errors. However, if the final /i/ on *berry* is lost, the meaning is radically changed, and could be regarded a "fatal" error as far as acquiring the meaning of that word for that instance. While this exact formulation of error is not used in other models' evaluation schemes, similar distinctions are drawn in de Marcken (1994) regarding semantic equivalences and Batchelder (1997) distinguishing types of morphological errors.

Finally, Brent and Siskind (2001) compare the list of words that occur in isolation (one-word utterances) in their corpus with the results of a passive vocabulary inventory (Fenson et al. 1993) and find a correlation. Given the use of that corpus, a natural next step in evaluation is to compare the correlations of the various conditions in Simulations 5 and 6 with those words actually learned by the children, according to Brent and Siskind's (2001) results.

In all these more sophisticated measures, another aspect of model evaluation that can receive further attention is the matter of thresholds. Since the thresholds used by AWLB96 and CAS98 are essentially arbitrary and often non-optimal, they do not always give the best indication of a model's performance. The optimization used in Chapter 6 also may not be the most accurate reflection of what infants will do. Moreover, when two variant models or conditions differ from each other in both recall and precision, adjustments of the threshold may be the only way to determine which performance is superior. However, since a more appropriate method of threshold optimization is not obvious from the experimental literature, questions of thresholding and optimization were left as an open issue in Chapter 6.

Earlier studies (Rytting 2006a,b) address this issue for boundary precision and recall by using receiver operating characteristic (ROC) curves to track the model's performance over each possible threshold with distinct precision and recall (or equivalent measures). However, since performance on word and lexicon precision and recall is not guaranteed to have monotonic tradeoffs (the one increasing and the other decreasing as the threshold moves, as boundary precision and recall do), ROC curves are not an appropriate measure for them, and the standard assumptions and measures for statistically comparing two curves do not apply. In

addition, the treatment in Simulation 6 of spurious segmentations immediately after utterance boundaries and stop closures, while useful for eliminating spurious "words" consisting only of silence, also makes creating boundary ROC curves less straightforward. Resolving these issues is also left for future work.

7.6 Conclusion

Many models of word segmentation by infants use transcriptional data instead of speech as input. This leads to a spurious assumption (sometimes implicit, sometimes acknowledged as a convenient approximation) that all instances of a given word are pronounced alike. This dissertation addresses this spurious assumption of stimulus invariance below the level of the word, and particularly below the level of the segment, that arises from using transcriptional input. Thoughtful researchers have acknowledged the problems arising from this assumption, and some attempts (e.g. Christiansen and Allen 1997) have been made to compensate for it. However, until the results of a representative model using idealized input was compared with results on actual speech, the extent of the overestimation remained unknown.

The simulations in this dissertation have shown that the consequences of assuming stimulus invariance are far greater than the results of Christiansen and Allen 1997 suggest, and proposes a closer approximation to the stimuli that children actually experience in the course of language acquisition. Despite the limitations of earlier studies' assumptions, however, the results here show that the Christiansen model is still robust enough to degrade gradually to increased amounts of sub-word and subsegmental variation, and still remain above chance in its segmentation performance.

BIBLIOGRAPHY

ALLEN, JOSEPH AND MORTEN H. CHRISTIANSEN. 1996. Integrating multiple cues in word segmentation: A connectionist model using hints. *Proceedings of the Eighteenth Annual Cognitive Science Society Conference*, 370–375. Mahwah, NJ: Lawrence Erlbaum Associates. URL `citeseer.ist.psu.edu/allen96integrating.html`.

ANDERSON, JENNIFER L., JAMES MORGAN, AND KATHERINE S. WHITE. 2003. A statistical basis for speech sound discrimination. *Language and Speech*, 46(2-3).155–182.

ARVANITI, AMALIA. 2000. The acoustics of stress in modern Greek. *Journal of Greek Linguistics*, 1.9–39.

ASLIN, RICHARD N., JENNY R. SAFFRAN, AND ELISSA L. NEWPORT. 1998. Computation of conditional probability statistics by human infants. *Psychological Science*, 9.321–324.

ASLIN, RICHARD N., JULIDE Z. WOODWARD, NICHOLAS P. LAMENDOLA, AND THOMAS G. BEVER. 1996. Models of word segmentation in fluent maternal speech to infants. Morgan and Demuth (1996), 117–134.

BACCHIANI, MICHIEL. 1999. *Speech recognition system design based on automatically derived units*. Ph.D. thesis, Boston University. URL `citeseer.ist.psu.edu/bacchiani99speech.html`.

BATCHELDER, ELANOR OLDS. 2002. Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, 83.167–206.

BATCHELDER, ELEANOR OLDS. 1997. *Computational evidence for the use of frequency information in discovery of the infant's first lexicon*. Ph.D. thesis, The City University of New York.

BECKMAN, MARY E. 2003. Input representations (inside the mind and out). G. Garding and M. Tsujimura, editors, *WCFFL22 Proceedings*, 70–94. Somerville, MA: Cascadilla Press.

BENGIO, Y., P. SIMARD, AND P. FRASCONI. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2).157–166.

BORTFELD, HEATHER, JAMES L. MORGAN, ROBERTA MICHNICK GOLINKOFF, AND KAREN RATHBUN. 2005. Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, 16(4).298–304. URL `http://www.blackwell-synergy.com/doi/abs/10.1111/j.0956-7976.%2005.01531.x`.

BRENT, MICHAEL R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34.71–105.

BRENT, MICHAEL R. AND TIMOTHY A. CARTWRIGHT. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61.93–125. URL `citeseer.ist.psu.edu/brent96distributional.html`.

BRENT, MICHAEL R. AND JEFFREY M. SISKIND. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81.31–44.

BRENT, MICHAEL R. AND XIAOPENG TAO. 2001. Chinese text segmentation with MBDP-1: making the most of training corpora. *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 90–97. Morristown, NJ, USA: Association for Computational Linguistics.

BRISCOE, EDWARD J. 1989. Lexical access in connected speech recognition. *Proc. 27th Annual Meeting of the Association for Computational Linguistics*, 84–90.

CAIRNS, PAUL, RICHARD SHILLCOCK, NICK CHATER, AND JOE LEVY. 1997. Bootstrapping word boundaries: A bottom-up corpus based approach to speech segmentation. *Cognitive Psychology*, 33.111–153.

CAMPBELL, NICK AND MARY E. BECKMAN. 1997. Accent, stress, and spectral tilt. *The Journal of the Acoustical Society of America*, 101(5).3195–3195. URL `/cgi-bin/sciserv.pl?collection=journals&journal=00014966&issu%e=v101i0005&article=3195_asast`.

CARLSON, R., K. ELENIUS, B. GRANSTRÖM, AND H. HUNNICUTT. 1985. Phonetic and orthographic properties of the basic vocabulary of five european languages. *STL-QPSR 1/1985*, 63–94. Stockholm: Speech Transmission Laboratory, Dept. of Speech Communication, Royal Institute of Technology.

CARTERETTE, E. C. AND M. H. JONES. 1974. *Informal Speech: Alphabetic and Phonemic texts with statistical analyses and tables*. Berkeley, CA: University of California Press.

CHARNIAK, EUGENE, DON BLAHETA, NIYU GE, KEITH HALL, JOHN HALE, , AND MARK JOHNSON. 2000. Bllip 1987-89 wsj corpus release 1.

CHRISTIANSEN, MORTEN, JOSEPH ALLEN, AND MARK SEIDENBERG. 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13 (2/3).221–268.

CHRISTIANSEN, MORTEN H. AND JOSEPH ALLEN. 1997. Coping with variation in speech segmentation. A. Sorace, C. Heycock, and R. Shillcock, editors, *Proceedings of GALA*. URL `citeseer.ist.psu.edu/89572.html`.

CHRISTIANSEN, MORTEN H., CHRISTOPHER M. CONWAY, AND SUZANNE CURTIN. 2000. A connectionist single-mechanism account of rule-like behavior in infancy. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

CHRISTIANSEN, MORTEN H., CHRISTOPHER M. CONWAY, AND SUZANNE CURTIN. 2005. Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. James W. Minett and William S.-Y. Wang, editors, *Language acquisition, change and emergence: Essay in evolutionary linguistics*. Hong Kong: City University of Hong Kong Press. URL `http://www.isrl.uiuc.edu/~amag/langev/paper/christiansen05ACE%chapter.html`.

CHRISTIANSEN, MORTEN H. AND RICHARD A. C. DALE. 2001. Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition. *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*, 220–225. Mahwah, NJ: Lawrence Erlbaum.

CHRISTOPHE, ANNE, EMMANUEL DUPOUX, J. BERTONCINI, AND JACQUES MEHLER. 1994. Do infants perceive word boundaries? an empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society of America*, 95.1570–1580.

CHURCH, KENNETH W. 1987. *Phonological parsing in speech recognition*. Dordrecht; Boston, MA: Kluwer Academic Publishers.

CMU. 1993–2002. The Carnegie Mellon Pronouncing Dictionary. URL `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`, carnegie Mellon University.

COLLET, R. J. AND J.G. WOLFF. 1977. From phoneme to morpheme—revisited. *Bulletin of the Association for Literary and Linguistic Computing*, 5.23–25.

CONWAY, CHRISTOPHER M., MICHELLE R. ELLEFSON, RICHARD DALE, AND MORTEN H. CHRISTIANSEN. in press. Connectionist models of developmental disorders: A critical appraisal. D. L. Molfese and V. J. Molfese, editors, *Handbook of developmental neuropsychology*. Mahwah, NJ: Lawrence Erlbaum.

CURTIN, SUZANNE, T. H. MINTZ, AND D. BYRD. 2001. Coarticulatory cues enhance infants recognition of syllable sequences in speech. A. H. J. Do, L. Dominguez, and A. Johansen, editors, *Proceedings of the 25th Annual Boston University Conference on Language Development*, 190–201. Somerville, MA: Cascadilla.

CURTIN, SUZANNE, TOBEN H. MINTZ, AND MORTEN H. CHRISTIANSEN. 2005. Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, 96.233–262.

CUTLER, ANNE. 1994. Segmentation problems, rhythmic solutions. *Lingua*, 92.81–104.

CUTLER, ANNE AND S. BUTTERFIELD. 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31.218–236.

DAVIS, S. B. AND P. MERMELSTEIN. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4).357–366.

DE JONG, KENNETH. 1995. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, 91(1).491–504.

DE MARCKEN, CARL G. 1994. The acquisition of a lexicon from paired phoneme sequences and semantic representations. *International Colloquium on Grammatical Inference*, 66–77. Alicante, Spain.

DE MARCKEN, CARL G. 1996. *Unsupervised language acquisition*. Ph.D. thesis, MIT, Cambridge, MA. URL `http://xxx.lanl.gov/abs/cmp-lg/9611002`.

DEMUTH, KATHERINE. 1996. The prosodic structure of early words. Morgan and Demuth (1996), 171–184.

DENG, LI AND DON X. SUN. 1994. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *Journal of Acoustical Society of America*, 95.2702–2719.

DUBĚDA, TOMÁŠ AND JAN VOTRUBEC. 2006. Acoustic analysis of Czech stress: Intonation, duration and intensity revisited. *Proceedings of Interspeech'06*. Pittsburgh, PA.

ECHOLS, C., M. CROWHURST, AND J. CHILDERS. 1997. Perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language*, 36.202–225.

EIMAS, P. D. 1999. Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America*, 105.1901–1911.

EIMAS, P. D., E. R. SIQUELAND, PETER W. JUSCZYK, AND J. VIGORITO. 1971. Speech perception in early infancy. *Science*, 171.304–306.

ELLEFSON, MICHELLE R. AND MORTEN H. CHRISTIANSEN. 2000. Subjacency constraints without universal grammar: Evidence from artificial language learning and connectionist modeling. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 645–650. Mahwah, NJ: Lawrence Erlbaum. URL `http://citeseer.ist.psu.edu/ellefson00subjacency.html`.

ELMAN, JEFFREY L. 1990. Finding structure in time. *Cognitive Science*, 14(2).179–211. URL `citeseer.ist.psu.edu/elman90finding.html`.

FANO, ROBERT M. 1961. *Transmission of information; a statistical theory of communications*. New York: MIT Press.

FENSON, L., P. S. DALE, J. S. REZNICK, E. BATES, D. THAL, AND S. PETHICK. 1994. Variability in early communicative development. *Monographs of the society for research in child development*, 59(5). Serial 242.

FENSON, L., P.S. DALE, S. REZNICK, D. THAL, E. BATES, J.P. HARTUNG, S. PETHICK, AND J.S. REILLY. 1993. *MacArthur Communicative Development Inventories. Users Guide and Technical Manual*. San Diego, CA: Singular Publishing Group.

FISHER, W. 1996. *The `tsylb2` Program: Algorithm Description*. NIST. Part of the tsylb2-1.1 software package.

FOSLER-LUSSIER, ERIC AND C. ANTON RYTTING. 2005. A cost-benefit analysis of hybrid phone-manner representations for ASR. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 571–578. Vancouver, British Columbia, Canada: Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/H/H05/H05-1072`.

FOSLER-LUSSIER, ERIC, C. ANTON RYTTING, AND SOUNDARARAJAN SRINIVASAN. 2005. Phonetic ignorance is bliss: Investigating the effects of phonetic information reduction on ASR performance. *Interspeech 2005*. Lisbon, Portugal.

FOURAKIS, MARIOS. 1986. An acoustic study of the effects of tempo and stress on segmental intervals in modern Greek. *Phonetica*, 43(4).172–88.

FOURAKIS, MARIOS, ANTONIS BOTINIS, AND MARIA KATSAITI. 1999. Acoustic characteristics of Greek vowels. *Phonetica*, 56.28–43.

FRIEDERICI, A. D. AND J. M. WESSELS. 1993. Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception and Psychophysics*, 54(3).287–295.

GAROFOLO, J. S., L. F. LAMEL, W. M. FISHER, J. G. FISCUS, D. S. PALLETT, AND N. L. DAHLGREN. 1993. DARPA TIMIT acoustic phonetic continuous speech corpus CDROM.

GASSER, MICHAEL AND ELIANA COLUNGA. 1998. Linguistic relativity and word acquisition: A computational approach. *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, 244–249. Mahwah, NJ: Lawrence Erlbaum.

GLAVE, R. D. AND A. C. M. RIETVELD. 1975. Is the effort dependence of speech loudness explicable on the basis of acoustical cues? *Journal of the Acoustical Society of America*, 58.875–879.

GOLD, KEVIN AND BRIAN SCASSELLATI. 2006. Audio speech segmentation without language-specific knowledge. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society (CogSci06)*. Vancouver, BC. URL `http://pantheon.yale.edu/~kg253/Gold-CogSci-06.pdf`.

GRAHAM, L. W. AND A. S. HOUSE. 1971. Phonological oppositions in children: A perceptual study. *Journal of the Acoustical Society of America*, 49(2).559–566.

GREENBERG, STEVEN. 1996. Understanding speech understanding: Towards a unified theory of speech perception. *Proceedings of the ESCA Workshop on The Auditory Basis of Speech Perception*, 1–8. Keele University.

HALBERSTADT, ANDREW K. AND JAMES R. GLASS. 1997. Heterogeneous acoustic measurements for phonetic classification. *Proceedings of Eurospeech '97*, 401–404. Rhodes, Greece. URL `citeseer.ist.psu.edu/article/halberstadt97heterogeneous.html`.

HÄMÄLÄINEN, K. A., L. W. J. BOVES, AND J.M. DE VETH. 2005. Syllable-length acoustic units in large-vocabulary continuous speech recognition. *SPECOM 2005*, 499–502. Patras, Greece: University of Patras Wire Communication Laboratory.

HAMMERTON, JAMES. 2002. Learning to segment speech with self-organising maps. Tanja Gaustad, editor, *Language and Computers, Computational Linguistics in the Netherlands*, 51–64. Rodopi.

HARRINGTON, JONATHAN, SALLYANNE PALETHORPE, AND CATHERINE I. WATSON. 2000. Monophthongal vowel changes in received pronunciation: An acoustic analysis of the queen's christmas broadcasts. *Journal of the International Phonetic Association*, 30.63–78.

HARRIS, ZELIG S. 1954. Distributional structure. *Word*, 10.146–162.

HARRIS, ZELIG S. 1955. From phoneme to morpheme. *Language*, 31.190–222.

HAUSER, M. D., E. L. NEWPORT, AND R. N. ASLIN. 2001. Segmenting a continuous acoustic speech stream: Serial learning in cotton-top tamarin monkeys. *Cognition 78,*, 78.B53–B64.

HERMANSKY, HYNEK, NELSON MORGAN, ARUNA BAYYA, AND PHIL KOHN. 1991. RASTA-PLP speech analysis. ICSI technical report TR-91-069. Technical report, ICSI, Berkeley, California.

HOCKEMA, STEPHEN A. 2006. Finding words in speech: An investigation of American English. *Language Learning and Development*, 2.119–146.

HOHNE, ELIZABETH AND PETER W. JUSCZYK. 1994. Two-month-old infants' sensitivity to allophonic differences. *Perception and Psychophysics*, 56.613–623.

HOUSTON, DEREK M. AND PETER W. JUSCZYK. 2000. The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, 26(5).1570–1582.

HOUSTON, DEREK M. AND PETER W. JUSCZYK. 2003. Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance*, 29(6).1143–1154.

JAKOBSON, ROMAN AND MORRIS HALLE. 1956. *Fundamentals of language*. The Hague: Mouton.

JOHNSON, ELIZABETH K. AND PETER W. JUSCZYK. 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44 (4).548–567.

JORDAN, MICHAEL I. 1986. Serial order: A parallel distributed processing approach. Technical Report 8604, Institute for Cognitive Science, University of California, San Diego.

JOSEPH, BRIAN. 2002. The word in modern Greek. R.M.W. Dixon and A.Y. Aikhenvald, editors, *Word: A cross-linguistic typology*. Cambridge: Cambridge University Press.

JUSCZYK, P. W., A. D. FRIEDERICI, J. M. I. WESSELS, V. Y. SVENKERUD, AND A. M. JUSCZYK. 1993a. Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, 32.402–420.

JUSCZYK, P. W., A. M. JUSCZYK, L. J. KENNEDY, T. SCHOMBERG, AND N. KOENIG. 1995. Young infants' retention of information about bisyllabic utterances. *Journal of Experimental Psychology: Human Perception and Performance*, 21.822–836.

JUSCZYK, P. W. AND E. J. THOMPSON. 1978. Perception of a phonetic contrast in multisyllabic utterances by two-month-old infants. *Perception and Psychophysics*, 23.105–109.

JUSCZYK, PETER W. 1997. Finding and remembering words: Some beginnings by English-learning infants. *Current Directions in Psychological Science*, 6(6).170–174. URL http://www.blackwell-synergy.com/doi/abs/10.1111/1467-8721.ep%10772947.

JUSCZYK, PETER W. AND RICHARD N. ASLIN. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29(1).1–23.

JUSCZYK, PETER W., ANNE CUTLER, AND NJ REDANZ. 1993b. Infants preference for the predominant stress patterns of English words. *Child Development*, 64(3).675–687.

JUSCZYK, PETER W., E. A. HOHNE, AND A. BAUMAN. 1999a. Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, 61.1465–1476.

JUSCZYK, PETER W., DEREK HOUSTON, AND MARY NEWSOME. 1999b. The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39.159–207.

KAPPA, IOANNA. 2002. On the acquisition of syllabic structure in Greek. *Journal of Greek Linguistics*, 3.1–52.

KENYON, JOHN S. 1950. *American Pronunciation*. Ann Arbor, Mich.: Wahr.

KORMAN, MYRON. 1984. Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First Language*, 5.44–45.

LADEFOGED, PETER. 1993. *A course in phonetics*. Fort Worth, TX: Harcourt, Brace, and Jovanovich, 3rd edition.

LI, PING AND BRIAN MACWHINNEY. 2002. PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, and Computers*, 34.408–415.

LIEBERMAN, PHILIP. 1996. Some biological constraints on the analysis of prosody. Morgan and Demuth (1996), 55–65.

LIN, YING. 2005. *Learning Features and Segments from Waveforms: A Statistical Model of Early Phonological Acquisition*. Ph.D. thesis, UCLA department of linguistics. URL http://www.humnet.ucla.edu/humnet/linguistics/faciliti/resear%ch/research.html/YLin_diss.pdf.

LONG, DANIEL AND PETER TRUDGILL. 2005. The last Yankee in the Pacific: Eastern New England phonology in the Bonin Islands. *American Speech*, 79(4).356–367.

LU, XIAOFEI. 2006. *Hybrid Models for Chinese Unknown Word Resolution*. Ph.D. thesis, The Ohio State University. URL http://www.ohiolink.edu/etd/view.cgi?acc_num=osu1154631880.

LUPYAN, GARY AND MORTEN H. CHRISTIANSEN. 2002. Case, word order and language learnability: Insights from connectionist modeling. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 220–225. Mahwah, NJ: Lawrence Erlbaum. URL http://cnl.psych.cornell.edu/papers/LandC-cogsci2002.pdf.

MACWHINNEY, BRIAN. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Erlbaum.

MANDEL, D.R., P.W. JUSCZYK, AND D.B. PISONI. 1995. Infants' recognition of the sound patterns of their own names. *Psychological Science*, 6.314–317.

MANDELBROT, B. 1966. Information theory and psycholinguistics: A theory of words frequencies. P. Lazafeld and N. Henry, editors, *Readings in Mathematical Social Science*. Cambridge, MA: MIT Press.

MANNING, CHRISTOPHER D. AND HINRICH SCHUTZE. 1999. *Foundations of statistical natural language processing*. MIT Press.

MARCUS, GARY F., S. VIJAYAN, S. BANDI RAO, AND P. M. VISHTON. 1999. Rule-learning in seven-month-old infants. *Science*, 283.77–80. URL http://www.psych.nyu.edu/gary/marcusArticles/marcus\%20et\%20%al\%201999\%20science.pdf.

MATTYS, SVEN L. AND PETER W. JUSCZYK. 2001a. Do infants segment words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance*, 27(3).644–655.

MATTYS, SVEN L. AND PETER W. JUSCZYK. 2001b. Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78(2).91–121.

MATTYS, SVEN L., PETER W. JUSCZYK, PAUL A. LUCE, AND JAMES L. MORGAN. 1999. Word segmentation in infants: How phonotactics and prosody combine. *Cognitive Psychology*, 38.465–494.

MATTYS, SVEN L. AND A.G. SAMUEL. 2000. Implications of stress-pattern differences in spoken-word recognition. *Journal of Memory and Language*, 42.571–596. URL http://www.ingentaconnect.com/content/ap/ml/2000/00000042/000%00004/art02696.

MATTYS, SVEN L., LAURENCE WHITE, AND JAMES F. MELHORN. 2005. Integration of multiple segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, 134.477–500.

MCMURRAY, BOB AND RICHARD N. ASLIN. 2005. Infants are sensitive to within-category variation in speech perception. *Cognition*, 95(2).B15–B26. URL http://www.sciencedirect.com/science/article/B6T24-4F3FF3K-1/%2/719f951c06f9986c002689f3cf6ab6ff.

MCMURRAY, BOB, MICHAEL K. TANENHAUS, AND RICHARD N. ASLIN. 2002. Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86.B33–B42. URL http://www.sciencedirect.com/science/article/B6T24-475WPVX-5/%2/567c063ab192123a424d837d7d8dccdd.

MEHLER, J., P. JUSCZYK, G. LAMBERTZ, N. HALSTED, J. BERTONCINI, AND C. AMIEL-TISON. 1988. A precursor of language acquisition in young infants. *Cognition*, 29.143–178.

MIELKE, JEFF. 2004. *The emergence of distinctive features*. Ph.D. thesis, The Ohio State University.

MORGAN, JAMES L. 1996. A rhythmic bias in preverbal speech segmentation. *Journal of Memory and Language*, 35.666–688.

MORGAN, JAMES L. AND KATHERINE DEMUTH, editors. 1996. *Signal to Syntax*. Mahwah, NJ: Lawrence Erlbaum.

MORGAN, JAMES L. AND JENNY R. SAFFRAN. 1995. Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development*, 66.911–936.

NAZZI, T., J. BERTONCINI, AND J. MEHLER. 1998. Language discrimination by newborns: Towards un understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24.1–11.

NEWMAN, R. S., N. BERNSTEIN RATNER, A. M. JUSCZYK, P. W. JUSCZYK, AND K. A. DOW. 2006. Infants' early ability to segment the conversational speech signal predicts later language development: a retrospective analysis. *Developmental Psychology*, 42(4).643–655.

NEWMAN, ROCHELLE S. 2005. The cocktail party effect in infants revisited: Listening to one's name in noise. *Developmental Psychology*, 41(2).352–362.

NEWMAN, ROCHELLE S. AND PETER W. JUSCZYK. 1996. The cocktail party effect in infants. *Perception and Psychophysics*, 58(8).1145–1156.

NEWPORT, ELISSA L., DAN J. WEISS, ELIZABETH WONNACOTT, AND RICHARD N. ASLIN. 2004. Statistical learning in speech: Syllables or segments? Paper presented at the 29th Annual Boston University Conference on Language Development, Boston.

NORRIS, DENNIS. 1994. Shortlist: a connectionist model of continuous speech recognition. *Cognition*, 52(3).189–234. URL `http://www.sciencedirect.com/science/article/B6T24-4608WBT-2/%2/9b878b4c257c11f6e50f48ecdd9679dc`.

OHALA, JOHN J. 1990. The phonetics and phonology of aspects of assimilation. John Kingston and Mery E. Beckman, editors, *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, 258–275. Cambridge University Press.

OLIVIER, D. C. 1968. *Stochastic grammars and language acquisition mechanisms*. Ph.D. thesis, Harvard University, Cambridge, MA.

PEGG, JUDITH E. AND JANET F. WERKER. 1997. Adult and infant perception of two English phones. *Journal of the Acoustical Society of America*, 102(6).3742–3753.

PELLOM, BRYAN. 2001. SONIC: The University of Colorado continuous speech recognizer: Technical report TR-CSLR-2001-01. Technical report, University of Colorado.

PELLOM, BRYAN AND KADRI HACIOGLU. 2003. Recent improvements in the CU SONIC ASR system noisy speech: The SPINE task. *Proceedings of the IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP)*. Hong Kong.

PERCY, THOMAS. 1906. *Reliques of Ancient English Poetry*. London and Bungay: Richard Clay and Sons. [Originally published 1765.].

PERRUCHET, PIERRE AND ANNE VINTER. 1998. PARSER: A model for word segmentation. *Journal of Memory and Language*, 39.246–263.

PIERREHUMBERT, JANET B. 2003. Phonetic diversity, statistical learning and acquistion of phonology. *Language and Speech*, 46(2/3).115–154.

PITT, MARK A., KEITH JOHNSON, ELIZABETH HUME, SCOTT KIESLING, AND WILLIAM RAYMOND. 2005. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45.89–95.

PLAUT, DAVID C. AND CHRISTOPHER T. KELLO. 1999. The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. Brian MacWhinney, editor, *Emergence of Language*. Hillsdale, NJ: Lawrence Earlbaum Associates. URL `http://www.isrl.uiuc.edu/~amag/langev/paper/plaut99theEmergen%ce.html`.

POLKA, LINDA, MEGHA SUNDARA, AND STEPHANIE BLUE. 2002. The role of language experience in word segmentation: A comparison of English, French, and bilingual infants. Paper presented at the 143rd Meeting of the Acoustical Society of America: Special Session in Memory of Peter Jusczyk, Pittsburgh, Pennsylvania.

RABINER, LAWRENCE R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, volume 77, 257–285.

REALI, FLORENCIA, MORTEN H. CHRISTIANSEN, AND PADRAIG MONAGHAN. 2003. Phonological and Distributional Cues in Syntax Acquisition: Scaling up the Connectionist Approach to Multiple-Cue Integration. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, 970–975. Mahwah, NJ: Lawrence Erlbaum.

ROY, DEB. 1999. *Learning form sights and sounds: A computational model*. Ph.D. thesis, MIT Media Laboratory.

ROY, DEB AND ALEX PENTLAND. 2002. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26(1).113–146.

RYTTING, C. ANTON. 2004a. Greek word segmentation using minimal information. *Proceedings of the Student Research Workshop at HLT/NAACL 2004*, 207–212. Boston, Massachusetts: Association for Computational Linguistics. URL `http://acl.ldc.upenn.edu/hlt-naacl2004/studws/pdf/sw-8.pdf`.

RYTTING, C. ANTON. 2004b. Segment predictability as a cue in word segmentation: Application to modern Greek. *Current themes in computational phonology and morphology: seventh meeting of the ACL special interest group for computational phonology (SIGPHON'04)*, 78–85. Barcelona, Spain: The Association for Computational Linguistics. URL http://acl.ldc.upenn.edu/acl2004/sigphon/pdf/rytting.pdf.

RYTTING, C. ANTON. 2005. An iota of difference: attitudes to *yod* in lexical and social contexts. *The Journal of Greek Linguistics*, 6.151–185. URL http://www.benjamins.com/cgi-bin/t_articles.cgi?bookid=JGL\%2006&artid=503061682.

RYTTING, C. ANTON. 2006a. Finding the gaps: Applying a connectionist model of word segmentation to noisy phone-recognized speech data. *Proceedings of Interspeech'06*. Pittsburgh, PA. URL http://www.interspeech2006.org/papersearch/IS06papers/IS06206%2.PDF.

RYTTING, C. ANTON. 2006b. Is recurrence redundant? revisiting Allen and Christiansen (1996). *Proceedings of the 28th Annual Meeting of the Cognitive Science Society (CogSci06)*, 2065–2070. Vancouver, Canada: The Cognitive Science Society. URL http://www.cogsci.rpi.edu/CSJarchive/Proceedings/2006/docs/p2%065.pdf.

SAFFRAN, JENNY R., RICHARD N. ASLIN, AND ELISSA L. NEWPORT. 1996a. Statistical cues in language acquisition: Word segmentation by infants. G.W. Cottrell, editor, *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, 376–380. Hillsdale, NJ: Lawrence Erlbaum Associates.

SAFFRAN, JENNY R., RICHARD N. ASLIN, AND ELISSA L. NEWPORT. 1996b. Statistical learning by 8-month-old infants. *Science*, 274.1926–1928.

SAFFRAN, JENNY R. AND ERIK D. THIESSEN. 2003. Pattern induction by infant language learners. *Developmental Psychology*, 39(3).484–494.

SANSAVINI, A., J. BERTONCINI, AND G. GIOVANELLI. 1997. Newborns discriminate the rhythm of multisyllabic stressed words. *Developmental Psychology*, 33, 3-l 1.

SAUSSURE, FERDINAND DE. [1916] 1983. *Cours de linguistique gĄénĄérale [Course in General Linguistics]*. London: Duckworth. Trans. by Roy Harris.

SCHARENBORG, ODETTE, DENNIS NORRIS, LOUIS TEN BOSCH, AND JAMES M. MCQUEEN. 2005. How should a speech recognizer work? *Cognitive Science*, 29(6).867–918. URL /cgi-bin/sciserv.pl?collection=journals&journal=03640213&issu%e=v29i0006&article=867_hsasrw.

SEIDL, AMANDA AND ELIZABETH K. JOHNSON. 2006. Infant word segmentation revisited: edge alignment facilitates target extraction. *Developmental Science*, 9(6).565–573. URL http://www.blackwell-synergy.com/doi/abs/10.1111/j.1467-7687.%2006.00534.x.

SETHY, A., B. RAMABHADRAN, AND S. NARAYANAN. 2003. Improvements in ASR for the MALACH project using syllable-centric models. *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*. St. Thomas.

SHI, RUSHEN, JANET F. WERKER, AND ANNE CUTLER. 2003. Function words in early speech perception. *Proceedings of the 15th International Conference of Phonetic Sciences*, 3009 3012. Adelaide, Australia: Causal Productions.

SHIPMAN, D. W. AND V. W. ZUE. 1982. Properties of large lexicons: Implications for advanced isolated word recognition systems. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 82, 546–549. Paris, France.

SLUIJTER, AGAATH M. C., VINCENT J. VAN HEUVEN, AND JOS J. A. PACILLY. 1997. Spectral balance as a cue in the perception of linguistic stress. *The Journal of the Acoustical Society of America*, 101(1).503–513. URL /cgi-bin/sciserv.pl?collection=journals&journal=00014966&issu%e=v101i0001&article=503_sbaacitpols.

STEPHANY, URSULA. 1995. The acquisition of Greek. Dan I. Slobin, editor, *The crosslinguistic study of language acquisition*, volume 4. Lawrence Erlbaum.

SUDDARTH, S. C. AND Y. L. KERGOSIEN. 1990. Rule-injection Hints as a Means of Improving network Performance and Learning Time. L.B. Almeida and C.J. Wellekens, editors, *Proceedings of the EURASIP Workshop 1990 on Neural Networks*, 120–129. Berlin: Springer-Verlag.

SUN, JIPING AND LI DENG. 2002. An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition. *Journal of the Acoustical Society of America*, 111.1086–1101.

SVARTVIK, JAN AND RANDOLPH QUIRK. 1980. *A corpus of English conversation*. Lund: LiberLaromedel Lund.

SWINGLEY, DANIEL. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50.86–132. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db%=PubMed&list_uids=15556130&dopt=Abstract.

THIESSEN, ERIK D., E.A. HILL, AND JENNY R. SAFFRAN. 2005. Infant directed speech facilitates word segmentation. *Infancy*, 7.49–67.

THIESSEN, ERIK D. AND JENNY R. SAFFRAN. 2003. When cues collide: Use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental Psychology*, 39(4).706–716.

THIESSEN, ERIK D. AND JENNY R. SAFFRAN. 2004. Spectral tilt as a cue to word segmentation in infancy and adulthood. *Perception and Psychophysics*, 66(5).779–791.

THRAX, DIONYSIOS. 1883. *Grammatike Techne [Ars grammatica—The Art of Grammar]*. Grammatici graeci recogniti et apparatv, critico instructi, partis 1. Lipsiae: B. G. Tevbner.

TINCOFF, R. AND P.W. JUSCZYK. 1999. Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, 10.172–175.

TRUBETZKOY, NIKOLAI S. 1939. *Principles of Phonology. (1969 English translation)*. Berkeley, CA: University of California Press.

VOGEL, IRENE AND ERIC RAIMY. 2002. The acquisition of compound vs. phrasal stress: The role of prosodic constituents. *Journal of Child Language*, 29(2).225–250.

VOULOUMANOS, ATHENA AND JANET F. WERKER. 2004. Tuned to the signal: The privileged status of speech for young infants. *Developmental Science*, 7(3).270–276. URL http://www.blackwell-synergy.com/doi/abs/10.1111/j.1467-7687.%2004.00345.x.

WELBY, PAULINE S. 2003. *The Slaying of Lady Mondegreen, being a Study of French Tonal Association and Alignment and their Role in Speech Segmentation*. Ph.D. thesis, The Ohio State University.

WERKER, JANET AND R. C. TEES. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7.49–63.

WERKER, JANET F. AND SUZANNE CURTIN. 2005. PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, 1(2).197–234.

WERKER, J.F., C.T. FENNELL, K.M. CORCORAN, AND C.L. STAGER. 2002. Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, 3.1–30.

WILSON, MICHAEL D. 1987. Mrc psycholinguistic database: Machine usable dictionary. URL `citeseer.ist.psu.edu/wilson87mrc.html`, available from `http://www.psy.uwa.edu.au/uwa_mrc.htm`.

WINITZ, H., M. E. SCHEIB, AND J. A. REEDS. 1972. Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech. *Journal of the Acoustical Society of America*, 51.1309–1317.

WOLFF, J.G. 1977. The discovery of segments in natural language. *British Journal of Psychology*, 68.97–106.

WRIGHT, SYLVIA. 1954. The death of Lady Mondegreen. *Harper's Magazine*, 209.48–51.

WU, SU-LIN, MICHAEL L. SHIRE, STEVEN GREENBERG, AND NELSON MORGAN. 1997. Integrating syllable boundary information into speech recognition. *ICASSP97*. Munich, Germany: IEEE.

YOUNG, S., G. EVERMANN, T. HAIN, D. KERSHAW, G. MOO RE, J. ODELL, D. OLLASON, D. POVEY, V. VALTCHEV, AND P. WOODLAND. 2002. *The HTK Book*. Cambridge Unveristy Engineering Department. Http://htk.eng.cam.ac.uk.

YU, CHEN AND DANA H. BALLARD. 2002. A computational model of embodied language learning. Technical Report Tech. Rep. 791, Department of Computer Science, University of Robchester. URL `citeseer.ist.psu.edu/yu02computational.html`.

ZIPF, G.K. 1965. *The Psycho-Biology of Language, An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press.