UNIVERSITY OF CALIFORNIA,
IRVINE

The Role of Empirical Evidence in Modeling Speech Segmentation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Psychology

by

Lawrence Phillips

Dissertation Committee:
Associate Professor Lisa Pearl, Chair
Professor Michael Lee
Associate Professor Barbara Sarnecka

2015

DEDICATION

To my mother, who inspired me to always continue learning.
To Scott Lawler, who helped remind me of the value of life outside of work.
To Kathy Cote and Holt Clark, who showed me that the path to success is full of twists and turns.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work would not have been possible without the support, advice, and general wisdom of many others. For their technical expertise and advice throughout my graduate career, I would like to give special thanks to Lisa Pearl, Michael Lee, Barbara Sarnecka, Mark Steyvers, Alex Ihler, Jon Sprouse, and Jim White.

None of this would have been possible where it not for Clara Schultheiss, John Sommerhauser, Adam Cook, and Jessica Cañas-Castañeda and the administrative support they were able to provide.

I would also like to give special thanks to David Pisoni for loaning me a copy of *The Mind's New Science*, which introduced me to the field of Cognitive Science. Many of the lessons I learned working in the Speech Research Laboratory provided the foundations for my graduate career.

CURRICULUM VITAE

Lawrence Phillips

EDUCATION

Doctor of Philosophy in Psychology 2015
University of California, Irvine *Irvine, California*

Bachelor of Arts in Cognitive Science 2010
Indiana University *Bloomington, Indiana*

Bachelor of Arts in Linguistics 2010
Indiana University *Bloomington, Indiana*

Bachelor of Arts in Germanic Studies 2010
Indiana University *Bloomington, Indiana*

RESEARCH EXPERIENCE

Graduate Research Assistant 2011
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant 2010–2015
University of California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation 2015
Cognitive Science

Perceptual adaptation to sinewave vocoded speech across languages 2009
Journal of Experimental Psychology: Human Perception and Performance

REFEREED CONFERENCE PUBLICATIONS

Utility-based evaluation metrics for models of language acquisition: A look at speech segmentation Jun 2015
Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics

Bayesian inference as a viable cross-linguistic word segmentation strategy: It's all about what's useful Jul 2014
Proceedings of the 36th annual conference of the Cognitive Science Society

Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things Apr 2014
Proceedings of the Workshop on Cognitive Aspects of Computational Language Learning

"Less is More" in Bayesian word segmentation: When cognitively plausible learners outperform the ideal Aug 2012
Proceedings of the 34th annual conference of the Cognitive Science Society

Taking the child's view: Syllable-based word segmentation as a (more) plausible word segmentation strategy Jan 2012
Proceedings of the 86th annual meeting of the Linguistics Society of America

ABSTRACT OF THE DISSERTATION

The Role of Empirical Evidence in Modeling Speech Segmentation

By

Lawrence Phillips

Doctor of Philosophy in Psychology

University of California, Irvine, 2015

Associate Professor Lisa Pearl, Chair

Choosing specific implementational details is one of the most important aspects of creating and evaluating a model. In order to properly model cognitive processes, choices for these details must be made based on empirical research. Unfortunately, modelers are often forced to make decisions in the absence of relevant data. My work investigates the effects of these decisions. Looking at infant speech segmentation, I incorporate empirical research into model choices regarding model input, inference, and evaluation. First, I use experimental results to argue for syllables as a basic unit for early segmentation and show that the segmentation task is less difficult than previously thought. I then explore the role of various inference algorithms, each of which produces testable predictions. Lastly, I argue that standard methods of model evaluation make unrealistic assumptions about the goal of learning. Evaluating models in terms of their ability to support additional learning tasks shows that gold standard performance alone is an insufficient metric for measuring segmentation quality. In each of these three instances, I treat model design decisions as free parameters whose impact must be evaluated. By following this approach, future researchers can better gauge the success or failure of cognitive models.

Chapter 1

Modeling Language Acquisition

The acquisition of a native language is a curious feat. Nearly every human being has succeeded in this task, yet despite our collective linguistic prowess, we struggle to learn the languages of our neighbors. As we age, we gain an increased understanding of the world, building constantly upon our previous knowledge, but it is the newborn and not their parents who best succeeds in learning a new language. This raises the question of why there should be any area in which the ability of infants appears to outstrip that of those older, wiser, and otherwise more cognitively capable.

One method to address how children learn their native language is to propose a theory of learning and then demonstrate that the theory succeeds for actual human languages. Computational modeling exists as a tool which can address these types of questions. This method takes a theory of learning and instantiates it explicitly in a way which can be simulated by a computer – this is then called the computational model. This model can be given the same types of information that children receive and researchers can then measure what the model learns. While a powerful process, making an explicit model requires any number of assumptions to be made about the learning process. What types of information

are encountered by learners? What (if any) cognitive limitations are incorporated into the learning process? In what way is the model evaluated?

This dissertation is an attempt to explore the role of these design decisions. Models generally have what are known as *free parameters*, unknown values which must either be learned or set constant. For parameters which cannot be learned by the model, researchers generally explore how changing the value of a free parameter influences the behavior of a model. Failing to do so opens up possible criticisms. For example, if learning requires a particular parameter value, how would children come to know that specific value? In the same way, design choices are often set by the researcher and can be similarly manipulated. When this is not done, it is unknown whether learning depends crucially on the particular values chosen by the researcher. Working in the domain of speech segmentation, I show how exploring the impact of model design choices can lead to a better understanding of model behavior, which in turn impacts our understanding of human behavior. It also allows us to generate testable predictions for future research.

This chapter serves as an introduction to computational modeling and its role within the study of language acquisition. I discuss a number of common problems with computational models and propose possible solutions.

1.1 Understanding Computational Modeling

1.1.1 Difficulties with Experimental Techniques

One of the largest problems with experimental and observational research on language acquisition is that while these methods are well suited to describe both what is learned and at what age, probing the way in which learning occurs can be difficult. One such area of

research is that of segmenting fluent speech into smaller units, often referred to as either word or speech segmentation. While there are many possible cues to segmentation, one which has received much experimental attention is the use of statistical information. In Saffran et al. (1996) it was shown that 8-month-old infants, exposed to only two minutes of an artificial language, were able to properly segment its words using only the statistical information available in the input. This experiment was a landmark in the field because it helped to conclusively show that young infants are capable of tracking statistical regularities such as the co-occurrences between syllables. Experiments of this kind are incredibly useful in that they can show:

1. At what age infants are able to solve particular language acquisition tasks, such as speech segmentation.

2. What cues infants are able to use in solving a particular task.

Subsequent research has explored this ability at other ages (Teinonen et al., 2009; Johnson and Tyler, 2010), in other animals (Hauser et al., 2001), in other domains (N.Z. Kirkham, 2002), using naturalistic stimuli (Pelucchi et al., 2009), and in relation to other cues (Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003). Despite the large amount of research into this area, there are still many open questions.

First, how is the statistical information in the ambient language being calculated? Saffran et al. (1996) describe their data in terms of forward transitional probability (TP), the probability that one syllable will occur *after* having heard the previous syllable. Pelucchi et al. (2009) show, however, that the same task can be solved when the only useful statistical information is backwards TP, the probability of a syllable appearing *before* another syllable. Do infants calculate both statistics or are they tracking co-occurrences in some other manner that accounts for both findings? Because there are so many possibilities and designing

experiments to disambiguate between them is both difficult and expensive, there is little experimental evidence that can weigh in on this debate.

Second, given that statistical information in these experiments is being calculated in *some* manner, is this ability robust enough to account for the segmentation ability of infants learning an *actual* language? Later experiments showed that when the stimuli consisted of words of varying lengths, learning did not occur (Johnson and Tyler, 2010). This work suggests that the use of statistical information by language learners may be more complex than previous studies suggested. Further, a cue such as TP can be used in multiple ways. Gambell and Yang (2006) showed that a naive strategy using TPs fails in the segmentation of English, but a more complex strategy might find greater success. More research is needed to determine exactly what strategy children are actually making use of, but the controlled environment of the laboratory seems to be ill-suited to answering these kinds of questions. Instead, explicit computational models can be used to demonstrate whether a particular theory is actually capable of learning an actual language.

1.1.2 Computational Modeling as a Solution

The benefit of experimental research is that it provides constraints on the number of possible theories of language acquisition. Any theory of speech segmentation in infants must account for the kinds of experimental findings discussed previously. Any theory which does not can be discarded. Yet there still remains the problem that experimental data are rarely able to narrow the set of possibilities down to a single theory. In order to further refine our understanding of how language acquisition proceeds, researchers need to examine individual theories and test their merits.

One way to do this is to use computational modeling. This is not without its own challenges. Making a computational model requires filling in many of the details which are left unan-

swered by previous research (Pearl, 2014; Pearl and Sprouse, 2015). As discussed above, experiments suggest that infants are able to make use of TPs between syllables in order to segment words. What's unclear, however, is how exactly this is accomplished. Do infants calculate statistics in terms of TP or by some other metric, such as mutual information (Swingley, 2005)? How do infants make use of these statistics in order to identify word boundaries? Are they only positing boundaries or are they also keeping track of the words in between and storing them in memory? A computational model is forced to address these questions in ways that theories of learning may not. Once these decisions have been made and a model has been created, it can then be used to simulate an acquisition task. This provides a number of concrete benefits:

1. Theories can be tested to see if they account for experimental results.
2. Theories can be tested on naturalistic data similar to what actual infants are exposed to.
3. The results of these simulations can be examined in great detail, allowing for researchers to discover novel, sometimes unintuitive predictions of a theory.

Of course, just because a model is successful does not necessarily mean that it represents the way in which children *actually* learn, but it does serve as an existence proof that learning might proceed in this manner. Moreover, because language learning is such a difficult endeavor, a model completely without merit is unlikely to fit with experimental and observational data related to language learning (Hoff, 2008).

1.2 Questioning Computational Modeling

As discussed previously, creating a computational model entails making choices about the learning process, many of which are simplifying assumptions. For instance, a segmentation model may choose to ignore issues of word meaning. This simplifies the learning problem in a way which may or may not be true. While creating a model which captures the entirety of the learning process is ideal, in practice this is generally not feasible.

However, the closer a model is able to match reality, the easier it is to generalize the simulation's findings to real life. Making decisions which do not correspond to reality increases the chance that the simulation's results become difficult or impossible to interpret. For example, a theory of speech segmentation might posit that children make use of statistical regularities in order to segment speech. Here's a model that could be made of this process: the model is given a corpus of unsegmented, orthographic text and produces a segmentation which closely matches the true segmentation. An implicit assumption of this model is that children learn using written text, which is clearly not the case. Because the model's assumptions are unfounded, it tells researchers very little about the true segmentation problem faced by infants and how they might solve it. Further work must be done to improve the assumptions of the model, perhaps by training the model using data which matches the kind of data children actually receive: spoken language.

1.3 Improving Computational Modeling

Because of these issues, there is a growing interest in the "cognitive plausibilty" of computational models (Anderson, 1990; Shi et al., 2010; Bonawitz et al., 2011; Pearl et al., 2011; Griffiths et al., 2015). Models whose assumptions more closely match reality are more plau-

sible, and therefore the results of these models are taken more seriously. This is intricately tied to the notion of external validity, the ability to take the results of a study and generalize them to other groups. What counts as plausible in models of language acquisition is a matter of some debate. The reason for this is that many assumptions exist without evidence for or against their validity. In the absence of any confirmation, how should these assumptions be judged?

Once a model has been created, there are three important areas where decisions must be made. First, the model must be given some *input*, data from which it can learn. Second, the model must have some method of performing learning once it encounters this data; this process is generally referred to as model *inference*. Finally, the model will produce some *output* and it must be evaluated to determine whether the model was successful. In this section, I go over these three areas, covering common concerns regarding these assumptions and discussing briefly how improvements might generally be made in each area.

1.3.1 Improving Model Input

Determining what data to give to a model is a question of great importance. As the computer science quote goes, "Garbage in, garbage out" (McRae, 1964). What then is the best way to ensure that the data given to a language acquisition model most closely reflects what children actually encounter?

With the advent of widespread computing and access to the internet, large collections of child-directed speech (CDS) are now available via the CHILDES database (MacWhinney, 2000). These corpora are made up of acoustic, orthographic, and phonological representations of actual speech that was made in the presence of actual children. Having large quantities of speech that is representative of the input given to actual children is an essential requirement for properly modeling language acquisition. Still, the process of generating a

CDS corpus is quite involved and even large corpora do not come close to approximating the amount of language a child actually encounters. A corpus might be made up of tens of thousands of utterances, yet that pales in comparison to the amount of speech a child will encounter, even in just the first year of life. Take, for example, the Brent corpus which is comprised of 100 hours worth of speech produced in the presence of children (Brent and Siskind, 2001). The amount of CDS in that corpus is only 144,474 utterances. If a child is awake and in the presence of adults for 10 hours a day, that implies the entire corpus represents only 10 days worth of language. Giving a model the amount of language that a child might encounter in the first year of life is simply impossible with current corpora (although the work of Roy et al. (2006) may eventually change this).

While it may be impossible to train a model using the same amount of data a child receives, a model can at least be trained using the kind of data that a child receives. In particular, this requires using not just the speech produced near a child, but focusing only on the speech directed to the child, since research shows that children both prefer CDS (Fernald, 1985) and the amount of this kind of speech they encounter best predicts their later language ability (Weisleder and Fernald, 2013). Further, not all CDS is the same. Caretakers vary their speech as a child ages (Bernstein-Ratner, 1984; Gleitman et al., 1984; Kitamura and Burnham, 2003) and therefore it is important to ensure not only that the input is CDS, but that it is speech aimed at children of the same age which is being modeled. Because it is unknown exactly how differences in CDS at various ages might interact with the results of a model, deviating from the ideal age range might have unforeseen consequences in terms of model results.

An additional complication is that before the model can be given this input, the researcher must make a decision about how that input should be represented to the model. First, a decision must be made regarding which cues in the input are available to the learner. For example, in segmentation do children make use of the stress patterns placed on syllables

or is that information ignored? Second, a decision must be made regarding how these cues are represented by a child learner. For example, do children perceive speech as a string of phones, phonemes, or some other unit such as syllables? While it is often convenient to represent a corpus in the most convenient fashion (typically phonemically), this does not always match what is known about child learners. Therefore, so far as possible, researchers should encode their model input in a manner which matches what is experimentally known about child language processing at the age being modeled.

1.3.2 Improving Model Inference

Once an appropriate model input has been chosen, the researcher must then determine how learning proceeds for the model. Because the model is drawing conclusions based on the data it encounters, this is referred to as model inference. How inference is carried out generally relates to the goals of the researcher. If a researcher simply wants to know whether or not a strategy *could possibly* solve a particular linguistic task (akin to the *computational level* of Marr, 1982), then inference should be done as optimally as possible[1]. This often forms the first step in analyzing a theory of learning and is done quite frequently (Johnson et al., 2007; Feldman et al., 2009; Frank et al., 2009; Goldwater et al., 2009; Christodoulopoulos et al., 2011).

Once a model has been shown to succeed at a task with optimal inference, it is often interesting to determine whether the model might be modified to perform more "plausible" inference. The goal of this is to show that children could use the strategy given the limited resources available to them (akin to the *algorithmic level* of Marr, 1982). Generally, researchers attempt to do this in one of three fashions:

[1]Optimal solutions are common in Bayesian methods, where inference algorithms, such as Gibbs sampling (Geman and Geman, 1984), often have guarantees of convergence.

1. Incorporate online inference

2. Incorporate non-optimal decision making

3. Incorporate memory constraints

1.3.2.1 Online Inference

Optimal inference strategies often operate in *batch*, gathering some large amount of data and then performing inference all at once. Incorporating *online* inference refers to the fact that individuals, children and adults alike, do not wait to process information. Instead, they process it as it comes in. Online inference methods generally perform worse than their batch counterparts, which is to be expected since they are forced to make decisions using much more limited information. In some cases however, they possess unique benefits such as learning quickly and avoiding local optima (Liang and Klein, 2009). Because online inference has many practical uses, there are many standard online inference algorithms available to researchers and it has been widely used by many cognitive models (Brent, 1999; Venkataraman, 2001; Vallabha et al., 2007; Wang and Mintz, 2008; Blanchard et al., 2010; Pearl et al., 2011; Lignos, 2011).

1.3.2.2 Non-Optimal Decision Making

A lack of optimality may be inherent to the process of human inference (Tversky and Kahneman, 1974; Kahneman et al., 1982), and in such a case it makes sense to model this sub-optimality directly. This can be straightforwardly accomplished by having the model choose options proportional to their probabilities, rather than always choosing the best option. This process is known as probability matching and can be found in both infant and adult learning (Kam and Newport, 2005; Kam and Chang, 2009; Denison et al., 2013).

1.3.2.3 Memory Constraints

Other, more specific cognitive constraints may also be modeled. It is well known, for instance, that humans have limited short-term memory (Miller, 1956; Atkinson and Shiffrin, 1968). Likewise, a model might be forced to make decisions while only being able to retain a certain amount of information. This type of limitation is less commonly implemented than other cognitive limitations but is still occasionally used (Wang and Mintz, 2008). Another method for implementing this style of constraint is to make use of the fact that limited memory focuses attention on recently encountered items (Murdock Jr, 1962). This type of shift in processing resources can be mimicked through the use of certain inference algorithms, such as the Decayed Markov Chain Monte Carlo process (Marthi et al., 2002; Pearl et al., 2011).

1.3.3 Improving Model Output Evaluation

Once a model has generated output, the researcher is then faced with determining whether learning has been successful. While this may seem to be a straightforward process, in practice it is often unclear what level of success should be expected from a model. If the goal is to match the behavior of children, then having baselines for performance would be ideal for comparison. However, the output of the model is often an entire segmented corpus. Obtaining experimental evidence as to how the corpus would be segmented by young infants is simply infeasible. Researchers are then left with making a decision, based on very little evidence, about whether a model with, for example, 80% accuracy counts as "successful". To avoid this scenario, it is important for the researcher to determine how the model output might best be evaluated and in what ways it might be compared to actual data from children.

One fact about language acquisition which a model can be compared against is that acquistion succeeds across all natural human languages. Therefore, a good model of language

acquisition should likewise succeed regardless of the language being learned. While a model which has succeeded on a single language is impressive and worthy of further research, the success of the model may be due to idiosyncratic properties of the language. For instance, imagine a speech segmentation model which assumes that words tend to have very few syllables. This model will perform quite well on languages such as English or Mandarin, which are heavily monosyllabic. The model will do much more poorly on a language such as Japanese or Hawaiian, which have words made up of many more syllables. A good segmentation model should be resilient to such changes and this can only be ensured if the model is tested across a wide range of languages.

In the most straightforward case, the model might be trained using experimental data where the output can be easily compared with results from children (Frank et al., 2009; Sanborn et al., 2010; Kolodny et al., 2015). This serves as a useful proof that the model is capable of accounting for experimental findings. On the other hand, there may be many models capable of accounting for a single experimental finding. Successfully modeling experimental results also does not address the larger question of whether the model is capable of learning actual languages outside of the laboratory setting.

In other cases, there may be qualitative patterns that the model can be compared against. For instance, specific types of segmentation errors have been reported in children 2-3 years of age, long after speech segmentation has begun (Brown, 1973; Peters, 1983). In particular, children appear to treat groups of function words as single units (e.g. *what's that* segmented as *what'sthat*), while in other cases function words are segmented out of larger words (e.g. *behave* segmented as *be have*). While there is no quantitative standard, models which produce these types of long-lasting errors might be preferred. Similarly, it's known that infants learn the predominant stress pattern in their language relatively early (7.5 months: Jusczyk et al., 1999b). One could compare the words that a segmentation model produces in order to see

what stress pattern they predict[2]. A model which captures the stress patterns that children learn would, in this case, be preferred over one which predicts a different pattern.

1.4 Conclusion

This chapter has covered a number of issues and possible solutions in the modeling of infant language acquisition. In Chapter 2, I introduce in more detail the problem of speech segmentation and cover the empirical work relevant to this particular learning task. Then, in Chapter 3, I introduce a number of existing segmentation models, covering their strengths and weaknesses. Having identified areas in need of improvement, I then present my simulations regarding model input (Chapter 4), model inference (Chapter 5), and model evaluation (Chapter 6). Together, this work demonstrates that concerns about model design decisions are not purely academic, but have an important impact on model results, which in turn impact our understanding of language acquisition. Further, this work represents a potential path to address these problems in the computational modeling of cognition.

[2]The practice of using the output of one model in order to train a second model is commonly referred to as "downstream evaluation".

Chapter 2

Experimental Evidence in speech segmentation

One of the first tasks a learner must tackle in acquiring a language is recognizing where words[1] begin and end. This task, commonly referred to as speech or word segmentation, lays the foundation for future language learning. Without having first identified the words of a language, how might a child go about learning word meanings, parts of speech, or syntax? Proper segmentation underlies the acquisition of all of these fundamental linguistic elements.

Although to an adult ear word boundaries are quite obvious, it can be difficult to identify boundaries in a non-native language. This leads to the common, but incorrect, perception that these languages are spoken much more quickly than one's native language (Pellegrino et al., 2011). A further issue for literate speakers is that in many languages, segmentation is indicated orthographically through spaces, a notation which gives the illusion that the speech stream might likewise be separated through pauses. This is, however, not the case (Cole and

[1]Although researchers often speak of segmentation in terms of producing words, it should be noted that this need not be the case. Morphemes are a more meaningful linguistic unit in many languages and might instead be the goal of segmentation. In order to avoid confusion by using abstract terms such as "unit", the product of segmentation throughout this chapter will be referred to as words.

Jakimik, 1980). Words blur against one another and are better thought of as a stream of sound rather than as discrete, separable chunks. Further compounding the issue, although we often think of segmenting speech into words, there is no general consensus regarding what a word actually is (Di Sciullo and Williams, 1987).

Because learning to segment speech is non-trivial, researchers have attempted to discover how infants might begin to solve this task. In order to produce a model of segmentation, we must first understand the empirical facts which form the basis for decisions regarding model input, the model inference, and the model evaluation. In this chapter, I review the empirical foundations of segmentation by infant learners.

For speech segmentation, there are a number of difficult areas where questions must be answered:

1. **Input:** What information do infants consider in segmentation? How are these speech cues represented?
2. **Inference:** What cognitive constraints are placed on the learning process?
3. **Evaluation:** What should be the outcome of segmentation? What evidence can we compare our model against?

2.1 Input to speech segmentation

Infants make use of many different cues in order to segment words in their native language. They use these cues both in isolation and in combination to form an increasingly accurate picture of what makes a word in their language. Given the finding that pauses only inconsistently mark word boundaries (Cole and Jakimik, 1980), much work has been done to investigate other cues that infants might use to segment words. These cues can be split into

those which are *language-dependent* (i.e., their use varies from language to language) and those which are *language-independent* (i.e., their use does not vary across languages).

2.1.1 Language-Dependent Cues

2.1.1.1 Stress Patterns

In many languages of the world, individual lexical items possess one or more syllables that are acoustically prominent relative to others. This feature is generally known as *stress*, although the specifics of what constitutes acoustic prominence vary from language to language (Lehiste, 1976; Dauer, 1983; Dogil and Williams, 1999). There are two general types of stress: stress can be placed on syllables inside of a larger word, or stress can be placed on words inside of a larger sentence. The former is known as *lexical* stress, and although not every language possesses it, almost all developmental work on stress focuses on this particular type (Jusczyk et al., 1993, 1999b; Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003; Thiessan and Saffran, 2007). Stress can come in varying degrees. Syllables with the largest stress are marked as having *primary* stress, while lesser degrees of stress are marked as either *secondary* or *tertiary*[2].

Languages that do possess lexical stress can be further split into two categories. Some languages, such as Hungarian, possess what is known as *fixed* stress, meaning that a particular syllable of all words is consistently given primary stress. For example, in Hungarian any word pronounced on its own will always have its first syllable stressed. Other languages, such as English, have *variable* stress. While every word has a unique syllable that is given primary stress, the exact syllable varies from word to word. For instance, the English word ***à****pple* receives stress on its first syllable. The word *ba****nà****na*, however, receives stress on its

[2]Although linguists generally distinguish between primary, secondary, and tertiary stress, psychologists use the term "stress" to refer to primary stress in almost all cases.

second. Still, languages with variable stress are not without patterns. For example, English tends towards word-initial stress, particularly in nouns[3]. A child learning a language with lexical stress can then identify the general pattern of primary stress in the language and begin using that information to segment out words from fluent speech.

Because languages differ in their stress patterns, infants must first identify enough words so that they can learn the stress pattern in the first place. Once an infant recognizes a pattern, its application would be quite simple. For example, if an English-learning infant recognizes that English words often begin with a stressed syllable, she might reasonably segment the sentence "The**dòg**gie**àte**theba**nà**na" as "The **dòg**gie **àte**theba **nà**na". While this strategy often identifies words correctly (e.g. The **dòg**gie), segmentation often fails (e.g. **àte**theba **nà**na).

A number of studies have shown that infants take advantage of stress-based cues to speech segmentation as early as 7.5 months (Jusczyk et al., 1999b). In a series of experiments, Jusczyk and colleagues presented infants with sentences of ambiguous segmentation. The passage could be segmented to create words with initial or with final stress. English-learning infants preferred to segment word-initially stressed words, indicating a knowledge of the general stress patterns in English. This preference to match dominant stress patterns is strong enough that infants at this age will missegment actual English words that take word-final stress (e.g. gui**tàr**). By 10.5 months infants have started avoiding these mistakes. Confirming that infants have learned the general stress pattern and are not simply memorizing word pronunciations, Jusczyk et al. (1993) showed that infants recognize dominant stress patterns even for low-pass filtered speech, where only prosodic information is available.

[3]Pearl et al. (2011) and Phillips and Pearl (2015) report word-initial stress for bisyllabic words in CDS corpora as 89% and 89.9%, respectively.

2.1.1.2 Phonotactics

Phonotactics is the general term used to describe the ways in which sounds can be combined to form words in a language. In particular, it identifies what are valid sound sequences in a language and what are not. For instance, the nonsense word *splont* (/splant/) is a possible English word which happens to have no actual meaning. It's a possible word because it follows the phonotactic rules of English. The nonsense word *stwont* (/stwant/), on the other hand, does not follow these rules and therefore is not a possible (native) English word.

Importantly, specific sequences of sounds may be prohibited only in certain locations. For instance, /stw/ does appear in English, but only in cases where the cluster is split by a syllable boundary as in the word *eastward* (/ˈist.wərd/)[4]. If infants are able to recognize that a cluster such as /stw/ can never appear at the beginning or end of a word, then they might leverage that in identifying word boundaries.

This specific type of strategy was investigated by Mattys et al. (1999). They presented 9-month-olds with bisyllabic nonwords where each syllable was made up of a consonant-vowel-consonant (CVC) sequence. The stimuli were carefully designed such that the medial C.C cluster either had high probability of appearing across a word boundary and low probability of appearing word-internally, or the cluster had low probability of appearing across a word boundary but high probability of appearing word-internally. For example, the sequence /mk/ as in *nomkuth* ([ˈnɔm.kʌθ]) was likely to occur across a word boundary, but unlike to occur word-internally. On the other hand, the sequence /ŋk/, as in *nongkuth* ([ˈnɔŋ.kʌθ]), represents a second set of stimuli being much more likely to occur word-internally, while only rarely occurring across a word boundary. Mattys et al. (1999) found that infants were capable of making use of the phonotactic cues, but that their relative importance was quite

[4]In the International Phonetic Alphabet (IPA), syllable boundaries are marked by a period. Stressed syllables are preceded by the symbol /ˈ/ with this taking precedence over the period notation when both would otherwise occur.

weak in comparison to stress-based cues. When the two types of cues were made to conflict, infants strongly preferred to use the stress-based cue.

2.1.1.3 Allophonic Variation

Another potential cue to speech segmentation is what is known as *allophonic variation*. Allophones are the varying pronunciations for a single sound. For instance, the sound /t/ can be realized phonetically in a number of different ways. Word-initially it appears as [t^h] with aspiration as in *tart*. When it follows /s/ at the beginning of a syllable it appears as [t] without aspiration as in the word *start*. A third, common pronunciation in North American English occurs word-medially as [ɾ] as in the word *butter*. If a learner recognizes that certain allophones only occur in a specific position within a word, they might take advantage of that fact to identify word boundaries.

Jusczyk et al. (1999a) showed that 10.5-month-olds, but not 9-month-olds are capable of using these allophonic cues in order to segment out familiarized words. In this experiment, 9-month-old and 10.5-month-old infants were tested on their ability to discriminate between two phrases that varied in their phonetic form due to allophonic variation. An example of this is the phrase *night rates* versus *nitrates*. In particular, the /t/ in *night rates* is unreleased and unaspirated, while the same phoneme in *nitrates* is both released and aspirated. The /r/ in *night rates* contains no frication, while the same phoneme in *nitrates* does (Jusczyk et al., 1999a). Although infants as young as two months are able to discriminate between these sounds (Hohne and Jusczyk, 1994), they found that only at 10.5 months were English-learning infants able to leverage these cues to segment words. Crucially, the 9-month-olds inability to discriminate was not due to external factors such as processing limitations, implying that infants at this age may not yet have identified how to make use of the cues.

2.1.1.4 Coarticulation

Phonetic coarticulation is another cue which infants may use in order to segment their native language. Whenever two sounds are produced one after another, there is some degree to which the articulation of one sound influences the articulation of the other. This process is called coarticulation, and is well documented in the phonetic literature (Ladefoged, 1993). Evidence suggests that coarticulation is weaker when two sounds appear across a word boundary compared to when the same sounds appear within a single word (Cole et al., 1978; Fougeron and Keating, 1996). Similarly, Fujimura (1990) argues that both syllable- and word-initial consonants are produced more forcefully than consonants in other locations. In English, Pierrehumbert and Talkin (1992) found evidence for this kind of articulatory strengthening for [h] phrase-initially (e.g. *hogfarmers* vs. *mahogany*). Further, consonants are generally produced stronger word-initially than word-finally (Cooper, 1991).

Based on this type of evidence, Johnson and Jusczyk (2001) attempted to determine whether infants might be able to use coarticulation cues to segment, and if so how does coarticulation interact with other strategies such as TPs. They replicated the results of Saffran et al. (1996), except infants were exposed to naturally, rather than artificially, produced speech. Further, they created a stream of speech where coarticulation cues indicated one segmentation, while statistical cues favored an alternative segmentation. They found that 8-month-old infants preferred to use coarticulation cues over statistical cues, suggesting that coarticulation cues may be used around the same time as stress-based cues.

2.1.2 Language-independent Cues to speech segmentation

2.1.2.1 Statistical Cues

Because all of the above cues vary in their use across languages, it has been an open question as to how an infant, not knowing the particulars of their language, might determine the correct language-specific cues while learning to ignore inappropriate cues. In 1996, however, the first experimental evidence surfaced that infants as young as 8 months might be able to track *statistical information* in their environment (Saffran et al., 1996). This study became a hallmark in the language acquisition literature because it showed infants are much more sensitive to the language around them than previously thought.

In particular, Saffran and colleagues created a simple artificial language made up of four nonce words, each three syllables in length (e.g. [bi.da.ku]). Stimuli was produced by combining words in an order such that the likelihood of hearing any two syllable combination was controlled. The probability of hearing one syllable after another is known as *transitional probability* (TP). Within words, transitional probabilities were held at 100%, such that if the infant heard [bi] there was a 100% chance that the next syllable would be [da]. After the final syllable of a word, however, the transitional probabilities were much lower, held at 33%. Because the stimuli were produced with a voice synthesizer, all other acoustic qualities (e.g. coarticulation, stress, duration) could be controlled for. They found that after only 2 minutes of presentation, infants were able to pick out the actual words at a higher rate than would have been expected by chance.

This finding spurred an immense amount of work to try and discover the uses and limits of infant statistical learning (Aslin et al., 1998; Johnson and Jusczyk, 2001; Maye et al., 2002; Thiessen and Saffran, 2003; Maye et al., 2008; Xu and Garcia, 2009; Xu and Denison, 2009; Pelucchi et al., 2009; Hay et al., 2011; Denison et al., 2013). In the context of speech

segmentation, researchers hoped that statistical cues might explain how infants begin to segment before they've acquired the knowledge necessary for language-specific cues. An important property of statistical learning in this context is that while the particulars of any language may vary quite a bit, the way in which statistical learning proceeds varies not at all. For example, if a learner tracks transitional probabilities in English, they may find that certain sequences are impossible (e.g. [bʏ]) while others are common (e.g. [bæ]). In German, however, the opposite might be true, [bʏ] might be frequent while [bæ] is never encountered. The probabilities vary quite a lot between languages, but the way in which they are used remains identical: units that co-occur frequently are more likely to exist within a single word than sequences which do not.

2.1.3 Timeline of Segmentation

The first behavioral evidence for speech segmentation in infants comes at six months (Bortfeld et al., 2005). At this age, infants have a small set of very frequent words which forms the basis of their proto-lexicon. They recognize these words in speech, making the segmentation of neighboring units easier. Unfortunately, more is not known regarding the way in which infants segment at six months.

It's only by around the age of seven or eight months that infants first show a more robust segmentation ability. At this point, we have the first evidence that children are segmenting words based on their stress patterns (Jusczyk et al., 1999b). At the same time, statistical cues are being utilized as well (Saffran et al., 1996; Thiessen and Saffran, 2003). Neither of these cues, however, can be used to properly segment all words in a language. Stress-based cues often don't apply to all items, and words may lack any kind of stress in fluent sentences. For example, words in English tend to have word-initial stress, but there are many counterexamples (e.g. *ba**nà**na*, *to**gè**ther*, *com**plète***). Likewise, statistical cues can be used

generally to identify words, but any individual word may not be properly segmented. For example, words which frequently occur together may be best treated, statistically, as a single unit (i.e. *who's that* segmented as *who'sthat*). This can also lead to conflicts between cues, where stress may indicate one segmentation, while transitional probabilities would suggest another (e.g. statistics might correctly segment that the phrase *gui**tàr** is*, while the stress cues indicate the segmentation should be *gui **tàr**is*.

During the time period from seven to nine months, infants are learning how to integrate these various cues (Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003). Johnson and Jusczyk (2001) found that when 8-month-olds are presented with a sequence of syllables with conflicting stress and transitional probability cues, they prefer to segment using the stress-based cues. Thiessen and Saffran (2003) investigated this further and found that while older infants (8- to 9-month olds) prefer to segment according to their language's predominant stress pattern, younger infants (6.5- to 7-month olds) ignore stress-based cues and segment according to the transitional probabilities alone.

Beyond nine months, infants begin making use of many language-specific cues. These include phonotactics (Mattys et al., 1999), allophonic variation (Jusczyk et al., 1999a), and coarticulation (Johnson and Jusczyk, 2001). These results appear to indicate that infants begin segmentation largely with cues that are independent of the structure of any given language (i.e. statistical cues, utilizing familiar words). It is only later that they begin to acquire the knowledge necessary to make use of language-specific cues such as stress patterns. It takes even longer before infants are able to make use of more detailed, language-specific phonetic cues in order to segment.

2.1.4 Unit of Representation

Given that speech segmentation begins around 6 months of age (Bortfeld et al., 2005) and that statistical segmentation appears to wane in importance around 9 months of age (Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003), it is necessary to identify how infants of this age actually perceive the speech around them. The literature broadly splits into arguments either for a segmental representation (e.g. phonemes) or for a non-segmental representation (e.g. syllables). This issue, in both adults and infants, has been of some consternation for researchers of speech perception for at least the past fifty years (Liberman et al., 1967; Pisoni, 1972; Massaro, 1972, 1974, 1975; Mehler et al., 1981; Jusczyk, 1997; Goldinger and Azuma, 2003). For the purposes of investigating infant speech segmentation, the exact nature of adult perception is somewhat irrelevant. Infants need not perceive speech in the same manner as adults, and therefore I examine only the evidence from young infants.

A number of researchers have investigated the ability of young infants to discriminate words based on the properties of these words' internal structure (both syllabic and phonemic). The goal of this research is to determine how well infants are able to perceive and make use of speech in terms of its individual segments (i.e. phones or phonemes) and its non-segmental units (i.e. syllables). Looking at two-month-olds, Jusczyk et al. (1995) trained infants to recognize a set of syllables which either shared a common segment (e.g. [bi], [ba], [bu]) or did not (e.g. [si], [ba], [tu]). After a two minute delay, the infants were able to recognize the trained syllables, but accuracy did not depend on the internal structure of the syllables. This suggests that infants do not perceive the segmental similarities between syllables (e.g. a shared initial consonant), but rather treat each syllable as an atomic unit.

This interpretation is strengthened by additional experiments in the literature (Jusczyk and Derrah, 1987; Bertonicini et al., 1988) which also find that similarities between syllables do not aid infants in categorization. Perhaps the most direct interpretation of these results is

that two- to three-month-old infants perceive speech as a string of syllables, each treated as an atomic unit. On the other hand, it is also possible that infants do perceive speech segmentally, but that for one reason or another infants lack the ability either to recognize segmental similarities or to process these similarities in a way which might be reflected in their behavior.

If infants perceive speech in terms of segments or syllables, one might expect that they would be able to differentiate words based on the number of segments/syllables they possess. Bijeljac-Babic et al. (1993) presented newborns with utterances with either a shared number of segments (4 vs. 6) or syllables (2 vs. 3) and examined whether or not infants noticed when stimuli shifted from one number of segments/syllables to another. Infants were surprised when bi-syllabic utterances were replaced with tri-syllabic utterances (or vice versa), indicating that infants were sensitive to the differences of syllable length. This finding remained even when the stimuli were of similar duration. Infants were also presented with bi-syllabic utterances made up of either 4 or 6 segments[5]. When bisyllabic utterances with 4 segments were replaced with bisyllabic utterances with 6 segments (or vice versa), infants did not change their response, indicating that they did not recognize a systematic difference between the two sets of stimuli.

Evidence for either representation can also be found by looking for categorization based on shared initial or final features. Jusczyk et al. (1995) found that infants have more difficulty recognizing the difference between two bisyllabic words that share an initial syllable (e.g. /ba.lo/ vs. /ba.nʌl/) as opposed to words which do not share an initial syllable (e.g. /ba.lo/ vs. /pa.mʌl/). To rule out the possibility that infants are tracking the repetition of two phonetic segments rather than syllables, they also tested infants on sequences that crossed

[5]To ensure infants detected a change in the number of segments, rather than a particular ordering of consonants and vowels, a wide range of syllabic structures were utilized. 4 segment items were of the form CVCV, VCCV, CCVV, VCVC, CVVC, or VVCC. 6 segment items were of the form CVCCCV, CCVCCV, VCCCCV, CVCCVC, CVCVCC, CCVCVC, CCVVCC, VCCCVC, or VCCVCC. Consonat clusters were chosen to conform with the phonotactics of the infants' target language, French.

syllable boundaries. For instance, when presented with stimuli that had the sequence /a.b/ (e.g. /la.bo/, /za.bi/), infants were unable to show the same effect, responding similarly to stimuli with the shared sequence (/na.bʌl/) and stimuli which lacked the shared sequence (/ba.nʌl/). These results favor the syllabic hypothesis for 2- to 3-month-olds.

A similar experiment by Eimas (1999) found evidence for categorical representation of bisyllabic words with a shared initial syllable (e.g. /ba.pi/, /ba.tad/). For bisyllabic words that shared final syllables (e.g. /pi.ba/, /tad.ba/), a weaker effect was found. When investigating the effect for utterances with shared initial phonemes (e.g. /bɪd/, /bæd/), they were unable to find evidence for categorization. The combined result of this work with 2- to 4-month-olds indicates that infants at this age do not find shared segmental information salient, although they are much more likely to recognize similarities among syllables. In his review of the subject, Jusczyk (1997) notes that "there is no indication that infants under six months of age represent utterances as strings of phonetic segments".

Evidence for segmental representations in infants is generally found in infants older than 6 months. One line of evidence for this is based around what sounds an infant is able to discriminate. Young infants are able to recognize the phonetic difference between all speech sounds (Werker and Tees, 1984; Werker and Lalonde, 1988; Best et al., 1988; Kuhl et al., 1992) with some exceptions (Aslin et al., 1981; Polka et al., 2001). After six months, infants begin to ignore sounds which do not appear in their native language, losing the ability to distinguish them from similar sounds. This shift begins around 6 months with vowels (Kuhl et al., 1992; Polka and Werker, 1994), with consonants being lost slightly later, between 8 and 12 months (Werker and Tees, 1984; Werker and Lalonde, 1988; Best et al., 1988, 1995). This implies that infants are only beginning to identify meaningful sounds in their native language at 6 months and all sounds in the language are identified somewhere around 12 months.

There is also evidence that infants at this age have some representation of abstract phonetic features. Maye et al. (2008) were interested in the ability of 8-month-olds to recognize how sounds are distributed. They were able to train infants on a non-native voice-onset time (VOT) contrast by exposing them to two sounds in a bimodal distribution. After exposure to the bimodel distribution, infants deduced that there were two separate phonetic segments. Interestingly, although the infants were trained on the VOT contrast using only dental sounds, they generalized the contrast to bilabial sounds as well, indicating that they recognized the importance of VOT in the contrast.

Still, many of these findings, both with younger and older infants, can be interpreted both with a segmental and a non-segmental hypothesis. What can be said, however, is that as infants age, they are better able to learn information about segments. Children at all ages, in contrast, appear relatively comfortable with syllables or other non-segmental units. At the time that segmentation is just beginning, around 6 months, infants might still be representing speech as a string of syllables. If infants perceive speech segmentally, they have still not learned the full set of sounds in their language, meaning they must represent speech phonetically and not phonemically. A safe assumption, therefore, would be that infants represent speech using syllables, especially based on evidence that they track syllables in order to segment (Saffran et al., 1996; Aslin et al., 1998).

2.2 Inference for Speech Segmentation

One of the facts that makes child language acquisition so impressive is that it is done in the face of the many cognitive constraints placed by the developing brain. One method for investigating optimal solutions to problems in statistical learning is Bayesian modeling. The Bayesian approach updates a learner's beliefs by explicitly modeling prior assumptions (a learner's *prior*) as well as how they react to incoming data (a learner's *likelihood* function).

While Bayesian modeling is a tool for discovering optimal solutions, the fact is that children have to solve problems without the luxury of unlimited memory and time. Faced with this fact, Bayesian cognitive modelers are increasingly incorporating cognitive considerations into their models (Anderson, 1990; Shi et al., 2010; Bonawitz et al., 2011; Pearl et al., 2011; Griffiths et al., 2015).

A Bayesian model defines a set of equations which reflect the assumptions of the learner. These equations generally do not have a simple exact solution and instead rely on statistical inference methods such as Markov chain Monte Carlo[6] in order to produce an approximation of the exact answer. Cognitive constraints can then be built into this learning process resulting in less "optimal" solutions. This separates the model itself from the inference process, allowing their roles to be independently investigated.

Despite increasing interest in cognitively-constrained modeling, there remains little experimental evidence to suggest exactly what kinds of constraints should be imposed and in what manner. Because the true inference process used by children (and/or adults) is unknown, a wide variety of procedures can be investigated in order to better understand the role of the inference process. Of the possible constraints which might be added, I focus on three in particular which have been incorporated into past models:

1. Online processing: Data should be processed as it is encountered.
2. Non-optimal decision making: Inference does not always choose a locally optimal choice.
3. Memory constraints: Inference may focus on recently encountered data.

It is generally well accepted that learning, both in infants and adults, proceeds in an online fashion, with information being processed as it is encountered. Harder modeling decisions,

[6]More information about Markov chain Monte Carlo (MCMC) can be found in Section 3.2.2.3.

however, have to be made in terms of both incorporating non-optimal decisions and in implementing memory constraints. Optimal behavior in many probabilistic domains involves choosing the highest probability event, but this is inconsistent with experimental evidence from infants and children (Köpcke, 1998; Kam and Newport, 2005, 2009; Davis et al., 2011; Denison et al., 2013). For example, imagine there is a bag filled with marbles, of which 75% are red and 25% are blue. If the goal is to predict the color of the next marble drawn from the bag, the optimal decision would be to always choose red. A sub-optimal decision would be to choose red 75% of the time and blue 25% of the time, an option known as *probability matching*. In some cases, infants appear to probability match (Davis et al., 2011; Denison et al., 2013) and in others they appear to generalize, but not always to the highest probability outcome (Köpcke, 1998; Kam and Newport, 2005, 2009).

In terms of memory, there is a long-standing hypothesis that limited memory actually aids in the child learner in acquiring a native language (the "Less is More" hypothesis: Newport, 1990). Memory effects certainly appear to play some role in infant learning (Kam and Chang, 2009; Perfors, 2011), and ignoring the role of memory may lead to model performance deviating from actual learners.

Although there is a lack of definitive experimental evidence regarding cognitive constraints on learning, what is clear is that some constraints should be expected. By modeling the impact of various constraints, the researcher can better understand what processes underly the model's success or failure. A detailed investigation of the differences in model behavior between various inference schemes is necessary in order to understand which process best matches the infant behavior detailed in Section 2.1.

2.3 Model Evaluation for Speech Segmentation

In order to properly determine whether or not a model was successful, it is important to evaluate the model's output against all relevant experimental and observational data. This ensures that the model conforms with everything which is known about the way in which a child solves the same task. For example, if infants are known to solve a particular task (e.g. the statistical learning demonstrated in Saffran et al., 1996), then the model should likewise be shown to pass the same task. Not all comparisons are experimental in nature. Observational data is also useful in many cases. For instance, it is well-known that any infant is capable of learning any of the world's languages. Because of this, a good model of acquisition should operate regardless of the language being learned.

2.3.1 Experimental Comparisons

There are a number of experiments against which models of speech segmentation can be compared. For models of segmentation which do not take into account stress-cues, the number of experiments, however, is somewhat limited. Experimental studies demonstrate that infants are able to solve trivial segmentation problems when the only relevant cues are forwards TP (Saffran et al., 1996; Aslin et al., 1998) or backwards TP (Pelucchi et al., 2009), but these represent relatively simple hurdles. Nonetheless, a model which cannot account for these experimental findings is clearly deficient.

Another possible point of comparison comes from Bortfeld et al. (2005). In this study, it is demonstrated that infants are better able to segment words when they appear frequently next to an already learned word. For instance, if the word *mommy* is known, then an unknown word (e.g. *feet*) will be more easily learned if it occurs next to *mommy* as in the sentences *The girl laughed at mommy's feet* and *Mommy's feet were different sizes.* If *feet* occurs

equally often, but in non-adjacent positions to the known word, then it will be learned less well. We should expect then that a good model of segmentation should correctly segment words which neighbor these kinds of familiar words more often than it correctly segments words which are not commonly neighbors.

Seidl and Johnson (2006) demonstrate that infants are also better at segmenting words when they are either utterance-initial or utterance-final. Interestingly, they find that infants are equally good at segmenting these words regardless of whether they appear at the beginning or end of sentences. This suggests that not only should a good model of segmentation be more accurate on utterance-initial and utterance-final words, but that the benefit should be equal across each type. Such an analysis was performed by Pearl et al. (2011) and provides a useful point of comparison with experimental data.

2.3.2 Observational Comparisons

One of the most commonly assumed outcomes of language learning is that eventually children should achieve adult competence. If this were not the case, then language as we know it could not continue to exist. In the same vein, it is well known that this should not be true of just some subset of languages, but must, in fact, be true of all human languages. Therefore, if our goal is to compare a model's performance against infants', then the model should be tested on a range of languages to ensure not only that adult competence can be achieved, but that it can be done on any human language.

Another point of comparison for models of speech segmentation relates to the kinds of errors that children produce. While there is some experimental evidence related to particular segmentation errors, these typically relate to how children make use of stress cues (e.g. *gui**tàr** is* segmented as *gui **tàr**is,* Jusczyk et al., 1999b). Another way to detect segmentation errors comes in the form of production errors in older children. Determining whether or not the

productions of a child represent a single unit or are decomposable is a challenging problem. Brown (1973) briefly discusses evidence from his own diary data, suggesting that there are "a great many unanalyzed 'chunks' . . . ". Brown derives a set of criteria from which this kind of judgment can be made, which were subsequently fleshed out in the work of Peters (1983). The work of both of these scholars suggests that children, even as old as three years, do frequently missegment the language around them. These errors can be widely split into two types, function word collocations (e.g. *that'sa*, *it'sa*) and function word oversegmentations (e.g. *a nother*, *be have*) and were used as an evaluation metric by Lignos (2012).

2.4 Conclusion

A wide range of experimental evidence lays the foundation for improving current models of speech segmentation. Experiments aimed at young infants demonstrate that at the time segmentation is just beginning infants have not learned the full phonology of their native language. This rules out the possibility of using a phoneme-based corpus to train and test a model of early segmentation. Syllables may make a better unit of representation given that infants use them for statistical learning, whereas infants at that age do not use phonetic segments in the same manner. Experimental evidence does not constrain the type of model inference used, but by examining a number of possible options, a better sense of model performance can be achieved. Lastly, experimental and observational results provide an alternative to gold standard evaluation for speech segmentation. In particular, segmentation error patterns may provide a useful qualitative benchmark to use in both understanding and evaluating model behavior.

Chapter 3

Previous Models of Speech Segmentation

Over the past twenty years there have been many learning strategies, implemented with various algorithms, proposed as methods infants might use to perform speech segmentation. These models make use of a variety of cues in order to achieve this task, but can be broadly categorized into three types:

- Phonotactic Models - Track phone-to-phone (or phoneme-to-phoneme) probabilities in order to discover boundaries.
- Bayesian Models - Leverage specific generative assumptions about how children view language in order to discover boundaries.
- Heuristic Models - Follow a set of predefined, heuristic steps in order to discover boundaries rather than attempting to find an optimal solution.

Because there is no evidence for early phonotactic segmentation (see Section 2.1.1.2), I focus only on the Bayesian and heuristic approaches describing important previous models and

discussing their strengths and weaknesses. I then go over how the work in Chapters 4 - 6 addresses the weaknesses of previous studies by making assumptions which better match experimental evidence. First, in order to compare the models, I describe the general methods of evaluation which have become standard in the field.

3.1 Gold Standard Model Evaluation

The vast majority of studies in speech segmentation report results in terms of a model's ability to recreate a gold standard segmented corpus. To compare a model's output with a gold standard, one first has to decide what unit to measure. Traditionally, researchers have presented results over three distinct items: 1) word tokens, 2) boundaries, and 3) lexical types.

For example, the sentence *The doggy is on the floor* might be segmented as *Thedoggy is onthe floor*. There are six word tokens (individual words) in the original sentence, but the model only identifies two of those tokens (*is* and *floor*). The same sentence has five word boundaries (excluding the utterance boundaries) and the model correctly identifies three of those (*thedoggy is*, *is onthe*, and *onthe floor*). Lastly, lexical types refers to the number of unique words. Although the original sentence has six word tokens, there are only five types because the word *the* appears twice. Of these five, the model only correctly recovers two types (*is* and *floor*).

For each of these measures, a comparison between the gold standard and the model output is typically made using the metrics of *precision* and *recall*. If we think of the problem as one of signal detection, then we can divide the model results into four categories as in Table 3.1.

Precision, also known as accuracy, is used to measure the percentage of units created by the model which were made correctly. High precision means that if the model believes there is a

		Response	
		Present	**Absent**
Stimulus	**Present**	Hit	Miss
	Absent	False Alarm	Correct Rejection

Table 3.1: Possible outcomes for a signal detection trial.

word token/word boundary/lexical item, it is likely to be correct. In signal detection terms, precision can then be defined as:

$$Precision = \frac{Hits}{Hits + False\ Alarms} \tag{3.1}$$

Recall, also known as completeness, is used to measure the percentage of true units which were correctly identified. High recall means that the model correctly identified most of the true word tokens/word boundaries/lexical items. Recall is defined as:

$$Recall = \frac{Hits}{Hits + Misses} \tag{3.2}$$

In many cases, it is trivial to create a model with very high precision (at the cost of low recall) or very high recall (at the cost of low precision). For example, consider a model which *always* inserts a word boundary whenever possible. Such a model will have very low boundary precision (because many of the boundaries it inserts are false) but 100% boundary recall (because it inserted a boundary whenever a true boundary existed). Likewise, consider a model that inserts a boundary only when it is absolutely necessary. The model would have very high precision (because all of the boundaries it inserts are true), but low recall (because most boundaries were not identified).

Because of these issues, the two scores are often aggregated into what is known as an *F-score*, the harmonic mean of precision and recall. F-score is defined as:

$$F\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.3}$$

All of precision, recall, and F-score lie on the range $[0, 1]$ with higher values indicating closer adherence to the gold standard. These values can also be reported as percentages rather than probabilities on the range $[0, 100]$, as they will be throughout this document.

Precision, recall, and F-score can be reported on each of the three measures described previously. Each set of measures provides a different perspective on model performance. Boundary scores are typically much higher than word token or type scores[1], but the relative performance of boundary precision and recall gives an indication of the general types of errors made by the model (Pearl et al., 2011). In particular, models with high boundary recall, but low precision, insert too many boundaries and therefore tend to *oversegment*. Models which have low recall, but high precision, insert too few boundaries and therefore *undersegment*. Because word types do not take into account frequency, they reward models that identify infrequent words more so than word tokens would.

[1]Correctly segmenting a word requires identifying two boundaries (excluding utterance-initial or final words) which makes word token identification a harder task than identifying single boundaries.

3.2 Bayesian Models

3.2.1 MBDP-1

One of the first Bayesian models of speech segmentation was the Model-Based Dynamic Programming (MBDP-1)[2] algorithm (Brent, 1999). The model assumes that each utterance in the corpus is made up of some sequence of words and each word is made up of a sequence of phonemes. As with all Bayesian models, MBDP-1 contains both a prior and likelihood term. The likelihood term takes on one of two values, 0 in the case that the sequence of words does not conform to the observed, unsegmented corpus, and 1 in the case that it does. In this way, only segmentations which match the observed data are considered. The posterior, then, is driven largely by the prior probability of the segmented corpus.

The prior assumes that the corpus was generated in the following fashion:

1. Generate the total number of word types in the lexicon
2. Generate a token frequency for each word type
3. Generate the phonemes that make up each word type (excepting the distinguished type, $ representing utterance boundaries)
4. Generate an ordering for the set of tokens

In order to learn the proper parameters for each of these steps, Brent (1999) developed an online algorithm to search the parameter space. The model performs reasonably well on an English corpus of child-directed speech (word token F-score: 68.2). MBDP-1 lacks, however,

[2]The 1 in MBDP-1 was added as a version number in the hopes that others might update the model to create an MBDP-2. Although this never came to be, the model is still referred to with its original version number.

in that modifications to the generative assumption are quite difficult. In theory, the model's prior, likelihood, and inference method are all separable, but in practice they are tied together such that changes to one would require changes across the entirety of the model. MBDP-1 makes a unigram assumption, that words appear independently of the words around them, and could in theory be adapted to have a bigram assumption, that a word is chosen based only on the word in front of it. In practice, such a change to the model would be quite difficult. A more detailed discussion of the peculiarities involved in the MBDP-1 model can be found in Goldwater et al. (2009). The model operates over phonemes, although it could reasonably be adapted to operate over syllables without changing the underlying model itself.

3.2.2 DPSEG

MBDP-1 builds off of two fundamental insights: 1) that frequent words should be preferred over infrequent and 2) that shorter words should be preferred over longer words. These parsimony biases form the basis for a more general Bayesian model of segmentation known as Dirichlet Process segmentation (DPSEG) (Goldwater et al., 2009), which is the primary model examined in the rest of this dissertation. The model represents a step forward from MBDP-1 in that it is more easily adapted both in terms of the generative assumptions and in the inference process used for learning. Because this model will be used repeatedly throughout the remaining chapters, I will discuss it in detail.

3.2.2.1 Unigram DPSEG

The simplest instantiation of DPSEG makes use of a unigram language model, assuming that every word is chosen independently of the words around it. In order to differentiate the unigram version of DPSEG, I will refer to it as DPSEG-1. As with MBDP-1, DPSEG-1 assumes that the role of the likelihood function is only to rule out segmentations which do

not match the observed corpus, while the majority of work is done by the prior probability. Essentially, this assumes that learners have expectations as to what the language they encounter should look like. The unigram model assumes that for any utterance, each word w_i is generated by first deciding whether it is a novel lexical item:

1. If w_i is a novel lexical item, generate a phonemic form $(x_1 \dots x_M)$ for w_i.
2. If w_i is not a novel lexical item, choose an existing form l for w_i.

Note that the model does not rule out the possibility of homophones, different words with the same pronunciation (e.g. *two* and *too*). Therefore even if a word has been previously encountered, that does not mean it cannot be a novel lexical item. In order to determine the probability of a word, the model combines these two possibilities into Equation 3.4:

$$P(w_i = l|w_{-i}) = \frac{n_l + \alpha P_0(w_i = l)}{i - 1 + \alpha} \tag{3.4}$$

$$P_0(w_i = x_1 \dots x_M) = p_\#(1 - p_\#)^{M-1} \prod_{j=1}^{M} P(x_j) \tag{3.5}$$

The parameter α is a positive real number, n_l is the number of times the lexical item l appears in the first $i - 1$ words, $p_\#$ is the probability of generating a word boundary, and $w_{-1} = w_1 \dots w_{i-1}$. This process represents what is known as a *Dirichlet process*, from which the model is named (Ferguson, 1973). A Dirichlet process (DP) is a non-parametric stochastic process which results in a probability distribution that is often used in Bayesian modeling as a prior. The DP takes two parameters, the first of which is the concentration parameter α and

the second is a base distribution P_0. In this case the concentration parameter determines how many word types are expected (higher values of α indicate a preference for greater numbers of types) while the base distribution determines the probability of any novel lexical item.

It should be noted that the model as described is somewhat incomplete in that it fails to model utterance boundaries. The model assumes that the speaker chooses to end an utterance with some probability, $p_\$$. Whenever a word is generated, there is some probability u_i that an utterance will be drawn. Because $p_\$$ is unknown, we can calculate the value of u_i by integrating over the possible values of $p_\$$. It is assumed $p_\$$ is generated from a symmetric Beta($\frac{\rho}{2}$) prior. Although I do not give the derivation here, this results in a standard form (Gelman et al., 2004):

$$P(u_i = 1 | u_{-i}, \rho) = \frac{n_\$ + \frac{\rho}{2}}{i - 1 + \rho} \tag{3.6}$$

where $n_\$$ is the number of utterance-final words encountered in the first $i - 1$ words of the corpus and ρ is a free parameter of the model. Goldwater et al. (2009) explored the impact of different values of ρ and found that its value did not impact segmentation performance. I therefore follow their decision to treat ρ as a fixed parameter with value of 2.

3.2.2.2 Bigram DPSEG

The DPSEG-1 model can be expanded to include a bigram language model (word are chosen depending on the immediately previous word) by making use of a heirarchical Dirichlet process (HDP, Teh et al., 2006). As with the unigram model, the bigram model (DPSEG-2) makes assumptions about how sentences are generated. Instead of generating words and

then phonemic forms, DPSEG-2 generates bigrams, then words, then phonemic forms in the following manner:

1. If the pair $< w_{i-1}, w_i >$ is a novel bigram:

   If w_i is a novel lexical item, generate a phonemic form $(x_1 \ldots x_M)$ for w_i.

   if w_i is not a novel lexical item, choose an existing form l for w_i.

2. If the pair $< w_{i-1}, w_i >$ is not a novel bigram:

   Choose an existing form l for w_i from those that have been previously generated after w_{i-1}.

As with DPSEG-1, the probability associated with each step must be defined. The bigram model is defined with the following equations:

$$P_2(w_i|w_{-i}) = \frac{n_{<w_{i-1},w_i>} + \beta P_1(w_i|w_{-i})}{n_{w_{i-1}} + \beta} \tag{3.7}$$

$$P_1(w_i|w_{-i}) = \frac{t_{w_i} + \gamma P_0'(w_i)}{t + \gamma} \tag{3.8}$$

$$P_0'(w_i = x_1 \ldots x_M) = \begin{cases} p_\$ & \text{if } w_i = \$ \\ P_0(w_i = x_1 \ldots x_M) & \text{if } w_i \neq \$ \end{cases} \tag{3.9}$$

The bigram model differs somewhat from the unigram model. First, equation 3.7 defines the probability of the bigram $< w_{i-1}, w_i >$. This equation describes a DP with concentration parameter β and base distribution P_1. In this case, the base distribution is generating individual lexical items and the parameter β controls how often the model expects to encounter novel bigrams (higher values of β indicate the model expects to see many different bigrams). So, for speech segmentation a learner with a higher β would rate more probable a pair of words which had not been seen before as opposed to the same learner with a lower β.

Equation 3.8 is similar to the unigram equation 3.4 in that it generates lexical items. Again we have a base distribution P_0' and a concentration parameter γ, but instead of using the number of tokens encountered n_l, the bigram equation makes use of t_{w_i}, which represents the number of times a bigram with w_i has been newly generated. t represents the total number of times any bigram was newly generated.

Lastly, because of the complication of utterance boundaries, it is useful to think of \$ as a separate word type. In generating novel lexical forms, the bigram model then makes use of equation 3.9. If w_i is an utterance boundary, then it is generated with probability $p_\$$. If not, then the word is generated as in P_0 from the unigram model and multiplied by $1 - p_\$$, the probability of not generating an utterance boundary.

3.2.2.3 DPSEG Inference

Batch Inference

The original inference algorithm used for DPSEG was a batch process known as Gibbs sampling (Geman and Geman, 1984). Gibbs sampling requires the parameters of a model to first be initialized, typically through some random process. Then, each parameter is updated conditioned on the current value of all other parameters. This is done for each parameter and then repeated for a number of iterations until convergence is achieved.

In the case of DPSEG, each possible boundary location is a parameter which takes either value 0 or 1 indicating either the lack or existence of a boundary. Boundaries are initialized in a random fashion and then the sampler goes through each boundary location in the corpus deciding based on the locations of other boundaries, whether or not to place a boundary at the location in question. This is repeated for s iterations. For all batch learners presented in later simulations, s was set to 20000, as in Goldwater et al. (2009), which is generally sufficient for convergence of the Gibbs sampler.

Gibbs sampling requires knowing the probability of any boundary given all the others in the corpus. This essentially reduces down to a choice between creating a single word out of the nearby phonemes (H_0), or to create one word from the phonemes before the boundary location and a second word from the phonemes after (H_1). For example, in segmenting the utterance *go play*, H_0 would represent the possibility of a single word *goplay* while H_1 would represent the possibility of the two words *go* and *play*. Because these are the only two possible options, the probability of inserting a boundary (H_1) can then be defined as:

$$P(H_1) = \frac{P(H_1)}{P(H_1) + P(H_0)} \tag{3.10}$$

The probability of H_0 and H_1 is defined by the generative models described previously. If no boundary is placed, then only a single word must be generated. If a boundary is placed, then two words must be generated. On the one hand, the model prefers H_0 because it has fewer words, but on the other it prefers H_1 because this avoids having a longer word. The exact trade-off depends on the model parameters and, most importantly, on the frequency with which each word (and possibly bigram) has been observed.

This process of choosing H_0 or H_1 is repeated for each boundary position. Once every boundary in the corpus has been sampled, that constitutes a single iteration, which is then repeated $s = 20000$ times.

Online Optimal Inference

Pearl et al. (2011) describe a number of alternative online inference methods for the DPSEG model. The first of these is based off of the MBDP-1 model and Pearl et al. (2011) therefore refer it as *Dynamic Programming Maximization*. In order to better highlight the particular cognitive constraints that each online learner possesses, I will refer to this inference method as the *OnlineOpt* learner highlighting that it is both online and makes optimal choices under uncertainty.

The OnlineOpt learner operates by looking at each utterance in turn and using the Viterbi algorithm (Viterbi, 1967) in order to calculate the highest probability segmentation. The Viterbi algorithm is a classic example of what is known as dynamic programming. A dynamic program takes a complex task and decomposes it into smaller subproblems, each of which only needs to be solved once. In the context of segmentation, the Viterbi algorithm takes the larger problem of determining an optimal segmentation and splits it into deciding at each possible boundary, what the optimal segmentation would be.

For example, let $W_{1...i}$ represent the optimal segmentation for an utterance from the beginning of the utterance until current index i. If $W_{1...i-1}$ is already solved, then solving for $W_{1...i}$ involves identifying whether inserting a boundary at position i is more probable than not. Making this calculation is straightforward and allows the model to quickly arrive at the optimal segmentation. Once the utterance is segmented, the words produced are added to the lexicon and the token counts of each word are updated accordingly[3].

[3]Note that because each utterance is segmented in turn, the OnlineOpt learner does not technically adhere to the model used by the BatchOpt learner, which requires that word counts be updated on a word-by-word basis. This same issue holds for the OnlineSubOpt learner as well, as described below.

Online Suboptimal Inference

An alternative to the OnlineOpt learner is the *OnlineSubOpt* learner (refered to by Pearl et al. (2011) as the *Dynamic Programming with Sampling* learner). This learner does not always choose the highest probability segmentation, and instead chooses segmentations in proportion to their probabilities. The probability of each possible segmentation is calculated using the Forward algorithm (part of the Forward-Backward algorithm) (Baum, 1972).

The Forward-Backward algorithm is similar to the Viterbi algorithm in that it is a dynamic programming method typically used on problems such as hidden Markov models. The model operates in two phases, a Forward portion which goes forwards through the path, and a Backward portion which goes backwards through the path. Once finished, the model has calculated the posterior marginal of all hidden state variables.

The Forward algorithm calculates the probability of observing a given sequence up to an index i, similar to the Viterbi algorithm (although the Viterbi algorithm calculates this value only for the maximally probable segmentation). The Forward algorithm makes this same calculation for every possible segmentation of the utterance up to index i and then calculates the probability of the possible segmentations for index $i + 1$. Because this solves for the probabilities that the DPSEG model cares about, the Backward portion of the full algorithm is not necessary.

As noted previously, once the model has calculated the probability of all possible segmentations for the utterance, the resulting probability distribution over segmentations is used to choose a single segmentation. This style of weighting options based on their probabilities is non-optimal, since a rational learner would always choose the highest probability segmentation. Infants, and adults, however, appear in some cases to not be optimal learners, a fact which the OnlineSubOpt learner attempts to replicate (Denison et al., 2013).

Online (Decayed) Memory Inference

The last learner investigated by Pearl et al. (2011) attempts to focus processing resources on recently encountered items. This is done using what is known as a Decayed Markov Chain Monte Carlo (DMCMC) method (Marthi et al., 2002). I will refer to this learner as the *OnlineMem* learner because it performs inference online and with a form of short-term memory. As with the BatchOpt algorithm, this learner samples individual boundary locations and updates them conditioned on the value of all other previously encountered boundaries. In Gibbs sampling, this would be done once per iteration for every possible boundary. For the OnlineMem process, locations to sample are chosen based on a decaying function from the current location. The model stops at the end of every utterance and chooses a potential boundary b_a to sample, where the probability of choosing that boundary given that there are a boundaries between b_a and the end of the current utterance is given by the formula:

$$P(b_a) = \frac{a^{-d}}{\sum a_i^{-d}} \tag{3.11}$$

where d is a parameter of the model. The larger the value of d the stronger the recency effect of the model. For all simulations, I choose $d = 1.5$ which represents a strong recency bias. For example, on the Brent UCI Syllables corpus (Phillips and Pearl, 2015; MacWhinney, 2000) this corresponds to 83.6% of sampled boundaries occurring in the current utterance, 11.8% in the previous utterance, and only 4.6% located in any other utterance. This aspect of the model implements a form of memory which matches the fact that adults frequently show recency biases in experimental settings (Murdock Jr, 1962; Baddeley and Hitch, 1993; Davelaar et al., 2005).

3.2.3 Adaptor Grammars

Just as the DPSEG algorithm expands upon the capabilities of the MBDP-1 model, *adaptor grammars* (Johnson et al., 2007) are an additional formalism which is more flexible than DPSEG. An adaptor grammar (AG) is a generalization of probabilistic context-free grammars (PCFGs). PCFGs are a particular type of language model which represents language as a series of context-free rules, each of which has an associated probability. A simple example set of rules is given in Table 3.2.

Rule	**Prob.**
S $\rightarrow$ NP VP	1.0
VP $\rightarrow$ V NP	0.6
VP $\rightarrow$ V PP	0.4
PP $\rightarrow$ P NP	1.0
NP $\rightarrow$ D NP	0.8
NP $\rightarrow$ N	0.2

Table 3.2: An example set of rules for a PCFG. Each rule indicates that the left-hand item can be made up of the item(s) on the right-hand side. Note that PCFGs also typically include rules mapping each terminal node (e.g. N) to individual lexical items. These are excluded here as their details are irrelevant for the purposes of the current explanation.

Adaptor grammars generalize PCFGs by introducing the idea of an *adaptor*, a function which takes a distribution and maps it onto a distribution with the same support. If we let G be the distribution over syntactic trees defined by a PCFG, then an adaptor C can be applied in order to create a new distribution over the same syntactic trees. Importantly, this allows adaptor grammars to relax the assumptions of a PCFG in numerous ways. Perhaps most importantly, an adaptor grammar need not treat each PCFG rule as independent. Johnson et al. (2007) show how this framework can make use of Pitman-Yor processes as adaptor functions, and in doing so are able to replicate the unigram DPSEG model using general PCFG rules as in Table 3.3.

The model assumes that a sequence is made up of one or multiple words, but that there is no relationship between those words (i.e. unigram assumption), and that each word is likewise made up of one or more characters (i.e. phonemes). A word with three phonemes would need to apply the Chars → Chars Char rule twice and the Chars → Char rule once in order to terminate. Because the probability of each rule is multiplied together, this ensures that long words have low probability and likewise utterances made up of many words also have lower probability, just as with the unigram DPSEG model.

Words → Word
Words → Word Words
Word → Chars
Chars → Char
Chars → Chars Char

Table 3.3: A set of CFG rules which can be used to replicate the unigram DPSEG model. As with the PCFG, each rule indicates that the left-hand item can be made up of the item(s) on the right-hand side.

Further changes to the allowed PCFG rules can allow for the modeler to incorporate more complex assumptions about how language is structured. For instance, if we assume children expect to find morphology, then one might add rules such as Word → Stem Suffix. In a model without morphology the word *walking* would need to be treated as an atomic unit, losing out on the ability to recognize the similarity between *walk* and *walking*. With suffix rules, *walking* could be decomposed into a stem, *walk*, and a suffix *ing*. This allows the model to capture additional sub-word regularities in a language.

At the same time, incorporating more complex assumptions about the structure of language requires the modeler to believe that the infants being modeled also make similar assumptions. It is perhaps an empirical question as to what language model might best describe the knowledge of 6-month-olds. Unfortunately, there exists no current adaptor grammar model which is capable of learning the structure of PCFG rules themselves while also learning their parameters. Altogether, this means that while adaptor grammars are an improvement upon

the DPSEG model, their use requires assumptions for which there is currently no empirical evidence either to support or disprove. However, this may be an exciting area for future work.

3.3 Heuristic Models

3.3.1 Unique Stress Constraint

The previous models formally define the probability of an utterance and produce a segmentation which attempts to maximize that probability. An alternative style of model is one which eschews optimization in favor of speed, known as a heuristic. These models have a long history both in psychology, where they were popularized by the work of Tversky and Kahneman (1974), as well as in computer science (Hart et al., 1968; Newell and Simon, 1976; Russell and Norvig, 2013).

A heuristic model is one in which a series of (generally) simple steps are chosen. These steps do not optimize a solution in the same fashion as the algorithms described previously. Instead, the heuristic model aims to arrive quickly at a good answer, trading guarantees of optimality for speed. In terms of cognitive modeling, heuristic models exist at Marr's algorithmic level. This makes them most comparable to the online Bayesian learners which similarly attempt to model how a process is solved, rather than how it might optimally be solved.

The first prominent heuristic segmentation strategy is the Unique Stress Constraint (USC) learner of Gambell and Yang (2006). The USC assumes that the learner inherently knows something about the way languages operate, namely that any word can have *at most* one

primary stress [4]. The model combines stress with syllable-based TPs in order to quickly segment a corpus. The model evaluates each utterance in turn and starts from the left and moves rightwards through the utterance. After each syllable it decides whether there should be a boundary in two steps:

1. If there are two stressed syllables next to one another, insert a word boundary.
2. If there are two or more unstressed syllables in between two stressed syllables, a boundary is placed between the nearest stressed syllables at the location where the pairwise TP is lowest.

For example, take the sentence *The big dog is awake.* In IPA, the symbol ˈ indicates that the following syllable has primary stress, while . indicates a lack of stress. The previous sentence would then be represented as /θəˈbɪgˈdag.ɪz.əˈwek/. Let the TPs between each pair of syllables be defined as $P(big|the) = 0.3, P(dog|big) = 0.2, P(is|dog) = 0.6, P(a|is) = 0.8, P(wake|a) = 0.3$. The USC learner would then go through the following steps:

1. No boundary is inserted between /θə/ and /ˈbɪg/ because only one of the syllables is stressed and the conditions for step 2 are unfulfilled.
2. A boundary is placed between /ˈbɪg/ and /ˈdag/ because both syllables are stressed.
3. Because there are two unstressed syllables in a row (/ɪz/ and /ə/, a boundary is placed between /ə/ and /ˈwek/ because these syllables have the lowest TP in that sequence of syllables.

Essentially, this means that the model inserts a single word boundary between any two stressed syllables. In cases where there are multiple unstressed syllables in a row, the model

[4]Linguists often make the distinction between primary and secondary stress. Syllables with primary stress are more acoustically pronounced than any other syllable within the word. Longer words, however, often have additional syllables with a somewhat lesser, secondary stress.

takes advantage of TP minima in order to insert a single boundary. This strategy builds off the idea that TPs of syllables within a word are relatively high, given that the syllables form a unit, while TPs of syllables between words are relatively low, because they do not form a unit.

The USC on its own achieves relatively good performance on English (token F-score: 72.3). The USC learner faces a number of challenges however. The most troubling concern is how stressed syllables are placed within the corpus. Gambell and Yang (2006) make use of the CMU pronunciation dictionary (Weide, 1998) to place stress within the corpus. Pronunciation dictionaries, including the CMU, tend to place stress on every word, unless that word appears unstressed in almost all instances. For example, a natural reading of the sentence *Who is the boy near the bridge* might put stress on *who*, *boy*, and/or *bridge* depending on the speaker and the item they intend to place focus on. The CMU dictionary places stress on every word in that sentence excluding *the*. The additional stressed syllables placed by this method aid the model in discovering word boundaries, but would not be available to an actual infant learner. Therefore it is unclear to what degree the model might succeed on the actual language children are exposed to.

3.3.2 Subtractive Segmentation

Later work on the USC learner incorporated more complex assumptions about the steps the learner might use in order to segment. This eventually resulted in the subtractive segmentation algorithm of Lignos (2011). Subtractive segmentation (SubtrSeg) builds off the idea that once a word has been recognized, it might be *subtracted* out of newly encountered utterances. This is compatible with the findings of Bortfeld et al. (2005), although it is not necessarily the same strategy used by infants. This simplifies the segmentation problem and allows the learner to treat the leftover pieces of the utterance as possible words.

Subtractive segmentation may be carried out either making use of the USC or not. A pure subtractive segmentation algorithm goes through each utterance and makes decisions about the segmentation as follows:

1. If the utterance begins with one or more recognized words, choose the highest scoring word and insert a boundary after it.
2. If a word is found, increment its score by 1 and advance by one syllable.
3. If a word was not found, advance by one syllable.
4. Repeat steps 1-3 until the end of the utterance is encountered.
5. Add the syllables between the last inserted boundary (or the beginning of the utterance if no words were found) and the end of the utterance to the lexicon with a score of 1.

For example, assume the learner has a lexicon with the following scores: {*the*: 3, *dog*: 1, *a*: 5}. Again, for the sentence *The big dog is awake*, the SubtrSeg learner would operate in the following manner:

1. The word *the* would be segmented because it is in the lexicon. Its score is incremented to 4.
2. The syllable *big* is skipped because it is not recognized in the lexicon.
3. The syllable *dog* is segmented because it is in the lexicon. Its score is incremented to 2. Note that because *big* was skipped, it does not enter the lexicon even though it was segmented.
4. The syllable *is* is skipped because it is not recognized in the lexicon.
5. The syllable *a* is segmented because it is in the lexicon. Its score is incremented to 5.

6. The syllable *wake* is added to the lexicon with score of 1 because it is between the last inserted boundary and the end of the utterance.

Essentially, this learner begins with an empty lexicon and treats newly encountered utterances as words. Later, if it encounters a word it recognizes from the lexicon it will segment it out by inserting boundaries and then add to the lexicon whatever appears *after* that word. Each time it encounters a word in the lexicon or adds a word to the lexicon, that words score is incremented by one. Note that the learner does not add to the lexicon the syllables which occur before the first identified word. Adding the USC to the SubtrSeg algorithm requires adding an additional step at the very beginning of the process. In going through the utterance, if the USC requires a boundary to be inserted, then a boundary will be placed and the resulting word will be added to the lexicon.

The model as presented is a greedy segmenter, making decisions as quickly as possible, but the model can be adapted so that it considers multiple possible segmentations in a non-greedy manner. This is done through beam search, the model keeps track of at most two separate segmentation hypotheses and evaluates each based on the geometric mean of the word scores. When choosing from competing segmentations, the lexical items in the winning segmentation have their scores incremented by one, and the lexical items in the losing segmentation have their scores decremented by one. This rewards words which participate in good segmentations, and penalizes words that appear in bad segmentations. The model progresses as before through the utterance, keeping track always of the two highest scoring segmentations and at the end of the utterance chooses the segmentation with the highest score.

The SubtrSeg algorithm performs quite well on English even without making use of the USC. It performs better with beam search than without, but both versions perform reasonably well (word token F-score with beam search: 84.9, without: 79.7). A benefit of the approach is

that it operates over syllables, which may be a good unit of representation for infants of the age being modeled. On the other hand, the model makes many assumptions regarding the process of segmentation without any systematic reasoning as to their inclusion. For instance, the model assumes that children initially treat all utterances as words. No experimental evidence indicates that this is the case, in fact, evidence from neonates suggests that already at this age infants do not treat utterances as atomic units (Teinonen et al., 2009). The model also assumes that infants keep track of a score for each lexical item, with that score reflecting not just the frequency of the word, but being increased and decreased according to whether or not the word was chosen when multiple possible segmentations existed. Again, this implementational detail was chosen in the absence of strong experimental evidence and without any specific reasoning given by the authors.

3.4 Results of Previous Work

Looking at Table 3.4, one can see that no single model performs near ceiling, i.e. token F-score = 100, on speech segmentation. The most consistently high-performing model is the AG which performs well on a number of different languages (token F-score: 55.6 - 77). It should be noted, however, that each reported score represents a different model within the AG framework. High performance on Sesotho requires the modeling of morphological units and their ordering within a word, as opposed to English where the relative lack of morphology makes its modeling unnecessary. Other AG frameworks report results with varying numbers of intermediary units between the utterance and phoneme levels. It would appear there is no single AG model which best solves the segmentation problem across all languages. This suggests that a learner might need to infer the correct structure in order to begin segmenting utterances, a process for which no solution has yet been proposed.

Model	Study	Language	CDS	Token F	Boundary F	Lexicon F
MBDP-1	Goldwater et al. (2009)	English	Yes	68.2	82.3	52.4
DPSEG-1	Goldwater et al. (2009)	English	Yes	53.8	74.3	57.2
DPSEG-2	Goldwater et al. (2009)	English	Yes	72.3	85.2	59.1
	Fleck (2008)	Arabic	No	32.6	63.8	9.5
	Fleck (2008)	Spanish	No	57.9	79.3	17.0
Adaptor Grammar	Johnson et al. (2010)	English	Yes	69.5		
	Johnson 2008	Sesotho	Yes	55.6		
	Johnson & Demuth (2010)	Chinese	Yes	77		
	Fourtassi et al. (2013)	Japanese	Yes	70		

Table 3.4: Word token, boundary, and lexicon F-scores as reported in previous phoneme-based work. Note that results for the MBDP-1 model are taken from Goldwater et al. (2009) as the original work from Brent (1999) does not present exact results. Empty cells represent results which were not reported in the original work. DPSEG-1 refers to the unigram DPSEG model, while DPSEG-2 refers to the bigram model. All DPSEG results are presented using the batch inference method of Goldwater et al. (2009).

Table 3.5 displays the word token, boundary, and lexicon F-scores reported by Pearl et al. (2011). Each of the four inference algorithms examined is presented both with the unigram and bigram language model. The BatchOpt results replicate the findings of Goldwater et al. (2009) with the bigram model outperforming the unigram on word token F-scores (DPSEG-1: 54.8, DPSEG-2: 71.5). Interestingly, Pearl et al. (2011) find that in some cases the online learners outperform the BatchOpt learner. This is particularly the case for the unigram learners, where all online learners outperform the BatchOpt on word token and boundary F-scores (e.g. BatchOpt: 54.8, OnlineMem: 67.8), with the OnlineMem learner also outperforming the BatchOpt on lexicon F-scores (BatchOpt: 62.3, OnlineMem: 65.0). Using the bigram language model, this advantage largely disappears, although the bigram OnlineMem still achieves a higher word token F-score than the bigram BatchOpt learner

Model	Learner	Token F	Boundary F	Lexicon F
DPSEG-1	BatchOpt	54.8	74.4	62.3
	OnlineOpt	65.9	81.0	59.3
	OnlineSubOpt	58.5	76.7	51.8
	OnlineMem	67.8	82.6	65.0
DPSEG-2	BatchOpt	71.5	85.0	69.1
	OnlineOpt	69.4	83.5	64.5
	OnlineSubOpt	39.8	66.8	36.5
	OnlineMem	73.0	85.7	62.6

Table 3.5: Word token, boundary, and lexicon F-scores on the Bernstein-Ratner corpus as reported in Pearl et al. (2011).

(BatchOpt: 71.5, OnlineMem: 73.0). A further description of why online learners sometimes outperform the BatchOpt learner will be given in Chapter 5.

Model	Paper	Corpus	Token F	Boundary F	Lexicon F
USC	Gambell & Yang (2006)	Brown	72.3		
SubtrSeg	Lignos (2011)	Brown		90.7	
	Lignos (2012)	Brown	79.7		
SubtrSeg + USC	Lignos (2011)	Brown		93.0	

Table 3.6: Word token and boundary F-scores as reported in previous syllable-based work. Empty cells represent results which were not reported in the original work.

In contrast to phoneme-based work, there has been little investigation into strategies which might apply over syllables. This work has almost exclusively focused on heuristic models, making use of either the USC or subtractive segmentation or some combination thereof. Performance is often only measured either over word tokens or individual boundaries, making model comparison somewhat difficult. Because correctly segmenting a word token generally requires placing two boundaries properly (i.e. the boundaries immediately preceding and following the token), word token scores tend to be somewhat lower than boundary scores. The highest boundary performance is given by a combination of subtractive segmentation and the USC (see Table 3.6), while a pure subtractive segmentation algorithm obtains the highest token F-score performance.

3.5 Future Work

In spite of the wide variety of work which has previously been done in modeling speech segmentation, there remain a number of open questions about how best to incorporate experimental evidence into the modeling process. In the following chapters, I will describe my work and how it improves upon previous modeling attempts. Both the Bayesian and heuristic approaches show promising results but have never been directly compared. Of the Bayesian models, I will examine the DPSEG learner because of its ability to perform online inference and for its simplicity in comparison to the AG framework. I will compare this model to the heuristic SubtrSeg learner because it performs very well on English without incorporating a large amount of language-specific knowledge (as with the USC learner).

In Chapter 4 I investigate the role of the unit of representation (phoneme vs. syllable) in segmentation behavior for each model. The models are tested on seven distinct languages, ensuring that model behavior is not dependent on the idiosyncrasies of any one particular language. I will show that the Bayesian approach successfully segments each language regardless of the unit of representation. The SubtrSeg model relies on syllables as a unit of representation and succeeds on only a portion of the languages tested.

In Chapter 5 I show that the discrepancy in performance between the Bayesian and heuristic models is not due to a difference of batch versus online processing. Even when the Bayesian model learns in an online fashion it still successfully segments each language. By performing a detailed error analysis, I show that each inference process makes predictions about the types of errors one might expect from a child.

Lastly, in Chapter 6 I address the difficulties with evaluating models based entirely on gold standard results. I propose the use of joint and downstream modeling techniques in order

to better understand how a model's behavior relates to its utility in the general language acquisition process.

Chapter 4

Segmentation with Infant-like Representations

4.1 Representing Data for Speech Segmentation

As discussed in Chapter 3, before any model can be implemented, decisions must be made regarding both what information the model makes use of and how that information is structured. Since we are attempting to model the very beginnings of statistical speech segmentation, we must base these decisions, so far as possible, on experimental data from that age, approximately six to seven months. Based on the experimental data reviewed in Chapter 2, this means that the model should make use of statistical regularities in the data, but should avoid making use of cues used by older infants such as lexical stress, phonotactics, or coarticulation effects.

Previous work has largely assumed that infants perceive speech as a string of phones or phonemes and that the problem of segmentation involves placing word boundaries between these units (Brent, 1999; Venkataraman, 2001; Johnson and Goldwater, 2009; Goldwater

et al., 2009; Pearl et al., 2011). As discussed in Chapter 2, there's good evidence that infants at six months do not know the full phonology of their native language and therefore using phonemes can be ruled out (Werker and Tees, 1984; Werker and Lalonde, 1988; Best et al., 1988). Experimental evidence does not rule out the possibility that infants perceive speech as a sequence of phones, as noted by Jusczyk (1997). Phones, however, are an unlikely unit of representation for the purposes of speech segmentation given that infants do segment using statistical information (Saffran et al., 1996; Aslin et al., 1998; Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003) but do not track statistical information over phones or phonemes until after segmentation has begun (Mattys et al., 1999). This suggests that for early statistical segmentation strategies, using any kind of segmental units is likely inappropriate.

Unfortunately, this throws into doubt the results of previous Bayesian models of speech segmentation, all of which assume segmentation operates either over phones (Fleck, 2008; Boruta et al., 2011) or phonemes (Brent, 1999; Johnson and Goldwater, 2009; Goldwater et al., 2009; Johnson and Demuth, 2010; Pearl et al., 2011; Fourtassi et al., 2013). Syllable-based models have also been proposed (Gambell and Yang, 2006; Lignos and Yang, 2010; Lignos, 2011, 2012), but are all heuristic models.

In this chapter, I explore the ability of both Bayesian and heuristic models to operate over syllables[1]. I examine the Bayesian DPSEG models (unigram and bigram) of Goldwater et al. (2009) as well as the subtractive segmentation model of Lignos (2011) and compare each to a number of baseline models. In order to gain a better understanding of how these models operate, each model is tested on a variety of languages, ensuring that model results are robust across linguistic variation.

[1]All models were additionally tested on the phoneme-based corpora of each language for comparison purposes. The results from this analysis are presented in Appendix A.

4.2 Corpora

4.2.1 English Corpus

The English corpus used to evaluate these learners is the UCI Brent syllables derived corpus (Brent, 1999; MacWhinney, 2000; Phillips and Pearl, 2015).[2] This corpus is composed of naturalistic child-directed speech produced by mothers as they interacted with their children in their own homes. While the full corpus contains speech directed at infants aged six to fifteen months, I use only the portion of the corpus aimed at infants nine months of age or younger, following Pearl et al. (2011). This results in a corpus of 28,391 utterances.

In order to properly represent the child-directed speech for the model, I first had to take the orthographic text and convert it into a phonemic representation. I accomplished this by making use of the MRC Psycholinguistic Database (Wilson, 1988), which contains pronunciations for many common words in British English. Many words in the corpus do not appear in the MRC database, possibly because they represent names, nonsense words, or are non-standard, including those words common to "motherese" (e.g. *toothie*, *footsie*, *baba*). For these missing items, phonemic representations were created by hand. If no pronunciation could be established, the words were removed from the corpus.

One difficulty with using phonemically-encoded corpora is that this automatically assumes that the learner knows the full phonology of their native language. Given that this is unlikely to be the case for children at six to seven months of age (see Section 2.1.4), this poses a significant problem. Unfortunately, creating more detailed phonetic representations from the original audio is both time-consuming and expensive. Difficulties with phoneme-based representations do not dissapear when using syllables, but their impact is in some cases reduced.

[2]The UCI Brent syllables corpus is available at `http://childes.psy.cmu.edu/derived/uci_brent_syl.zip`

For instance, phonetic variation is often conditioned within a syllable. As an example, vowels in English are often phonetically nasal when a nasal occurs in the coda ([k^hæt˺] versus [k^hæ̃n]). In this case, because each syllable is represented as a unique whole with no subsyllabic structure, this variation between [æ] and [æ̃] is automatically captured. Phonetic variation which occurs between syllables still remains a problem as does the phenomenon of free variation, where multiple realizations can be chosen at will.

Once the corpus is represented phonemically, it must then be syllabified. When possible, syllabifications were taken directly from the MRC database. When this was not possible, syllabification was done using the maximum-onset principle (MOP, Selkirk, 1981). This general principle states that the onset (or beginning) of a syllable should be as large as possible, so long as it does not violate any phonotactic rules in the language. For instance, in syllabifying the word *onset*, the largest possible onset for the second syllable would be the sequence *ns*, but this is ruled out as it is not a valid English consonant cluster (e.g. English words never begin with the sequence [ns]). The next largest onset would be *s*, which is a valid onset; therefore it is included in the beginning of the second syllable. This produces the syllabification *on set*. The MOP and hand-coded syllabifications produce the same result on 98% of the words in the corpus. They tend to differ in cases where there are multiple consonants between two vowels, especially with the consonant [s]. For example, the MRC syllabifies *escape* as /əsˈkep/, while the MOP prefers /əˈskep/. While this simple principle does not capture the full range of syllabifications which exist cross-linguistically, it provides a useful baseline for creating syllabifications.

4.2.2 Other Languages

Evaluating only on English poses a number of problems based primarily on the fact that the model results are not generalizable to the wide variety of languages which exist. In order

to remedy this fact, I took corpora from the CHILDES database in a number of different languages (MacWhinney, 2000). The task of collecting an appropriate cross-linguistic range of input data is non-trivial, even with this vast database, due primarily to two factors. First, even with its broad collection of child-directed speech corpora, the CHILDES database does not have a large amount of speech directed at children under one year of age, which is when early segmentation strategies would be in use. Second, even when age-appropriate data are available, they are often available only as orthographic transcripts, which then need to be converted into a syllabified, phonemic form. Taking into account the various factors needed for an appropriate speech segmentation corpus, I chose corpora from CHILDES in the following languages: German, Spanish, Italian, Farsi, Hungarian, and Japanese.

As with the English corpus, each of the different corpora must be converted both into a phonemic and a syllabified form. Fortunately, some derived corpora already exist in CHILDES that provide a phonemic and syllabified transcription (e.g., Hungarian, Italian). For languages whose corpora were purely orthographic (e.g., Japanese, Farsi, Spanish, and German) I used pronunciation dictionaries and linguistically-trained native speakers in order to create a proper phonemic transcription. For syllabification, I primarily used adult syllabification judgments, but when these were unavailable, syllabification was done using the MOP. As noted previously, although the MOP does not align with all adult judgments, it is a simple guideline which accurately describes a great deal of syllabification across the world's many languages.

Table 4.1 provide basic descriptive statistics regarding the various syllable-based corpora used for this analysis. Although care was taken to seek out corpora of an appropriate age range, not all languages have corpora available directed towards children younger than one year of age. The corpora also vary in their size, although the smallest corpus (German) is still large enough to provide good results. I also list the number of unique syllables in each corpus as this varies considerably by language (e.g. Spanish: 522, Hungarian: 3029). This

Language	Corpus	Ages	Utt	Syl types	Syls/Utt	B Prob
English	Brent	0;6-0;9	28391	2330	4.16	76.26
German	Caroline	0;10-4;3	9378	1682	5.30	68.60
Spanish	JacksonThal	0;10-1;8	16924	522	4.80	53.93
Italian	Gervain	1;0-3;4	10473	1158	8.78	49.94
Farsi	Family & Samadi	1;8-5;2	31657	2008	6.98	43.80
Hungarian	Gervain	1;11-2;11	15208	3029	6.30	51.19
Japanese	Noji, Miyata, & Ishii	0;2-1;8	12246	526	4.20	44.12

Table 4.1: Summary of the syllabified child-directed language corpora, including the CHILDES database corpora they are drawn from (Corpus), the age ranges of the children they are directed at (Ages), the number of utterances (Utt), the number of unique syllables (Syl types), the average number of syllables per utterance (Syls/Utt), and the probability of a word boundary appearing between syllables (B Prob).

results from the fact that some languages, such as Spanish and Japanese, have relatively strong phonotactic restrictions placed on syllables (e.g. in Japanese only the phoneme /ɴ/ may appear after a vowel), while languages such as English, German, and Hungarian allow much more complex syllable types (e.g. consider the English coda in *warmth*, /rmθ/).

Also presented are the average number of syllables per utterance. Corpora of older child-directed speech tend to have longer utterances, leading to a higher unit per utterance count. Languages with lower numbers of syllables per utterance should be easier to segment than languages with longer utterances. This is due to the fact that a higher percentage of words are either utterance-initial or utterance-final. Consider that words neighboring an utterance boundary are easier to segment since only one boundary must be identified correctly (because the utterance boundary is known by the model). In corpora with long utterances, this impacts only relatively few words, but for corpora with short utterances (e.g. English and Japanese) utterance boundaries will play a greater role in segmentation performance.

Lastly, I also present the percentage of syllables followed by a boundary. This boundary probability is one way of thinking about the average number of syllables per word. In terms

of random guessing, having a boundary probability of 0.50 is a worst case scenario. If the boundary probability is higher than 0.50, models which insert too many boundaries will be rewarded (since there is a high likelihood the inserted boundaries are correct). Similarly, if the boundary probability is lower than 0.50, models which insert too many boundaries will be punished (since there is a high likelihood the inserted boundaries are incorrect).

4.2.3 Model Training and Parameter Estimation

Because the models being evaluated are probabilistic in nature, each model was trained and evaluated five times with results averaged over each run. Although the models are all unsupervised, a train-test split was utilized in order to better evaluate how each model was able to adapt to new data. To achieve this, each corpus was split five times so that the model was trained on 90% of the corpus, while the remaining 10% served as a test set. The corpora are all organized by date, such that earlier utterances were presented before later utterances just as they would be encountered by an infant. The test sets were taken from any portion of the corpus, but the relative ordering of utterances was kept the same.

In order to run the Bayesian models, values for their free parameters must be set. For the unigram model, there is only a single parameter, α, whose value determines how probable new words are (higher values of α indicate an increased preference for novel words). For each language, the model was trained with a variety of values for α, ranging from 1 to 500. The value which resulted in the best word token F-score was selected for all unigram Bayesian learners for that language. This process was repeated for the bigram learners, selecting values for β and γ, where larger values of β indicate an increased preference for novel bigrams and larger values of γ indicate an increased preference for novel words. The values for β were chosen on the range 1 to 500, while the values of γ were chosen from a slightly larger range,

1 to 3000. These value ranges were chosen based on previous work (Goldwater et al., 2009; Pearl et al., 2011).

	α	β	γ
English	1	1	90
German	1	1	100
Spanish	1	200	50
Italian	1	1	90
Farsi	1	200	500
Hungarian	1	300	500
Japanese	1	300	100

Table 4.2: Parameter values for all unigram and bigram Bayesian models across each language. Values were selected based on previous research in the following ranges α: [1, 500], β: [1, 500], γ: [1, 3000].

Table 4.2 provides the full set of parameter values for each learner. Higher parameter values for α, β, and γ all increase the probability of novel words. In the unigram case, however, it appears that low values of α are always preferred. There is somewhat more variation in the values for the bigram model. How the proper values in this case might be identified is somewhat unclear. One possibility is that infants might learn proper parameter values given their own experience by making use of hyperparameters. Incorporating hyperparameter inference into the DPSEG model may be an important next step in segmentation modeling, although its importance is minimized if infants make use of a unigram assumption. The results presented in this chapter, however, focus only on performance with the optimized free parameters in Table 4.2.

4.3 Baseline Models

4.3.1 TP minima

One interpretation of the results from Saffran et al. (1996) is that infants calculate transitional probabilities between syllables and place boundaries when the TPs reach a low point or local minimum, hereon refered to as the TPminima model. Essentially, a boundary is placed between two syllables whenever both the preceding and following TP are greater than the TP currently under consideration. Gambell and Yang (2006) explore this possibility and find that it does not achieve high performance on English child-directed speech represented as syllables. They attribute this poor performance to the fact that two local minima cannot occur next to one another. This rules out the possibility of placing boundaries between strings of monosyllabic words. Since English is heavily monosyllabic, this ensures that the model places too few boundaries and therefore has low recall.

I include the TPminima model as a baseline which makes boundaries with relatively high precision, but low recall. Because the TPminima learner struggles with the monosyllabic nature of English, it is also an open question as to how successful the strategy might be on other languages.

4.3.2 Random Oracle

An alternative baseline explored by Lignos (2012) is a random oracle segmenter (RandOracle). The random oracle segments a corpus by treating each possible boundary location as a Bernoulli trial. The segmenter is an "oracle" in that it knows the true probability of a boundary in the corpus. The random oracle inserts boundaries with this same probability. This acts as a baseline for comparison, representing the performance that could be achieved

by randomly guessing at the true boundary rate. Any model which does not surpass the RandOracle model should have low enough performance that it could be ruled out as a possible model of early infant segmentation..

4.4 Syllable-based Results

In this section, I first present results as measured against the gold standard segmentation. After discussing the performance of each learner, I investigate the types of errors generated in order to better understand what causes the differences in performance found between each learner.

4.4.1 Gold Standard Results

4.4.1.1 Word Token Results

Word tokens are the most commonly measured unit for segmentation and therefore I first examine the ability of the various learners to produce proper word tokens from the syllable-based corpora.

	Eng.	**Ger.**	**Spa.**	**Ita.**	**Far.**	**Hun.**	**Jpn.**
DPSEG-1	53.1	60.3	55.0	61.9	66.6	59.9	63.2
DPSEG-2	77.1	73.1	**64.8**	**71.3**	**69.6**	**66.2**	**66.5**
SubtrSeg	**86.6**	**83.1**	44.9	42.2	34.1	51.5	33.5
TPminima	13.0	16.4	28.7	32.2	31.5	25.3	32.4
RandOracle	56.4	47.5	27.0	22.8	20.3	26.4	26.1

Table 4.3: Syllable-based segmentation results compared against the adult gold standard for Bayesian and non-Bayesian learners. The top learner for each language is given in bold.

Looking at the word token results for the syllable-based learners, we see very similar results to previous phoneme-based learners (Fleck, 2008; Goldwater et al., 2009; Pearl et al., 2011).

First, we find that the bigram assumption is useful: on every language the bigram DPSEG model outperforms the unigram. The utility of the bigram assumption is essentially the same as it is for previously investigated phoneme-based models, tracking bigrams allows the model to account for words which frequently co-occur (Goldwater et al., 2009). That this assumption holds true across all of our languages is unsurprising. It can also be confirmed that the DPSEG model successfully segments every language, regardless of whether the model used a unigram or bigram assumption. In fact, the worst performance for the DPSEG model comes from the DPSEG-1 on English (F-score: 53.1). In contrast, English is easier for the DPSEG-2 model to segment than other languages, matching previous research (Fleck, 2008; Fourtassi et al., 2013).

Looking at the SubtrSeg model, I find that it has a similar bias towards languages such as English and German, segmenting them much more easily than the other languages. An opposite pattern of results is found for the TPminima learner although in no case does this learner perform particularly well (e.g. best performance is found on Italian: 32.2). This matches the pattern described previously where the TPminima model performs poorly on languages with a monosyllabic bias (Gambell and Yang, 2006). Examining the last baseline learner, we see that the RandOracle performs similar to the DPSEG-2 and SubtrSeg learners in that it performs much better on English and German than on the other languages. What might be driving this behavior?

Fourtassi et al. (2013) have suggested that some languages are inherently more ambiguous with respect to segmentation than others. Specifically, even if all the words of the language are already known, some utterances can *still* be segmented in multiple ways (e.g., /al.rajt/ segmented as *alright* and *all right* in English). The degree to which this happens varies by language, with the idea that languages with high inherent ambiguity would be harder to correctly segment. If this is true, we might expect that low inherent segmentation ambiguity correlates to high performance by statistical segmentation strategies. With this in mind,

perhaps English and German have lower inherent segmentation ambiguity than the other languages.

In order to quantify this ambiguity, Fourtassi et al. (2013) proposed the normalized-segmentation entropy (NSE) metric:

$$NSE = -\sum_{i} P_i \log_2(P_i)/(N-1) \tag{4.1}$$

where P_i represents the probability of a possible segmentation i of an utterance and N represents the length of that utterance in terms of potential word boundaries (so this is determined by the number of syllables for our learners). To calculate the probability of an utterance, we use the unigram or bigram DPSEG generative model equations described in Section 3.2.2, since these represent the probability of generating that utterance under a unigram or bigram assumption. As an example, to calculate the NSE of a single utterance /al.rajt.ðɛn/, we use the unigram and bigram model equations to generate the probability of every segmentation comprised of true English words (P_i above). In this case, two segmentations are possible: *alright then* and *all right then*. The probabilities for each segmentation are then used in Equation 4.1 above, with $N = 2$ since there are two potential word boundaries among the three syllables.

Because a low NSE represents a true segmentation that is less ambiguous for the learners using the n-gram assumptions tested here, English and German should have lower NSE scores if inherent segmentation ambiguity was the explanation for the better segmentation performance. Table 4.4 shows the NSE scores for both unigram and bigram learners for all seven languages, with token F-scores for the respective BatchOpt learners for comparison.

Unigram	NSE	F-score	Bigram	NSE	F-score
German	0.000257	60.3	**German**	0.000502	73.0
Italian	0.000348	61.9	**Italian**	0.000604	71.3
Hungarian	0.000424	59.9	**Hungarian**	0.000694	66.2
English	0.000424	53.1	**English**	0.000907	77.1
Farsi	0.000602	66.6	**Spanish**	0.00103	64.8
Japanese	0.00126	55.0	**Farsi**	0.00111	69.6
Spanish	0.00128	63.2	**Japanese**	0.00239	66.5

Table 4.4: Average NSE scores across all utterances in a language's corpus, ordered from lowest to highest NSE and compared against the BatchOpt token F-score for a language. Results are shown for both the Unigram and Bigram models. Lower NSE scores represent less inherent segmentation ambiguity and higher token F-scores indicate a better segmentation performance.

From Table 4.4, we see that German fits with the hypothesis that low NSE predicts higher segmentation performance, having in both cases the lowest NSE scores. At the same time, English does not fit this pattern, ranking fourth overall for both a unigram and bigram learner in spite of having the lowest token F-scores for the DPSEG-1 model and highest token F-scores for the DPSEG-2 model. Because of this, the high segmentation performance on both German and English cannot simply be due to both having lower inherent segmentation ambiguity.

More generally, it becomes clear by looking at all seven languages that low NSE does not always lead to higher token F-scores. If it was, we would expect to find a significant negative correlation between NSE score and token F-score – but this does not happen (unigram: $r = -0.084$, $p = 0.86$; bigram, $r = -0.341$, $p = 0.45$). Examining individual languages in Table 4.4, this lack of correlation is apparent. The DPSEG-1 Farsi NSE score is ranked fifth lowest, but in fact has the highest F-score, while the DPSEG-1 Spanish NSE score is actually the worst, though it has the second best F-score. When we turn to the bigram learners, we see that Hungarian has the third best NSE score but the next to worst F-score, while English has the fourth worst NSE score but the best F-score. So, NSE cannot be the sole factor determining segmentation performance, though its role may still be non-trivial.

An alternative explanation comes from the performance of the RandOracle learner. Like the SubtrSeg learner, the RandOracle guesser performs better on English and German than the other languages. Because the RandOracle is guessing boundaries at the same rate as in the true corpus, it should do better when there are more true boundaries. The difference in performance is explainable by the fact that English and German have many more monosyllabic words than the other languages. This means that in English and German, the probability of a boundary appearing after any syllable is quite high (76.3% and 74.5% respectively compared to the next highest languages Spanish: 51.3% and Hungarian: 51.2%). The relationship between the probability of a word boundary and the success of the RandOracle learner is illustrated in Figure 4.1. The blue line represents the probability that the RandOracle learner would agree with the true corpus as a function of the boundary probability. The red crosses represent the actual RandOracle performance on individual languages. The fact that randomly guessing is more rewarding in these languages suggests that the increase in performance that many learners see for English and German is simply a result of model guesses resulting in correct segmentations.

The most interesting non-Bayesian learner is the SubtrSeg algorithm from Lignos (2012). The SubtrSeg learner actually outperforms the bigram DPSEG model on both English and German. While the DPSEG performance stays relatively constant across languages, the SubtrSeg sees a large drop-off in performance on all other languages. Again, this is likely due to the fact that guesses made by the SubtrSeg are more likely to result in correct segmentations on English and German as opposed to the other languages. This casts doubt on the viability of the SubtrSeg learner as its previous high performance is likely a result of particular aspects of English, rather than being inherent to the model itself.

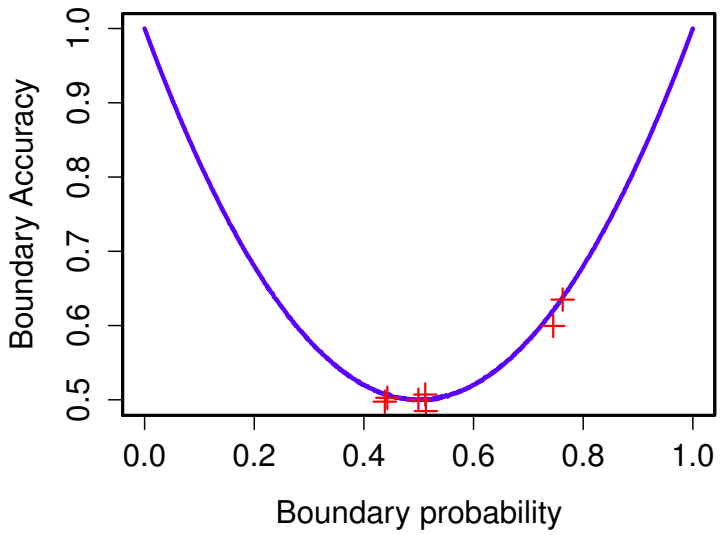

Figure 4.1: Expected boundary accuracy for an informed random guesser (a random oracle) given a corpus of 1,000,000 potential boundaries, with true boundaries appearing randomly with the given boundary probability. Crosses represent each of the seven languages used for evaluation.

	Eng.	**Ger.**	**Spa.**	**Ita.**	**Far.**	**Hun.**	**Jpn.**
DPSEG-1	56.3	63.4	54.5	**54.2**	**57.9**	49.7	**56.8**
DPSEG-2	71.9	**70.9**	**63.5**	53.4	57.8	**51.9**	55.2
SubtrSeg	**74.7**	67.1	39.6	23.8	27.7	36.3	25.8
TPminima	18.3	25.7	32.2	35.5	29.5	29.7	35.7
RandOracle	41.9	35.3	29.1	20.7	21.7	24.0	29.0

Table 4.5: Syllable-based lexicon F-score results for Bayesian and non-Bayesian learners. The highest scoring learner for each language is presented in bold.

4.4.1.2 Lexicon Results

I next examine the learners in terms of their lexicon F-score. While word tokens are a natural unit to measure performance, an important goal of segmentation is to produce a set of words from which later learning can occur. The quality of the lexicon, therefore, is of great importance to actual learners. Lexicon results do not take into account word frequencies, which may be important given that children ignore frequency when making

linguistic generalizations (Yang, 2005; Perfors et al., 2014). Based on Table 4.5, one can see that both the unigram and bigram DPSEG models produce good lexicons regardless of the language. In contrast to the word token results, the bigram assumption is less helpful for producing good lexicons. Turning to the non-Bayesian learners, I again find that the SubtrSeg learner shows high performance only in the case of English and German. The lowest performing Bayesian learner is the DPSEG-1 on Hungarian (F-score: 49.7) while the highest performing SubtrSeg learner on a language other than English or German is found on Spanish (F-score: 39.6). The TPminima learner again shows an opposite pattern of results, performing worst on English and German while the RandOracle learner matches the pattern of the SubtrSeg learner. This shows that Bayesian models, both unigram and bigram, produce more reliable lexicons cross-linguistically than the non-Bayesian learners.

4.5 Conclusion

Taken altogether, these results indicate that syllables are indeed a viable unit of representation upon which to base speech segmentation for English. This is not just because syllables make the task somewhat easier by reducing the number of possible word boundaries because many of the non-Bayesian learners failed to properly segment all of the languages. Indeed, it may be the case that proper cross-linguistic segmentation is only possible for models which incorporate the biases which drive the Bayesian models, as all of the non-Bayesian syllable-based learners failed to segment well on all languages. In contrast, regardless of whether the DPSEG model uses a unigram or bigram assumption, it performs reasonably well on all languages.

The syllable-based results are presented in terms of word token F-scores as well as lexicon F-scores. Although one might assume that the bigram assumption is necessary because it leads to higher word token F-score performance, the lexicon results reveal that the unigram

assumption is sufficient to produce a high quality lexicon. This shows the robustness of the DPSEG strategy: not only does it succeed on all tested languages for syllables, it also succeeds both with the unigram and bigram generative assumptions.

The high performance of the DPSEG model across all seven languages stands in opposition to previous findings using a phoneme-based DPSEG model (Fleck, 2008). This indicates the importance of identifying proper units of representation. When modelers fail to account for actual infant segmentation constraints, they may come to conclusions which are potentially unwarranted. The source of the DPSEG models' robustness comes from the fact that it makes very few assumptions about the nature of the language that it is segmenting. It assumes that sentences are made up of some kind of repeated units (e.g. words), which are in turn made up of syllables. Because it is driven by basic parsimony biases, rather than by complex, language-specific assumptions, it seems reasonable that it might perform well on many languages.

Chapter 5

Model Inference

5.1 Model Inference for Speech Segmentation

The previous chapter explored the role of the unit of representation in speech segmentation, finding that syllables are a viable unit of representation. Here, I examine how incorporating cognitive constraints into model inference can likewise affect model results. While Bayesian models of cognition have had great success in recent years (Xu and Tenenbaum, 2007; Johnson et al., 2007; Frank et al., 2009; Goldwater et al., 2009; Perfors et al., 2011; Feldman et al., 2013; Denison et al., 2013), they often rely on methods of inference which process all of the linguistic data at once (e.g. Gibbs sampling) rather than processing data as it is encountered. Researchers often talk about these models as being on Marr's computational level (Marr, 1982), exploring rational solutions to problems faced by learners. There is growing interest, however, in incorporating more realistic forms of inference to bring these models closer to the algorithmic level, modeling the strategies used by actual learners (Sanborn et al., 2010; Shi et al., 2010; Pearl et al., 2011; Bonawitz et al., 2011; Abbott et al., 2013; Griffiths et al., 2015).

5.2 Cognitive Constraints on Inference

In order to examine the role of the inference process on the results of the DPSEG model, I take the four learners of Pearl et al. (2011) and investigate their performance on syllable-based input. As described in Chapter 3, the four inference methods are the *BatchOpt*, *OnlineOpt*, *OnlineSubOpt*, and *OnlineMem* learners. Table 5.1 summarizes which cognitive constraints are built into each learner.

	Online	**Sub-optimal decisions**	**Recency Effect**
BatchOpt			
OnlineOpt	✓		
OnlineSubOpt	✓	✓	
OnlineMem	✓	✓	✓

Table 5.1: Description of the varying assumptions built into each inference algorithm. Online refers to processing each utterance in turn. Sub-optimal decisions refers to not choosing the highest probability segmentation. Note that the sampling algorithm of the BatchOpt learner does not technically choose the highest probability segmentation 100% of the time. The use of annealing, however, guarantees that in most cases the algorithm will settle only on very high probability segmentations. Because the OnlineMem algorithm does not use annealing in this way, it is considered sub-optimal. Recency effect refers to the model giving extra processing resources to portions of the corpus which have just been encountered.

Three of the learners process the corpus utterance by utterance, making them online learners. The OnlineOpt always chooses the highest probability segmentation, meaning it does not make sub-optimal decisions, and it incorporates no form of memory which might implement a recency effect. The OnlineSubOpt learner chooses segmentations based on their relative probabilities, meaning that it will sometimes choose sub-optimal segmentations. It does not, however, have any form of memory or recency effect built in. Finally, the OnlineMem learner uses a Decayed Markov Chain Monte Carlo (DMCMC) process in order to implement a recency effect (i.e. items near the end of an utterance are processed more heavily than items further in the past). This process selects possible boundaries to sample, and updates them as with Gibbs sampling. Although the BatchOpt process of using Gibbs sampling plus

annealing guarantees that the learner will generally converge on an optimal segmentation, the DMCMC process does not. Instead, each boundary is sampled according to its probability, essentially doing on a boundary-by-boundary basis what the OnlineSubOpt learner is doing utterance-by-utterance. For that reason, I consider the OnlineMem learner to also make sub-optimal decisions.

5.3 Corpora

The same corpora are used here as in Chapter 4. I evaluate the models on the same set of languages: English, German, Spanish, Italian, Farsi, Hungarian, and Japanese, using the same train-test procedure as described in Section 4.2. Because syllables appear to be a more likely candidate for speech segmentation, I only examine the models on the syllabified versions of the corpora.

5.4 English Results

In order to better highlight the impact of online learning on model results, I first present the results for English alone. Although the results from a single language are not necessarily indicative of model behavior across languages, focusing on a single set of results allows the presentation of results in a more concise format. First, I discuss the gold standard results as was done in Chapter 4. In order to better understand the differences between the various online learners, I then analyze the errors made by each learner.

		Token F-score	Type F-score
DPSEG-1	**BatchOpt**	53.1	56.2
	OnlineOpt	58.8	43.9
	OnlineSubOpt	63.7	48.5
	OnlineMem	55.1	54.5
DPSEG-2	**BatchOpt**	77.1	71.9
	OnlineOpt	75.1	64.3
	OnlineSubOpt	77.8	65.2
	OnlineMem	86.3	**75.5**
Other	**Subtr.Seg.**	**86.6**	74.7
	TPminima	13.0	18.3
	RandOracle	56.4	41.9

Table 5.2: Syllable-based segmentation results compared against the adult gold standard for Bayesian and non-Bayesian learners, showing average word token and word type F-score. The best results for each measure are presented in bold.

5.4.1 Gold standard Results

Table 5.2 presents the word token and word type F-scores for all Bayesian and non-Bayesian learners tested on the English corpus. The best scores for each measure have been printed in bold. The word token results give a better indication of how the model does on frequent words, since tokens count every individual instance of a word. The word type results, on the other hand, better reflect how well the learner does on less frequent items. The highest word token results comes from the SubtrSeg learner, although the DPSEG-2 OnlineMem learner performs nearly as well (OnlineMem: 86.3, SubtrSeg: 86.6). The highest word type results, however, come from the DPSEG-2 OnlineMem learner with the SubtrSeg performing slightly lower (OnlineMem: 75.5, SubtrSeg: 74.7).

In the previous chapter, unigram and bigram BatchOpt learners were compared, finding that the bigram assumption is generally useful for segmentation. This holds true for the online learners as well: in no case does performance moving from unigram to bigram decrease. This indicates that the bigram assumption is useful in English not just when using optimal inference, but also when using cognitively constrained inference.

Turning to just the DPSEG-1 learners, I find that online learners actually outperform the BatchOpt learner on word token F-score. At the same time, these online learners produce somewhat lower word type F-scores when compared to the BatchOpt. In particular, the OnlineOpt and OnlineSubOpt learners produce much worse word type F-scores than the BatchOpt (OnlineOpt: 43.9, OnlineSubOpt: 48.5, BatchOpt: 56.2). The OnlineMem does slightly better than the other online learners, but is still worse than the BatchOpt (OnlineMem: 54.5, BatchOpt: 56.2). This appears to indicate that for the unigram assumption, using online learning actually increases performance on common words (aiding in word token scores), while decreasing performance on infrequent words. Of the online learners, the unigram OnlineMem does the worst in token F-score, but best in type, indicating that this learner is less biased towards frequent items.

Looking at the DPSEG-2 learners, there is clearly an improvement over their unigram counterparts. While all unigram online learners outperformed the DPSEG-1 BatchOpt learner, the bigram constrained learners show a less robust increase in performance. The only online learner to reliably outperform the BatchOpt is the OnlineMem learner, although it does so both for token and type F-scores. The OnlineMem learner sees a very substantial increase in word token F-score (BatchOpt: 77.1 vs. OnlineMem: 86.3). In contrast, the OnlineOpt and OnlineSubOpt learners do much more poorly, especially in terms of their lexicons, captured by their Type F-score.

One possible explanation for why online learners sometimes outperform the BatchOpt is that perhaps the BatchOpt learner is, for unknown reasons, not converging on an appropriate solution. If this were the case, we might expect to find that the online learners actually achieve higher log posterior scores than the BatchOpt. The log posterior of a Bayesian learner indicates the probability of a hypothesis (in this case a segmentation) given the data. Log posteriors closer to 0 indicate that the model was better able to create a segmentation which matched the underlying unigram or bigram generative model. Essentially, if the BatchOpt

does not achieve a better log posterior than its online equivalents, then some aspect of the model's inference is deficient.

		Token F-score	Log Posterior
DPSEG-1	**BatchOpt**	53.1	$\mathbf{-5.51 \times 10^5}$
	OnlineOpt	58.8	-6.46×10^5
	OnlineSubOpt	**63.7**	-6.45×10^5
	OnlineMem	55.1	-6.01×10^5
DPSEG-2	**BatchOpt**	77.1	$\mathbf{-5.53 \times 10^5}$
	OnlineOpt	75.1	-6.23×10^5
	OnlineSubOpt	77.8	-6.32×10^5
	OnlineMem	**86.3**	-5.78×10^5

Table 5.3: Log posterior and word token F-score presented for each of the Bayesian learners. Note that log posteriors closer to zero indicate better model fit.

Table 5.3 presents the word token F-score and log posterior results for each learner. Note that posteriors can only be compared between learners with the same underlying model (i.e. DPSEG-1 learners can all be compared with one another, but not against any DPSEG-2 learner). The BatchOpt learner does indeed produce a better model fit than the online learners, despite the fact that it achieves lower gold standard results in some cases. This mismatch can be explained by the underlying modeling assumptions built into both the DPSEG-1 and DPSEG-2 models. Both of these models make unrealistic assumptions about how language is structured. Because of this, an "optimal" segmentation that matches the underlying naive generative assumptions may not match the target language. Instead, the segmentation which optimizes the log posterior will be the one which best fits the unigram or bigram Dirichlet process, rather than the one which best matches the grammar of English (or any other natural language).

5.4.2 Error Patterns

In order to better establish what causes the differences in behavior between the batch and online learners on English, I look at the specific types of errors which each model produces.

5.4.2.1 Over- and Undersegmentation

		Over %	Under %	Token F
DPSEG-1	**BatchOpt**	1.7	98.3	53.1
	OnlineOpt	5.0	95.0	58.9
	OnlineSubOpt	6.5	93.5	63.7
	OnlineMem	9.0	91.0	55.1
DPSEG-2	**BatchOpt**	13.8	86.2	77.1
	OnlineOpt	13.2	86.8	75.1
	OnlineSubOpt	18.8	81.2	77.8
	OnlineMem	46.3	53.7	86.3
Other	**SubtrSeg**	95.3	4.7	86.6
	TPminima	0.5	99.5	13.0
	RandOracle	70.5	29.5	56.4

Table 5.4: Percentage of errors which resulted in over- or undersegmentations for all Bayesian and non-Bayesian learners in English. For the purposes of this analysis, errors which result in both an over- and an undersegmentation (e.g. *the doggie* segmented as *thedog gie*) are ignored.

There are stark differences between some of the learners in terms of whether they prefer to over- or undersegment in English. Oversegmentation refers to erroneously inserting an extra boundary within a word, while undersegmentation occurs when the model fails to insert a boundary, essentially combining two or more words together. Two of the non-Bayesian learners tend to drastically oversegment (e.g. SubtrSeg: 95.3% and TPminima learners: 99.5%), while the TPminima learner drastically undersegments (99.5%). There is also a divide between the unigram and bigram Bayesian learners, with the latter producing a much higher rate of oversegmentations. Generally, the online learners oversegment more regularly than the BatchOpt learners, both for the unigram and bigram case. The OnlineMem learner, in particular, has a much greater tendency to oversegment, especially in the bigram case, than any other Bayesian learner (46.3%).

To evaluate models based on their over- and undersegmentation performance, it first makes sense to understand what types of errors infants produce. Unfortunately, there are no definitive studies on this subject, but evidence from diary studies suggests that young children

tend to undersegment at least until the age of 3 years (Brown, 1973; Peters, 1983). This would suggest that SubtrSeg and RandOracle learners do not match evidence from children. That being said, diary studies do not quantify this undersegmentation and therefore models must be compared based on qualitative patterns only. The online Bayesian learners tend to oversegment more so than the BatchOpt learner, but in no case do oversegmentations make up the majority of errors.

5.4.2.2 Reasonable Errors

Another way to measure the errors a model makes is to group errors into different categories. Anecdotal evidence from previous studies has claimed that many errors made by the DPSEG model were relatively reasonable and might be similar to errors made by children (Goldwater et al., 2009; Pearl et al., 2011). In order to quantify these errors, I chose three categories to investigate:

1. Oversegmentations that result in **real words** (e.g., *alright* /ɑl ɹaɪt/ segmented as *all* /ɑl/ and *right* /ɹaɪt/)
2. Oversegmentations that result in **productive morphology** (e.g., segmenting off *–ing* /iŋ/)
3. Undersegmentations that produce **function word collocations** (e.g., segmenting *is that a* as *isthata*)

These three errors were chosen because they result in units that may be useful to early learners or because they are evidenced in actual infants. Errors which produce real words are potentially useful in the sense that they help the infant to learn actual words in their language. Similarly, errors which produce morphemes result in useful linguistic units which must be learned by the infant. Finally, function word collocations occur when infants treat

common sequences of words as a single unit. Errors of each type are attested in the productions of older children, around 2 to 3 years of age, although the rates of each error are not quantified (Brown, 1973; Peters, 1983).

In order to accurately count the number of these errors, a few guidelines were put in place for their identification. Real word errors are defined as those which resulted in at least one segmented word which occurred in the true corpus. Many CDS corpora contain non-words which appear with low frequency (e.g. transcription errors, babbled speech). To avoid counting segmentations which reproduce these non-words in the corpus, real word errors were limited to those which resulted in a word that occurred at least five times in the corpus.

To identify morphology errors it was necessary to create a list of common morphemes in English. Because the model treats each syllable as an atomic unit, it is impossible for the model to recognize sub-syllabic morphemes (e.g. plural *-s*) and these morphemes were not included in the analysis. Segmented items were counted as morphology errors only in cases where the morpheme was segmented from the proper location. For example, the morpheme *-ful* being segmented out of *helpful* would be treated as reasonable because a suffix was segmented off the end of a word. Segmenting *-ful* out of *fulfill* would not be treated as reasonable because the suffix was segmented off the beginning of a word.

Similar to morphology errors, identifying function word collocations requires creating a list of function words in English. This list included 179 common English determiners, prepositions, pronouns, conjunctions, particles, and auxiliary verbs. In order for an item to be treated as a reasonable function word collocation, it needed to be produced by undersegmenting a sequence of function words in the corpus.

Table 5.5 presents the percentage of errors which produce either 1) a real word, 2) a morpheme, or 3) a function word collocation.

		Real word	Morph.	Func. words	Total Errors
DPSEG-1	**BatchOpt**	5.5%	2.6%	14.7%	3124.2
	OnlineOpt	18.8%	3.4%	1.0%	7481.0
	OnlineSubOpt	16.2%	3.7%	1.0%	9383.0
	OnlineMem	8.5%	3.6%	12.1%	3301.6
DPSEG-2	**BatchOpt**	8.2%	3.0%	11.5%	2937.8
	OnlineOpt	21.0%	5.0%	2.1%	4751.2
	OnlineSubOpt	12.7%	3.1%	1.0%	9827.2
	OnlineMem	9.3%	4.2%	7.0%	3306.2
Other	**SubtrSeg**	18.9%	5.1%	0.1%	1386.8
	TPminima	0.0%	0.0%	7.8%	2642.6
	RandOracle	7.6%	1.7%	3.9%	2885.0

Table 5.5: Percentage of errors that resulted in reasonable errors averaged over five train/test splits. Because not all errors are reasonable, the percentages do not sum to 100%.

There are clear differences between the various non-Bayesian learners. The SubtrSeg learner produces a good deal of real words and morphemes, but only rarely produces function word collocations. Because function word collocations are believed to persist for quite some time, this might be taken as evidence against the subtractive segmenter. The TPminima learner, on the other hand, cannot insert boundaries next to one another and therefore frequently undersegments. This produces many function word collocations for the learner, but no real word or morpheme errors. The RandOracle guesser, on the other hand, produces a mix of all of the above errors.

Of the Bayesian learners, broadly speaking the unigram learners produce more function word collocations, while the bigram learners produce more real words and morphology. This is in line with the fact that the unigram learners tend to undersegment (producing longer units), while the bigram learners are more likely to oversegment (producing shorter units). Of the online learners, the OnlineMem learner most closely matches the BatchOpt, producing fewer real word errors than the OnlineOpt and OnlineSubOpt learners but many more function word collocations. In contrast to the SubtrSeg and TPminima learners, all Bayesian learners produce some amount of errors from each category.

5.4.3 Discussion

These results indicate that while syllable-based, online learning changes certain aspects of the segmentation process, segmentation is still possible without batch inference methods. In some cases (e.g. all unigram online learners and the bigram OnlineMem learner) cognitively constrained inference actually appears to help segmentation of English. This fact is not explained by poor model fit of the BatchOpt learner, as shown by the log posterior results in Table 5.3. Instead, the largest differences between the learners come from the types of errors they produce. While all Bayesian learners tend to undersegment, online learners produce oversegmentations at a higher rate than the BatchOpt learner. Online learners also produce different amounts of reasonable errors, making mistakes that result in real words, morphemes, and function word collocations. While all of these are attested in older children (Brown, 1973; Peters, 1983), it is unclear what exact quantitative pattern should represent good model behavior. Because these results come only from English, further cross-linguistic evaluation is necessary to better understand these model trends.

5.5 Cross-linguistic Results

5.5.1 Gold Standard Results

Table 5.6 presents the gold standard word token F-score results for each learner on all seven languages. The best performing learner on each language is presented in bold. The SubtrSeg outperforms all of the Bayesian learners on both English and German, but the DPSEG-2 BatchOpt learner outperforms all other learners on every other language.

Looking at the unigram versus bigram performance, it can be seen that the bigram assumption is useful for all learners on all languages with the exception of the OnlineSubOpt learner

		Eng.	Ger.	Spa.	Ita.	Far.	Hun.	Jpn.
DPSEG-1	**BatchOpt**	53.1	60.3	55.0	61.9	66.6	59.9	63.2
	OnlineOpt	58.8	50.7	55.9	59.9	67.8	52.9	62.2
	OnlineSubOpt	63.7	63.1	54.0	60.2	65.9	54.5	61.3
	OnlineMem	55.1	60.3	56.1	58.6	59.6	54.5	63.7
DPSEG-2	**BatchOpt**	77.1	73.1	**64.8**	**71.3**	**69.6**	**66.2**	**66.5**
	OnlineOpt	75.1	75.0	61.1	67.1	69.8	62.0	64.2
	OnlineSubOpt	77.8	76.0	52.8	61.3	55.3	51.1	51.9
	OnlineMem	86.3	82.6	60.2	60.9	62.5	59.5	63.3
Baseline	**SubtrSeg**	**86.6**	**83.1**	44.9	42.2	34.1	51.5	33.5
	TPminima	13.0	16.4	28.7	32.2	31.5	25.3	32.4
	RandOracle	56.4	47.5	27.0	22.8	20.3	26.4	26.1

Table 5.6: Word token F-scores for learners across English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. Higher token F-scores indicate better performance, with the best score for each language in bold.

on Spanish, Farsi, Hungarian, and Japanese and the OnlineMem learner on Japanese. This supports the idea that the bigram assumption is generally useful, although not necessary for acceptable segmentation performance.

Indeed, the general level of performance for the Bayesian learners is quite high, with no learner scoring below 50. Thus, it appears that segmentation can successfully occur with a variety of constraints on the inference process. The non-Bayesian segmenters, however, show that this is no easy feat. The best performing non-Bayesian segmenter, the SubtrSeg learner, struggles with languages other than English and German, performing worst on Japanese and Farsi (Token F-score of 33.5 and 34.1, respectively). In fact, the SubtrSeg only outperforms a single Bayesian learner on any language other than English or German (Hungarian SubtrSeg: 51.5, DPSEG-2 OnlineSubOpt: 51.1). As described in the previous chapter, both the TPminima and RandOracle learners likewise fail cross-linguistically.

5.5.2 Error Patterns

An interesting fact about the online learners in English was that in some cases they outperformed the BatchOpt learner. This pattern does not hold cross-linguistically, and to better understand why, a more detailed error analysis was conducted.

5.5.2.1 Over- and Undersegmentation

		Overseg Errors (%)						
		Eng.	**Ger.**	**Spa.**	**Ita.**	**Far.**	**Hun.**	**Jpn.**
DPSEG-1	**BatchOpt**	1.7	9.1	8.7	39.9	47.7	45.3	39.0
	OnlineOpt	5.0	4.3	19.1	31.1	58.3	17.0	60.8
	OnlineSubOpt	6.5	7.2	25.0	45.3	70.0	24.8	67.5
	OnlineMem	9.0	15.9	25.8	53.8	68.0	55.3	53.5
DPSEG-2	**BatchOpt**	13.8	26.0	33.0	73.1	59.8	58.0	58.4
	OnlineOpt	13.2	18.8	36.8	59.3	75.4	48.2	66.6
	OnlineSubOpt	18.8	20.4	75.7	74.9	92.8	71.2	85.0
	OnlineMem	44.8	60.6	72.8	89.9	93.4	82.7	79.9
Baseline	**SubtrSeg**	88.6	85.6	96.6	99.2	99.3	96.8	99.6
	TPminima	0.2	1.1	1.2	3.7	5.6	1.4	3.8
	RandOracle	51.7	60.0	57.7	57.7	58.7	56.9	54.5

Table 5.7: Percentage of errors which resulted in an oversegmentation as compared to adult orthographic segmentation.

Table 5.7 presents the percentage of errors resulting in an oversegmentation. The results show that bigram Bayesian learners tend to oversegment more than their unigram counterparts, matching the results from English in Table 5.4. Of the non-Bayesian learners, the TPminima is the only learner which heavily undersegments. This behavior is a result of the inability of the TPminima learner to posit two boundaries next to one another. The other non-Bayesian learners tend to oversegment when they make errors, especially so for the SubtrSeg learner.

Whereas the non-Bayesian learners all tend to segment similarly regardless of language, the Bayesian learners show a wider variety of behavior across languages. For example, the

bigram BatchOpt learner tends to undersegment on English (13.8% oversegmentation), but oversegments on Italian (73.1% oversegmentation). This makes a clearly testable prediction that infants in different languages should produce different kinds of errors, whereas the non-Bayesian learners predict the opposite pattern (that segmentation errors are of the same type across these languages).

5.5.2.2 Reasonable Errors

I use the same three categories for reasonable errors as in the English-only analysis. Each language varies in the structure of their morphology and the degree to which they contain function word collocations. Examples of actual errors made by the DPSEG models on the non-English languages are given in Table 5.8.

		True	**Learner**
Real words	**Spanish**	*porque* 'because'	*por que* 'why'
	Japanese	*moshimoshi* 'hello'	*moshi moshi* 'if if'
Morphology	**Italian**	*devi* 'you must'	*dev i* 'must' 2-PL
	Farsi	*miduni* 'you know'	*mi dun i* PRES 'know' 2-SG
Func words	**Italian**	*a me* 'to me'	*ame* 'to-me'
	Farsi	*mæn hæm* 'me too'	*mænhæm* 'me-too'

Table 5.8: Examples of reasonable errors (with English glosses) made by learners in different languages. *True* words refer to the segmentation in the original corpus, while *Learner* output represents the segmentation generated by a modeled learner.

Real Word Errors

Table 5.9 presents the percentage of all errors made by each learner that resulted in at least one true word. Higher rates of these errors are potentially useful for lexical models because

		Real Word Errors (%)						
		Eng.	**Ger.**	**Spa.**	**Ita.**	**Far.**	**Hun.**	**Jpn.**
DPSEG-1	**BatchOpt**	1.0	3.3	2.8	23.7	20.1	11.9	17.7
	OnlineOpt	3.8	2.2	8.9	21.4	37.4	10.0	30.6
	OnlineSubOpt	4.5	4.0	10.5	30.3	40.8	12.4	34.1
	OnlineMem	3.4	4.5	6.3	27.6	25.5	14.9	21.4
DPSEG-2	**BatchOpt**	5.8	7.9	11.2	38.3	24.6	17.8	26.7
	OnlineOpt	10.0	7.5	14.0	38.0	45.5	30.4	33.0
	OnlineSubOpt	14.3	8.7	15.1	47.9	33.7	31.8	34.0
	OnlineMem	29.6	17.6	15.1	57.0	41.5	27.7	34.6
Baseline	**SubtrSeg**	55.0	25.4	31.3	61.2	39.9	39.0	43.5
	TPminima	0.0	0.1	0.2	0.3	0.9	0.1	0.9
	RandOracle	17.5	7.7	13.6	14.9	10.0	8.6	12.6

Table 5.9: Percentage of model errors which produced at least one true word in the corpus, excluding true words occurring fewer than five times.

they increase the perceived frequency of the true word. For example, if a model segments *alright* as *all* and *right*, then the next time the word *right* is encountered in a sentence, it is more likely to be segmented because it has been previously seen.

In English, we found that the online learners were more likely to produce real word errors than the BatchOpt, which could be explained by the increased number of oversegmentations which produced common, short words. This pattern extends across all of the examined languages although there are some counter-examples. For instance, both the unigram and bigram OnlineOpt learners sometimes produce fewer real word errors than the BatchOpt (e.g. Hungarian DPSEG-1 OnlineOpt: 10.0%, BatchOpt: 11.9%). English and German show relatively few real word errors in comparison to the other languages, which again fits with the rates of oversegmentation presented in Table 5.7

All non-Bayesian learners do produce some amount of real word errors although they vary as in English. The SubtrSeg model produces real word errors at a much higher rate than any other learner, while the TPminima learner only rarely produces real words. This pattern of results corresponds neatly with the oversegmentation results, such that learners which oversegment more are also more likely to produce real word errors.

Morphology Errors

		Morphology Errors (%)						
		Eng.	**Ger.**	**Spa.**	**Ita.**	**Far.**	**Hun.**	**Jpn.**
DPSEG-1	**BatchOpt**	0.2	2.7	2.8	3.3	5.0	2.5	9.1
	OnlineOpt	0.2	1.4	4.4	2.8	6.2	1.1	17.7
	OnlineSubOpt	0.4	1.7	7.0	4.2	7.0	1.5	18.4
	OnlineMem	0.6	4.6	7.5	4.8	7.5	3.4	10.5
DPSEG-2	**BatchOpt**	1.0	7.7	10.4	6.3	8.4	3.3	10.4
	OnlineOpt	0.6	5.5	9.3	7.6	8.5	3.0	16.7
	OnlineSubOpt	0.8	5.5	27.5	9.6	15.1	4.7	23.8
	OnlineMem	2.6	24.9	20.4	6.7	13.0	4.6	16.9
Baseline	**SubtrSeg**	5.7	22.7	33.7	10.4	20.0	7.7	31.6
	TPminima	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RandOracle	2.2	12.1	10.6	3.0	5.1	3.0	10.1

Table 5.10: Percentage of model errors which produced at least one morphological affix. In order for the error to count, the produced morpheme must have occurred in the correct location (e.g. suffixes at the end of the error, prefixes at the beginning). Lists of morphemes for each language were produced by linguistically-trained native speakers.

Table 5.10 presents the percentage of all errors which produced at least one morphological affix in the language. For each language, a list of morphemes in the corpus was generated by linguistically-trained native speakers. As with the real word errors, learners that produce oversegmentations are also more likely to produce morphology. This makes intuitive sense: since morphological affixes are smaller than words, they generally are only produced through oversegmentations. This means that the DPSEG-2 learners produce more morphemes than the DPSEG-1 learners, and that the SubtrSeg learner also produces many morphemes.

Languages also vary in the degree to which they allow for morphological segmentations. German, Spanish, Farsi, and Japanese all produce many morphemes when segmented regardless of model. This might seem surprising, since Italian and Hungarian are also morphologically complex, but is explained by the fact that Italian and Hungarian have fewer syllabic morphemes. Because the syllable-based learners treat each syllable as atomic, they are incapable of segmenting sub-syllabic morphemes such as the English plural *-s*. This means that certain

types of morphology are not learnable until a learner is using phonemes as a basic unit of representation, which may be after segmentation has already begun.

Function Word Errors

		Function Word Errors (%)						
		Eng.	**Ger.**	**Spa.**	**Ita.**	**Far.**	**Hun.**	**Jpn.**
DPSEG-1	**BatchOpt**	8.8	27.2	8.9	6.4	4.3	2.3	6.9
	OnlineOpt	6.1	16.5	3.3	1.2	0.4	1.6	1.8
	OnlineSubOpt	7.4	17.4	2.8	1.0	0.2	1.5	1.9
	OnlineMem	10.2	26.7	7.7	5.8	3.1	2.3	7.7
DPSEG-2	**BatchOpt**	15.7	28.2	6.4	5.8	3.7	2.9	5.2
	OnlineOpt	9.7	17.8	2.5	1.6	0.1	0.7	2.7
	OnlineSubOpt	8.9	14.7	2.3	0.6	0.3	0.8	0.7
	OnlineMem	9.9	10.8	2.3	1.1	0.2	1.7	2.6
Baseline	**SubtrSeg**	0.2	2.2	0.1	0.0	0.0	0.0	0.0
	TPminima	4.9	18.2	10.9	9.1	2.6	2.3	5.8
	RandOracle	3.6	8.7	4.2	3.4	1.0	1.1	2.1

Table 5.11: Percentage of model errors which produced a function word collocation. Lists of function words for each language were developed by linguistically-trained native speakers for each language.

The last type of reasonable error are function word collocations. The percentage of errors which were made up entirely of joining together function words are presented in Table 5.11.

As with the other reasonable errors, the degree to which a learner produces this error is related to their over- and undersegmentation preference. Because function word collocations are produced through undersegmentation, their appearance is somewhat the opposite as that found for other reasonable errors. Unigram Bayesian learners tend to produce more function word collocations than their bigram counterparts (although there are exceptions, notably the bigram English BatchOpt). The SubtrSeg only rarely produces function word collocations, but the TPminima produces many.

Looking across languages, English and German both produce many function word collocations regardless of learner. The other languages possess fewer collocations, possibly because

they also have a more complex morphology which reduces the need for strings of function words.

Discussion

The results of the reasonable error analysis clearly indicate that many of the errors made by each learner are potentially less harmful than would otherwise appear. A large percentage of the errors made by Bayesian learners, for instance, are either potentially useful for later segmentation (e.g. real word errors), useful for learning morphology (e.g. morpheme errors), or are in accord with evidence from diary studies (e.g. function word collocations).

While the non-Bayesian learners also produce many of these errors, they tend to focus on only one or two types of errors. For example, the SubtrSeg learner produces many real words and morphemes, but almost never produces function word collocations. Also, while the non-Bayesian learners are consistent across all languages, the Bayesian learners shift the types of errors they make depending on the language. That is to say, on languages with lots of syllabic morphology, the Bayesian learners produce many morphological errors, while on languages with many strings of function words they produce function word collocations. This matches what one might expect from rational infant learners.

One advantage to the reasonable error analysis is that it makes detailed, quantitative predictions about the kinds of errors that infants should be making. These types of errors might be elicited in future experimental studies to provide confirmation as to which type of model best fits the errors made by actual children.

5.6 Conclusion

Chapter 4 showed that syllables were a valid unit of representation for Bayesian segmentation, and the results of this chapter have shown that Bayesian segmentation is successful regardless of the type of inference used to produce a segmentation. Because children possess many kinds of cognitive constraints, this evidence suggests that Bayesian segmentation may well be an appropriate tool for modeling the beginnings of the segmentation process.

The various online learners evaluated in this chapter generally produce lower word token F-scores than the batch inference method, but the online learners also possess other interesting qualities. Online learners produce more oversegmentation errors, resulting in a greater rate of errors involving real words and morphemes. Future experimental work might shed light on whether this is reflective of the errors made by actual children. The online learners also vary in the types of errors they make on different languages, providing yet another point against which experimental evidence might be compared.

Chapter 6

Model Evaluation in Speech Segmentation

The goal of any model is to determine whether or not a theory of acquisition is able to explain empirical data. This requires, of course, a judgment regarding whether the model has succeeded in this attempt. Previous work has largely focused on the ability of the model to recreate a gold standard segmentation, typically based off of orthographic standards (Brent, 1999; Goldwater et al., 2009; Blanchard et al., 2010; Pearl et al., 2011; Lignos, 2011). However this represents only one method for measuring success. The work in Chapters 4 and 5 explored how model input and inference can be improved to make models of speech segmentation more meaningful. In this chapter, I will explore alternative methods for analyzing the success of models of speech segmentation.

6.1 Importance of Model Evaluation

In order for any model of language acquisition to have any impact on our understanding of language learning, the model output must be examined to measure how closely the model mimics data from actual acquisition. Model output can be measured in a number of different ways which broadly split into two categories, intrinsic and extrinsic (Galliers and Jones, 1993). Intrinsic evaluations measure the model output with regard to the model's direct objective, i.e. the task for which it was trained. This would include measures made against a gold standard, such as F-score, as well as measures of model fit such as log posterior probability. Extrinsic evaluations measure how well the model output performs in relation to alternative tasks.

To make this distinction clearer, consider a company trying to predict what items a customer will enjoy based on previous ratings. The company trains the model on some subset of the data and then tests the model on a withheld portion. The company could evaluate the model based on intrinsic measures by measuring the similarity between the model predictions and the actual withheld ratings (a gold standard measure). The company might also measure the degree to which the model found parameters which optimized the solution (a model fit measure). Both of these measures, however, are indirectly measuring what the company cares about, which is whether the model will improve their customer satisfaction. To measure this, the company needs an extrinsic evaluation, perhaps by implementing the model for a subset of users and creating a survey or measuring how much customers spent. Depending on the goal of the modeler, intrinsic or extrinsic evaluations might be of greater or lesser importance, although it is generally wise in practice to examine both.

Turning to segmentation, intrinsic measures evaluate the degree to which the segmenter recreated the gold standard segmentation (typically based off of orthography). This can either be done by measuring things such as precision, recall, and F-score, but can also be

done by measuring model fit, the ability of the model to explain the training or test set. Just as with the example of the company, intrinsic measures capture important information for the cognitive modeler, but the end goal of segmentation is not simply to insert boundaries between words. Instead, the goal is to produce lexical items which will aid in the future learning tasks which infants face. Rephrased, the goal is not to segment, the goal is to learn a native language. Researchers must turn towards extrinsic measures in order to determine how well a model of segmentation fits within the greater language learning problem.

6.2 Intrinsic Evaluation

6.2.1 Gold Standard Analysis

Comparing model output to a gold standard has long been the most commonly practiced method of model evaluation. A researcher determines *a priori* what the output of the model should be and then measures how well that output was achieved. Depending on the task, this can be measured in many different ways, but in speech segmentation the most commonly practiced are precision, recall, and F-score as measured primarily over word tokens (Brent, 1999; Blanchard et al., 2010; Johnson, 2008; Johnson et al., 2010; Johnson and Demuth, 2010; Fourtassi et al., 2013).

Such a direct comparison is not necessarily the only gold standard measurement which can be made. Models can be measured in terms of their ability to predict seen or unseen data. In some cases there can be no direct mapping between model output and the gold standard. Consider the problem of identifying the sounds in a language. This can be thought of as a clustering problem over acoustic space, but one of the difficulties is that the child does not know *a priori* just how many clusters, or sounds, there should be. Attempting to directly map model clusters onto the gold standard is a poor metric of model performance and instead

researchers might need to use alternative information-theoretic metrics such as variation of information or V-measure (Rosenberg and Hirschberg, 2007).

Given that gold standard results were widely presented in Chapters 4 and 5, they will not be further discussed here.

6.2.2 Measures of Model Fit

An alternative to direct measures against a gold standard are measurements of model fit. This refers to the ability of the model to accurately capture the data it was given. For Bayesian models, this is generally measured in the form of the posterior probability which measures how well the model can explain the data. If the model's segmentation fits its underlying generative model well, then the model will have a large posterior probability, indicating good model fit. On the other hand, if the model's segmentation is not well explained by the generative model the posterior probability will be very low, indicating poor model fit. For models without a likelihood function, such as the heuristic models previously examined, there is no way to measure the probability of the data. This means that model fit cannot be measured in these models and gold standard measures must be relied upon to indicate performance.

An advantage of measuring model fit is that it can be done without concern for the accuracy of the gold standard. However, in the case of speech segmentation, model fit does not necessarily indicate a successful segmentation. Because our Bayesian learners operate with either a unigram or a bigram assumption, neither generative model is actually indicative of the true language model for any of the languages evaluated. If the model is able to very closely fit a bigram model, that does not mean it has closely fit the true language model. In this way, a model can have very good log posterior, while still having poor gold standard performance. In fact, we see this dissociation between model fit and gold standard

performance in English, where some online learners show poorer model fit than the BatchOpt but achieve higher word token F-scores.

6.2.3 Comparison to Experimental Results

A third alternative is to compare the results of the model to experimental results. In many cases, this requires modeling the input of the experiment and measuring the output in some manner similar to the experimental results (for a good example of this technique, see Frank et al., 2009 and Kolodny et al., 2015). For example, if an experiment reports a certain percent accuracy, the model might be compared in its ability to recreate that same accuracy. More generally, however, qualitative comparison can be made, examining the model's ability to capture the general trends of the experimental results. This is necessary because for experiments which report measures such as reaction time, there may not be a good method for comparing model results quantitatively given that many models do not simulate the entire decision making process and therefore do not make predictions that would match those from experimental studies.

In the case of speech segmentation, only rarely have models been compared based on their ability to capture experimental results. Frank et al. (2010) find that both the unigram DPSEG model and a Bayesian TP model are able to reasonably capture experimental data from adult segmentations which varied in sentence length, amount of exposure, and number of word types. Pearl et al. (2011) likewise show that the various DPSEG learners are in many cases capable of replicating the finding of Seidl and Johnson (2006) that infants more easily segment words that appear utterance-initially or utterance-finally. Unfortunately for modelers of speech segmentation, the available experimental evidence from infants is relatively scarce, limited to a small number of experiments, most of which are readily solved even by simple TP learners (Kolodny et al., 2015).

6.3 Extrinsic Evaluation

Looking at all three types of intrinsic measures, it becomes clear that there is no one intrinsic evaluation which is optimal for speech segmentation. Gold standard measures are influenced by arbitrary orthographic standards and do not reflect the knowledge of young infant segmenters. Indicators of model fit measure how well the segmentation fits the language model used by the learner rather than how well it matches the target language. Relevant experimental evidence is often difficult to obtain and in many cases experimental results can be replicated even by models which poorly segment actual languages.

Because of this, it makes sense to widen the number of evaluation metrics to ensure that no single metric has undue sway in determining the success of the model. An alternative is extrinsic evaluation, using the output of speech segmentation in order to perform secondary tasks. This can be done in one of two ways, either by modeling segmentation along with another task (joint modeling), or by using the output of the segmentation process to perform the second task in isolation (downstream evaluation).

6.3.1 Joint Modeling vs. Downstream Evaluation

Increasingly, there has been interest in exploring the degree to which modeling two tasks jointly can aid in the learning of both tasks (Feldman et al., 2009; Blanchard et al., 2010; Jones et al., 2010; Dillon et al., 2013; Doyle and Levy, 2013). Indeed, joint modeling often has the benefit that the two tasks can bootstrap off of one another, resulting in performance which is higher than in either individual task. Because of the popularity of this method, one might wonder why downstream evaluation, which is unable to take advantage of these task synergies, would even be considered.

When decisions are made with regard to the training or structure of a model, these decisions necessarily entail assumptions about the learning problem. If a model is trained on phonemes, it assumes the learner represents speech in terms of phonemes. If a model is trained in batch, it assumes either that the learner is capable of batch inference or the modeler is forced to admit that the model is searching for an optimal solution to the problem rather than attempting to recreate a child learner. In this same way, modeling two tasks jointly assumes that the learner solves both tasks at the same time.

There are two main reasons why this may not be an appropriate strategy. First, there may be some delay between the two tasks. For instance, if segmentation begins before infants start to learn the phonotactics of their language, then jointly modeling segmentation and phonotactic learning is inappropriate. Although there are synergies between the two tasks, it would appear that infants do not make use of them initially. A joint model in this case may underestimate how difficult the task of segmentation is because it is learning with information that infant learners do not have access to.

Second, when creating a joint model, decisions must be made regarding how the two tasks interact with one another. By connecting two tasks into a single model, one assumes that the learner likewise knows *a priori* how the two tasks are related and is able to take advantage of that during learning. This may or may not be a valid assumption given the tasks at hand and the way in which they are linked. Imagine a joint model of segmentation and stress pattern learning. If the model assumes that words are generated and each word takes on some fixed stress pattern such that ['da.gi] and [da'gi] are treated as separate words, then the modeler assumes that infants expect their language to have lexical stress. Many languages of the world do not possess this property (e.g. French, Yoruba, Bella Coola) and deviations in stress pattern do not indicate a change in meaning. If the joint model's assumptions are correct, then we should expect to find evidence that children learning languages without lexical stress display an early bias towards assuming its existence.

Downstream evaluation, on the other hand, represents the assumption that the first task is more or less solved before the second task even begins. This would be more appropriate in cases where there is some gap between tasks, as with segmentation and phonotactic learning. It may also be appropriate for cases where infants do not appear to take advantage of the synergies between two tasks either because the infant does not know how to connect the tasks or because they are learned entirely separately.

In a sense, joint modeling represents a *best* case scenario for the learner. The learner knows that the two tasks are connected and knows specifically how they are related, allowing for joint learning to take advantage of the two processes. Downstream evaluation, on the other hand, represents a *worst* case scenario. The learner does not realize the two tasks are connected and instead has to solve the first on its own and then solve the second on its own without ever realizing how to connect the two. In reality, neither scenario may be true for infant learners, but by utilizing both approaches researchers can set upper and lower bounds on success, developing a richer picture of how useful a particular learning strategy is to the infant learner.

6.3.2 Stress Pattern Induction

As discussed in Chapter 2, stress patterns are learned at some point directly after segmentation has begun. This naturally raises the question: Are stress patterns learned at the same time as words are being segmented, or do children only begin to make use of stress patterns after an early segmentation strategy has already created a seed pool of words from which to learn? Unfortunately, this is an area where experimental evidence has very little to say.

Certainly, stress cues seem to be utilized first at 7.5 months (Jusczyk et al., 1999b) while segmentation clearly begins before then (Bortfeld et al., 2005). This could be viewed as being in favor of the downstream model, but it does not rule out the possibility of joint

stress learning and segmentation. It may be the case, for instance, that stress patterns are being learned by younger infants, but that the strength of those cues does not, for whatever reason, provide sufficient evidence to produce a stress-based segmentation under experimental settings. For this reason, it makes sense to investigate the ability of various segmentation strategies to succeed in producing appropriate stress patterns either with a joint or downstream approach.

6.3.2.1 Joint Model

In the best case scenario, infants are already aware that they may be learning a language with predictable stress patterns and know that stress patterns play a role in segmentation. The only question for the learner is what type of stress patterns exist in the language they are learning. This scenario is explored by the work of Doyle and Levy (2013). They extend the syllable-based DPSEG-2 model by incorporating the assumption that when a word's syllables are generated, a stress pattern is also generated. The probability of a stress pattern is conditioned on the total number of syllables in the word, such that 2-syllable words have a stress distribution independent of 3-syllable words.

$$P_0(w_i, s_i) = P_W(w_i)P_S(s_i|M) \tag{6.1}$$

This is incorporated into the equation for P_0 in the bigram model as shown in Equation 6.1. P_W represents the probability of the syllables within a word which is the same as in the DPSEG model described earlier, while P_S represents the probability of a stress pattern s_i given the total number of syllables M. P_S is calculated as a multinomial over all possible stress patterns of a given length with parameter values based on the observed frequency with

plus-one smoothing. Stress patterns are not restricted in any way, such that a word might possess multiple stressed syllables or might have no stressed syllables at all. The model incorporates a uniform prior over all possible stress realizations.

Doyle and Levy (2013) run their model on the Korman corpus (Korman, 1984) as modified by Christiansen et al. (1998). This corpus is made up of CDS aimed at infants aged 6-16 weeks, totaling 24493 words. Phonemic forms, syllabification, and stress patterns were all taken from the MRC Pscyholinguistic Database (Wilson, 1988). As with all corpora of CDS in English, the corpus is heavily monosyllabic (87.3%) with longer words possessing a strong word-initial bias (89.2% for all multisyllabic tokens). No token in the corpus has more than three syllables.

	T.P.	**T.R.**	**T.F.**	**B.P.**	**B.R.**	**B.F.**	**L.P.**	**L.R.**	**L.F.**
Without Stress	76	60	67	99	69	82	72	84	77
With Stress	76	**61**	**68**	99	**70**	82	**75**	**87**	**80**

Table 6.1: Precision (P), recall (R), and F-score (F) for word tokens (T), boundaries (B), and lexicon items (L) from the simulations of Doyle and Levy (2013) both with and without stress.

Their results, given in Table 6.1, suggest that incorporating stress into the model does improve performance, although not drastically (Token F-score increases by 1 point, Lexicon F-score by 3 points). Because lexicon scores increase more so than token scores, this indicates that the model's performance is improved primarily on rarer words. Intuitively, this makes sense because the model has little incentive to segment words it has rarely seen, but if the word fits a predominant stress pattern, that provides additional weight to its segmentation. The model is also able to learn the word-initial stress preference of English, segmenting word-initially stressed bisyllabic words (as compared to word-finally stressed bisyllabic words) at a ratio of 6.77:1, which is a slight underestimate of the true ratio in the corpus (7.86:1).

This seems to suggest that a joint Bayesian learner will easily be able to begin segmentation and quickly identify the basic stress patterns of their language. The model assumes, however,

that any word has one and only one stress pattern, and that deviations of that stress pattern represent distinct words. As mentioned above, this works well on English which is a language with lexical stress, but not all languages function this way (e.g. French, Yoruba, Bella Coola; Hyman, 2010). If infants do assume lexical stress, then this should be apparent in difficulties with learning words in languages without lexical stress, although I'm aware of no evidence demonstrating this is the case.

This fact sheds some doubt on the ability of the joint model to accurately capture the learning trajectory of infants. Doyle and Levy (2013) also discuss a further difficulty of the model, namely that it fails to account for the results of Thiessen and Saffran (2003). In this work, Thiessen and Saffran (2003) show that infants at 7 months are biased to segment according to TP cues, while infants at 9 months are biased to segment according to stress cues. Presumably, by 9 months infants have noticed the fact that English has a strong word-initial stress bias roughly of the ratio 8:1 while the younger infants have yet to discover this bias and therefore rely on TP cues alone. If the joint model is initialized with no stress bias and trained on the same stimuli as in Thiessen and Saffran (2003), then it replicates the finding for younger infants that TP cues are preferentially used. On the other hand, if the model is initialized with the true English stress bias, it fails to replicate the fact that infants prefer to segment using stress. In fact, the bias towards word-initial stress must be orders of magnitude higher before the model will prefer stress-based over non-stress-based cues (Doyle and Levy (2013) report that stress cues are only utilized at biases of 10,000:1 or stronger).

This fact could be explained in one of three ways. First, perhaps it is the case that infants actually do possess a stronger bias than their native language would otherwise suggest. This explanation seems unlikely given that infants would need such a strong bias and that the direction of the bias would be language-dependent; therefore it could not be known innately. Doyle and Levy (2013) also propose that perhaps infants learn from much less data than we might otherwise expect. This too seems unlikely as the amount of data a child is exposed to

is much greater than any current model is trained on. Finally, the most likely alternative is simply that the joint model is in some manner incorrect, possibly because of the generative function or possibly because the joint model is otherwise inappropriate.

6.3.2.2 Downstream Evaluation

While the joint approach represents a possible best-case scenario for learning, it is also important to identify a worst-case scenario. In this case, speech segmentation occurs entirely in the absence of stress information, and only at a later point do infants realize they should pay attention to the regularities of stress within words. In this way, none of the synergies between stress cues and segmentation can be taken advantage of, which may be more appropriate for modeling very early stages of speech segmentation.

This process can be mimicked first by running the segmentation model as before, and then using the resulting words in order to identify the stress cue of the language. By measuring the resulting stress patterns, we can identify not only whether these stress patterns are learnable in the worst-case scenario, but we also can indirectly measure the quality of each segmentation.

Because the UCI Brent Syllables corpus does not mark stress, I made use of the English Callhome Lexicon (Kingsbury et al., 1997) to identify the main stress in words. For child-register words not found in standard dictionaries (e.g. *moosha*), I manually coded the stress when the proper stress was known. If a word was not familiar enough to be confident about its stress pattern (e.g., *bonino*), it was ignored for the purposes of this analysis. All words in the analyses presented below were given their dictionary stress patterns. In order to better approximate the stress of actual utterances, monosyllabic words were left unstressed. This approximation technique is necessary because conversational speech is not spoken using dictionary stress on every word.

Lexical items, rather than tokens, were used to measure the stress patterns identified by each learner. This follows from the fact that young learners appear to ignore frequency in making generalizations (Yang, 2005; Perfors et al., 2014). Table 6.2 presents the percentage of bisyllabic words with a single stress which had either word-initial stress (SW) or word-final stress (WS)[1]. As can be seen, of bisyllabic words with exactly one stress, 88.4% are word-initially stressed in the gold standard corpus, roughly in line with previous results (Doyle and Levy, 2013). All of the Bayesian learners recreate the stress pattern, ranging from 85.3% word-initial (unigram OnlineMem) to 90.6% initial (bigram OnlineMem). The SubtrSeg learner does almost as well as the Bayesian learners, producing 81.4% of bisyllabic types with word-initial stress, while the RandOracle fails to capture the pattern, achieving only 46.7% word-initially stressed.[2] Although the Bayesian learners show the highest quantitative match with the true proportions, it is unclear whether children achieve this level of performance as well. Without experimental evidence to indicate what level of performance is desired, it may suffice to show a bias of any kind towards the SW pattern. If the goal is only to match this qualitative pattern, then both the Bayesian learners and SubtrSeg learner perform equally well.

Interestingly, it is not the quality of the lexicon, in terms of F-score, which drives these differences. In fact, although the Bayesian learners vary wildly in terms of their lexicon F-scores (43.9 - 75.5), they all obtain roughly the same word-initial stress bias. Interestingly, although the unigram OnlineOpt produces a lexicon with roughly the same F-score as the RandOracle (43.9 vs. 41.9, respectively), they perform quite differently in reproducing the word-initial stress pattern (89.6% vs. 46.7%, respectively).

This seems to indicate that even in a worst-case scenario, the bisyllabic stress pattern of English is learnable so long as the learner is using a reasonable segmentation strategy. The

[1]I denote stressed syllables using the character S (strong) and unstressed syllables using the character W (weak).

[2]Note that results are not presented for the TPminima learner, given that it fails to successfully segment on any of the languages previously examined.

		SW	WS	Type F-score
	Adult Seg	**88.4%**	**11.6%**	**1.0**
Unigram	**BatchOpt**	87.3%	12.7%	56.2
	OnlineOpt	89.6%	10.4%	43.9
	OnlineSubOpt	88.8%	11.2%	48.5
	OnlineMem	85.3%	14.7%	54.5
Bigram	**BatchOpt**	88.4%	11.6%	71.9
	OnlineOpt	89.0%	11.0%	64.3
	OnlineSubOpt	89.0%	11.0%	65.2
	OnlineMem	90.6%	9.4%	75.5
Other	**SubtrSeg**	81.4%	18.6%	74.7
	RandOracle	46.7%	53.3%	41.9

Table 6.2: Stress pattern results for all learners on bisyllabic word types. Percentages are calculated out of all bisyllabic word types identified by each learner with exactly one stressed syllable.

question remains, however, as to whether this holds for other languages. To answer this, I trained the same models on all of the cross-linguistic corpora for which stress information was readily available, in this case German and Hungarian. As with English, stress was taken from the dictionary forms and then removed from monosyllabic items. The results are presented in Table 6.3.

		English SW (%)	German SW (%)	Hungarian SW (%)
	Adult Seg	**88.4%**	**90.3%**	**100%**
Unigram	**BatchOpt**	87.3%	90.8%	96.9%
	OnlineOpt	89.6%	89.5%	98.2%
	OnlineSubOpt	88.8%	90.4%	97.7%
	OnlineMem	85.3%	87.6%	93.1%
Bigram	**BatchOpt**	88.4%	90.9%	97.9%
	OnlineOpt	89.0%	91.2%	98.2%
	OnlineSubOpt	89.0%	91.4%	95.7%
	OnlineMem	90.6%	92.6%	96.4%
Other	**SubtrSeg**	81.4%	83.8%	89.1%
	RandOracle	46.7%	49.6%	52.5%

Table 6.3: Stress pattern results for all learners on bisyllabic word types across various languages. Percentages are calculated out of all the bisyllabic word types identified by the model.

As can be seen, the results are consistent across all three languages. Each language has a word-initial stress bias which is captured best by the Bayesian learners, with relatively little variation between specific learners. The SubtrSeg performs somewhat worse, but still recreates the pattern, while the RandOracle learner fails to capture the pattern, performing best on Hungarian (52.5%). As with the English results, the findings for German and Hungarian support the notion that the Bayesian learners best fit the corpus quantitatively. Since we do not know what level of performance infant learners evidence, qualitative fit may be a better indicator of success, in which case all learners succeed except for the RandOracle.

6.3.2.3 Discussion

The results both from joint modeling and from downstream evaluation suggest that identifying the basic stress pattern of a language is a relatively simple task, especially for Bayesian learners. Interestingly, stress patterns can be readily identified even when segmentations are created entirely without reference to stress. It remains an open question, however, as to exactly how infants combine segmentation with stress pattern learning. The most straightforward joint model is unable to explain how infants go from possessing a bias to segment with statistical cues to eventually preferring to segment using stress patterns (Thiessen and Saffran, 2003). The downstream evaluation method, on the other hand, has nothing to say in regards to how infants might integrate these two cues.

Although these questions are of great importance to identifying the greater process of speech segmentation, they also shed light on the various learning strategies for speech segmentation without stress. For instance, while the Bayesian learners vary quite considerably in terms of the quality of their lexicons, they all achieve roughly the same stress bias, indicating that perhaps even the lower quality lexicons are still "good enough". These results indicate a dissociation between the gold standard measures and the quality of the segmentation in

terms of its ability to support future learning, suggesting that future research might also benefit from investigating the quality of model output in this same way.

6.3.3 Word-Object Mapping

An alternative task which proceeds at the same time as speech segmentation is word-object mapping (Tincoff and Jusczyk, 1999, 2012; Bergelson and Swingley, 2012). This process refers to the ability of learners to begin associating word forms with concrete objects in the environment. Word-object mapping is a well-studied area of developmental psychology, where well-known phenomena such as fast-mapping (Carey, 1978; Markson and Bloom, 1997), the shape bias (Landau et al., 1988), and mutual exclusivity (Markman and Wachtel, 1988; Markman, 1989; Markman et al., 2003) have been investigated. An additional avenue of research in word learning has been motivated by the insight that words and their referents often co-occur in space and time. This phenomenon leads to what is known as cross-situational word learning, which is learning the meaning of a word as it is encountered in various situations (Yu and Smith, 2007; Smith and Yu, 2008; Yu and Smith, 2011).

Explaining how these various lines of research come together into a single process of word learning has been a great challenge for researchers over the last three decades. Recently, Frank et al. (2009) proposed a Bayesian model of word learning which was shown to account for effects across a wide range of experiments. In their model, infants are assumed to know what words are spoken in a sentence (W_s) as well as what objects (O_s) are in the vicinity and have to infer what items (I_s) the speaker intends to talk about as well as the adult lexicon (L) which maps the intended referent to a spoken word. A plate diagram for the model can be seen in Figure 6.1.

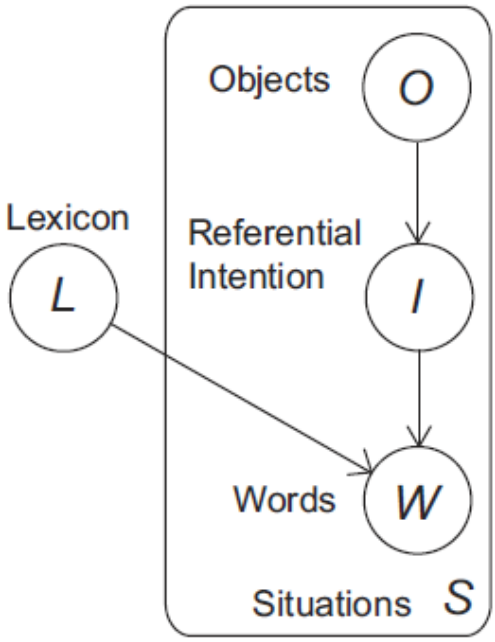

Figure 6.1: Plate diagram of the Frank et al. (2009) word-object mapping generative model.

Equation 6.2 describes the probability of a corpus of utterances given an adult speaker's lexicon:

$$P(C|L) = \prod_{s \in C} \sum_{I_s \in O_s} P(W_s|I_s, L) P(I_s|O_s) \tag{6.2}$$

The probability of intending to speak about any object, $P(I_s|O_s)$, is treated as uniform, such that all objects are equally likely to be referred to. The model is then driven by the probability of emitting a word given the speaker's intentions and the lexicon. The equation

for this term is defined below and is split into two portions, one in case the spoken word is referential and a second for the case that the word is non-referential.

$$P(W_s|I_s, L) = \prod_{w \in W_s} [\gamma \sum_{o \in I_s} \frac{1}{I_s} P_R(w|o, L) + (1 - \gamma) P_{NR}(w|L)] \tag{6.3}$$

The probability of using a word referentially, $P_R(w|o, L)$, assumes that the word is chosen uniformly from all words linked in the lexicon to the particular object. For instance, if *bear* is the only word linked to the object BEAR, then it has a probability of 1. If both *cat* and *kitty* are linked to the object CAT, then each has a probability of 0.5.

A word can be emitted non-referentially with probability $P_{NR}(w|L)$ regardless of whether it has a referential meaning (e.g. compare the referential noun *duck* to its non-referential verb *to duck*[3]). If the word is not in the referential lexicon, then it is emitted with probability proportional to the constant 1. If the word does exist in the lexicon it is emitted non-referentially with probability proportional to K, a free parameter of the model[4]. If K is less than one, then referential words are less likely to be used non-referentially than words that aren't in the lexicon.

Because this model has the nice property that it captures wide-ranging phenomena such as cross-situational word learning, mutual exclusivity, and fast mapping, there is strong support that such a model might reflect basic processes in early word learning. This model assumes, however, that children have already solved the segmentation problem and therefore begin with the gold standard segmented corpus. As with stress pattern learning, segmentation can be incorporated into the model either jointly or through a downstream approach. By inves-

[3]In the context of the Frank et al. (2009) model, non-referential simply means that a word is used without referring to a concrete noun currently visible to the infant.

[4]Although K does not appear in equation 6.3, K exists within the function $P_{NR}(w|L)$.

tigating a joint approach, we can establish a best-case scenario for segmentation, since the use of referential words would be a boost to the segmentation process. Likewise, evaluating a downstream approach would allow for the utility of various segmentations to be analyzed without the pitfalls of a gold standard analysis, establishing a lower-bound on segmentation performance.

6.3.3.1 Joint Model

One approach to jointly modeling segmentation and word-object mapping is explored by Jones et al. (2010). In this paper, the authors create a topic model (Blei et al., 2003) which integrates the DPSEG-1 model with a topic-based adaptation of the Frank et al. (2009) model. The model assumes that every utterance has a topic, which refers to a particular object in the environment. Each word is then generated either from a referential, topic-specific distribution, or from a shared, non-referential distribution. The probability of a word from either of these distributions is defined by the DPSEG-1 model with α_0 used for referential items and α_1 used for non-referential items.

The model differs from Frank et al. (2009) in that it assumes there may be only one referent, or topic, per utterance. The model is trained on the Fernald-Morikawa corpus (Fernald and Morikawa, 1993) which was phonemically transcribed using the VoxForge dictionary. Note that although segmentation is done over phonemes rather than syllables, the model achieves segmentation performance very similar to the syllable-based DPSEG-1 BatchOpt learner (Joint model token F-score: 54, DPSEG-1 BatchOpt: 53.1).

The results of their study are given in Table 6.4. Their findings suggest that if infants know how to join the problems of segmentation and word-object mapping, then neither problem appears to be insurmountable. However even with the gold standard segmentation, the model still performs much more poorly than the original Frank et al. (2010) model. To

	Lex. P	Lex. R	Lex. F	Acc.
Joint	**21**	45	28	**0.81**
Gold Std.	**21**	**60**	**31**	0.79
Frank et al. (2010)	67	47	55	0.83

Table 6.4: Performance of the joint word-object mapping algorithm, as well as the mapping algorithm using the gold standard segmentation, as reported by Jones et al. (2010). Accuracy measures the proportion of utterances where the topic was correctly recovered by the model. Results from the original Frank et al. (2010) model are presented for comparison.

some degree, this comparison, particularly in terms of topic accuracy, is unfair, given that the topic model assumes a single topic per utterance, while the previous model allowed for multiple. At the same time, this suggests that the Frank model is picking up on information which is lost to the topic model. Even with this relatively poor performance, however, the model represents a possible entry point into learning word meanings for young infants who are still in the early stages of speech segmentation.

6.3.3.2 Downstream Evaluation

Unfortunately, creating a joint model for other segmentation models is not so straightforward. The joint model of Jones et al. (2010) was forced to choose a relatively simple segmentation model (DPSEG-1) in order to make it fit with the word-object mapping algorithm, which was also based on a unigram distribution. Changing either portion of the joint model would likely require major revisions to the entire generative model. Further, the joint approach works best for Bayesian models. Combining a model such as the SubtrSeg learner with word-object mapping would require an entirely novel model. In order to better determine the ability of various segmentation models to support word-object mappings, it makes sense to investigate the worst-case scenario through a downstream evaluation.

In order to set up the downstream process, we need to have each learner segment the training corpus for the Frank et al. (2009) mapping model. We first train each segmenter on the

entirety of the UCI Brent syllables corpus, the same English corpus used in Chapters 4 and 5. Once the models have been trained, we test each model on the corpus used by Frank et al. (2009), a portion of the Rollins corpus hand-annotated for objects in the visual field (Rollins, 2003; MacWhinney, 2000). This test set is a small corpus of 619 utterances. The Brent and Rollins corpora are also of different types, the Brent corpus containing naturalistic CDS while the Rollins corpus was recorded during an experimental study where caregivers were asked to play with their children. In spite of these differences, segmentation performance transfers well onto the new Rollins corpus (see Table 6.5).

Looking at the results from the Rollins corpus in Table 6.5, we can see that in terms of word token F-score, the previous findings on the UCI Brent syllables corpus are replicated. A slight drop in performance is to be expected when transferring to a new corpus, but the general pattern of results does not change. I then examine the results of the word-object mapping model. The model produces a lexicon pairing objects with individual words. The goal of the model is not to find all possible lexicon pairings, because it is not assumed that young learners know the meanings of many words. In fact, certain principles of word learning such as mutual exclusivity (Markman and Wachtel, 1988; Markman et al., 2003) would keep infants from learning all possible lexicon pairings (e.g. both *rabbit* and *bunny* can both refer to the object RABBIT). Because of this, we compare the models in terms of their lexicon precision, which measures only how often a hypothesized mapping was correct.

These results show first that, as might be expected, no segmentation produces better results than the gold standard. The gold standard lexicon precision may appear somewhat low (58.3%), but this is still good performance in comparison to other, similar models (e.g. the Jones et al. (2010) model (21%)). Although the DPSEG segmenters do not produce word-object mapping performance as high as the gold standard, they still perform very well. The worst Bayesian learner for word-object mapping is the DPSEG-2 OnlineMem learner (38.8%), while the highest performing learner is the DPSEG-1 OnlineSubOpt learner (57.8%) which

		Segmentation			Mapping
		Brent	Rollins		
		Token F	Token F	Overseg.	Lex. P
	Adult Seg	**1.0**	**1.0**	**NA**	**58.3**
Unigram	**BatchOpt**	53.1	51.4	17.5%	46.5
	OnlineOpt	58.8	56.1	25.8%	48.5
	OnlineSubOpt	63.7	57.6	34.2%	57.8
	OnlineMem	55.1	52.4	17.0%	47.3
Bigram	**BatchOpt**	77.1	74.6	38.2%	54.4
	OnlineOpt	75.1	71.0	50.6%	55.2
	OnlineSubOpt	77.8	74.4	53.9%	43.0
	OnlineMem	86.3	81.3	74.5%	38.8
Other	**SubtrSeg**	86.6	83.3	95.3%	33.6
	RandOracle	56.4	57.6	49.9%	40.6

Table 6.5: Word-object mapping and segmentation results for Bayesian and non-Bayesian learners. Token F-scores are reported for segmentations both on the original Brent corpus and the Rollins corpus. Oversegmentation rates on the Rollins corpus are also given, as well as lexicon precision for the word-object mapping model trained on each learner's segmentation.

almost matches the gold standard performance. This behavior is interesting in that it shows a dissociation between word token F-score, where the DPSEG-2 OnlineMem performs highest, and lexicon precision, where it performs the lowest. Similarly, the DPSEG-1 OnlineSubOpt learner is outperformed on word token F-score by every DPSEG-2 learner, but still outscores them in lexicon precision. The worst mapping performance comes from the SubtrSeg model (33.6%), which is outperformed even by the RandOracle learner (40.6%).

What explains this behavior? Because the word-object mapping model produces only a small number of mappings (approximately 20-30), they can be investigated directly. There are two major types of errors. The first are driven by correlations between words and objects likely made because of the small corpus size (e.g. *bottle* - BEAR, *mmhmm* - HAND). The most common other type of error involves oversegmentations. Examples of these errors are given in Table 6.6. Many words for objects in CDS have two syllables (e.g. *doggie*, *piggy*, *bunny*) and if these words are segmented into two separate words, then the model often maps these sub-words onto the otherwise proper object (e.g. *do* and *ggie* mapped onto the

object DOG). In cases where both sub-words are mapped onto the same object, this does not appear particularly harmful. One can imagine a scenario where a child might recover from such an error by noticing the frequency with which the two words co-occur together. In other cases, however, only a single sub-word is mapped onto the correct object. For example, the subtractive segmenter splits the word *birdie* into two words, *bir* and *die*, but only *bir* is mapped onto the object DUCK. This represents a confusing scenario for learners. If they encounter the syllable *bir* elsewhere, they are more likely to incorrectly segment it, but they might also expect to encounter the object DUCK. Recovering from these errors is less trivial.

	Word	Object	% Over Errors
BatchOpt (Bi)	**bu**(nnies)	RABBIT	6.4%
	(bu)**nnies**	RABBIT	
OnlineMem (Bi)	(bir)**die**	DUCK	10.2%
	bir(die)	DUCK	
SubtrSeg	**bu**(nnies)	RABBIT	8.1%
	bir(die)	DUCK	
RandOracle	**pi**(ggy)	PIG	8.5%
	bir(die)	DUCK	

Table 6.6: Example oversegmentation errors from the four learners that make them for items in the referential lexicon. Oversegmented lexical items are shown in bold with the remainder of the correct word in parentheses. The percentage of all lexical mappings that were incorrect because of oversegmentation is also given.

Undersegmentation mappings also occur, but behave somewhat differently. First, fewer mapping errors are the result of an undersegmentation. The errors which do exist tend to undersegment a noun with a determiner (e.g. *the book* segmented as *thebook*). In this case, mapping *thebook* onto the object BOOK is not a perfect mapping, but if children hear *thebook* they are in fact likely to see a BOOK in their environment.

6.3.3.3 Discussion

The results of both the joint and downstream word-object mapping tasks indicate that creating a small lexicon of words is entirely plausible regardless of whether it is done jointly with

segmentation or not. The ability of segmentation models to support word-object mapping clearly indicates that in the worst-case scenario, early word-object mapping is still possible. If, however, it is the case that segmentation is done at the same time as word-object mapping, then both tasks can benefit from one another. The joint model of Jones et al. (2010) lacks some of the performance of the stand-alone Frank et al. (2009) word-object mapping model, indicating that perhaps future work must be done to create a better model of joint learning.

While the general outlook for word-object mapping is quite good, the true purpose of these analyses is to determine the quality of various segmentation models without taking into account their gold standard performance. In this case, the downstream evaluation results clearly indicate that gold standard performance does not necessarily result in good word-object mapping performance. In particular, models which heavily oversegment appear to be at a disadvantage for problems such as this, especially when they oversegment the concrete nouns children are most likely to learn the meanings of. This result should warn researchers from treating over- and undersegmentation equally. Each type of error impacts segmentation performance in different ways. Oversegmentations appear to threaten the learner's ability to correctly create word-object mappings, posing a much larger threat to later learning.

6.4 Conclusion

The end result of both the stress and word-object mapping tasks is that gold standard performance does not necessarily indicate that a segmentation is useful for a child learner. While one might hope that producing a segmentation closer to the gold standard would result in better performance on later learning tasks, I find that this is not the case. In some cases, as with stress pattern learning, the proper bias can be obtained even when a model has poor segmentation performance. Models which incorporate lexical biases, such as the

Bayesian learners, are all capable of obtaining the proper stress pattern, regardless of how they are trained.

In other cases, later learning is not so simple. With word-object learning, many of the evaluated learners struggled. In particular, oversegmentation errors make word-object mapping more difficult by confusing the word-object co-occurrences. Models which made many of these errors ended up producing worse word-object mappings than the random oracle guesser. These very same models, however, obtain very good token F-scores when evaluating against the gold standard. This throws into doubt our ability to judge the quality of a model based solely on its gold standard performance. Only by judging a model from a number of different directions can the true level of performance be gauged.

In many cases, researchers have difficulty evaluating their models against anything other than a gold standard. Both joint modeling and downstream evaluation, however, are alternative options which should be explored in addition to the typical gold standard. A joint model represents a good avenue of investigation in cases where two tasks are closely interrelated and it can be assumed that child learners have already deduced how the tasks fit together. Downstream evaluation, on the other hand, while less sophisticated, is an option for evaluating the quality of a model's output without making specific assumptions about what knowledge infants have regarding the linking of the two tasks. Downstream evaluation can also be applied using any two models, while making the models joint often requires model modifications that can prove technically challenging. Both approaches, however, give a better picture than using a gold standard alone of how well a learning strategy actually succeeds.

Chapter 7

Conclusion

Choosing specific implementational details is one of the most important aspects of creating and evaluating a model. Every aspect of the modeling process makes assumptions about how learning occurs and this is especially important for cognitive modelers who hope to model actual human learning. Unfortunately, the modeler is forced to make decisions about a model for which there is no clear empirical evidence to turn towards. Because of this, choices of model input, inference, and evaluation are often made without sufficient evidence. Instead of making important design and evaluation choices based solely on the modeler's intuition, these decisions can be treated as parameters of the model. The free parameters of any model must be explored so as to better understand their importance in generating model behavior. In this same way, the work described in this dissertation has attempted to explore the importance of these design decisions.

In each of the previous simulation chapters, assumptions regarding model input, inference, and evaluation were explored to assess their impact. By modifying each of these aspects of model design in turn, valuable information was gained concerning possible infant segmentation strategies. Changing the unit of representation from phonemes to syllables suggests

that the segmentation task is less difficult than previously thought. Altering how inference is carried out generated empirical predictions that can be experimentally investigated. Evaluating models in terms of their ability to support additional learning tasks showed that gold standard performance alone is an insufficient metric for the quality of a segmentation model.

In each case, multiple design decisions could reasonably be made based on the experimental literature. Exploring these choices, however, does not just increase the researcher's understanding of the model. Instead, by modeling each scenario explicitly one can find testable predictions. If, for example, segmenting with phonemes predicts one type of error, but segmenting with syllables predicts another, these errors can be probed for in young children, eventually leading to further evidence in support of one or the other learning strategy. Modelers can therefore play an important role in resolving long-standing theoretical debates regarding the nature of language learning and understanding.

The work involved in this dissertation is by no means restricted to the modeling of speech segmentation. For all cognitive modelers, the techniques used here are applicable. By exploring the impact of modeling decisions on model behavior, future work will lead to a better understanding not just of language learning, but of cognition more generally.

Bibliography

J. Abbott, J. Hamrick, and T. Griffiths. Approximating bayesian inference with a sparse distributed memory system. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 1686–1691, 2013.

J. Anderson. *The adaptive character of thought*. Erlbaum, Hillsdale, NJ, 1990.

R. Aslin, J. Saffran, and E. Newport. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9:321–324, 1998.

R. N. Aslin, D. B. Pisoni, B. L. Hennessy, and A. J. Perey. Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. *Child development*, 52(4):1135, 1981.

R. Atkinson and R. Shiffrin. Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, 2:89–195, 1968.

A. D. Baddeley and G. Hitch. The recency effect: Implicit learning with explicit retrieval? *Memory & Cognition*, 21(2):146–155, 1993.

L. E. Baum. An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.

E. Bergelson and D. Swingley. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258, 2012.

N. Bernstein-Ratner. Patterns of vowel modification in motherese. *Journal of Child Language*, 11:557–578, 1984.

J. Bertonicini, R. Bijeljac-Babic, P. Jusczyk, L. Kennedy, and J. Mehler. An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology*, 117(1):21–33, 1988.

C. Best, G. McRoberts, and N. Sithole. Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by english-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3):345–360, 1988.

C. Best, G. McRoberts, R. LaFleur, and J. Silver-Isenstadt. Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. *Infant Behavior and Development*, 18:339–350, 1995.

R. Bijeljac-Babic, J. Bertoncini, and J. Mehler. How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, 29(4):711–721, 1993.

D. Blanchard, J. Heinz, and R. Golinkoff. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, 37:487–511, 2010.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

E. Bonawitz, S. Denison, A. Chen, A. Gopnik, and T. Griffiths. A simple sequential algorithm for approximating bayesian inference. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2011.

H. Bortfeld, J. Morgan, R. Golinkoff, and K. Rathbun. Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, 16(4):298–304, 2005.

L. Boruta, S. Peperkamp, B. Crabbé, and E. Dupoux. Testing the robustness of online word segmentation: Effects of linguistic diversity and phonetic variation. pages 1–9, 2011.

M. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105, 1999.

M. Brent and J. Siskind. The role of exposure to isolated words in early vocabulary. *Cognition*, 81:31–44, 2001.

R. Brown. *A first language: The early stages*. Harvard University Press, 1973.

S. Carey. The child as word learner. In J. Bresnan, G. Miller, and M. Halle, editors, *Linguistic Theory and Psychological Reality*, pages 264–293. MIT Press, Cambridge, MA, 1978.

M. H. Christiansen, J. Allen, and M. S. Seidenberg. Learning to segment speech using multiple cues: A connectionist model. *Language and cognitive processes*, 13(2-3):221–268, 1998.

C. Christodoulopoulos, S. Goldwater, and M. Steedman. A bayesian mixture model for part-of-speech induction using multiple features. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011.

R. Cole and J. Jakimik. *Perception and Production of Fluent Speech*, pages 133–163. Erlbaum, Hillsdale, NJ, 1980.

R. Cole, J. Jakimik, and W. Cooper. Perceptibility of phonetic features in fluent speech. *Journal of the Acoustical Society for America*, 64:44–56, 1978.

A. Cooper. Laryngeal and oral gestures in english /p, t, k/. In *Proceedings of the XIIth International Conference of Phonetic Sciences*, volume 2, pages 50–53, 1991.

R. M. Dauer. Stress-timing and syllable-timing reanalyzed. *Journal of phonetics*, 1983.

E. J. Davelaar, Y. Goshen-Gottstein, A. Ashkenazi, H. J. Haarmann, and M. Usher. The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychological review*, 112(1):3, 2005.

S. J. Davis, E. L. Newport, and R. N. Aslin. Probability-matching in 10-month-old infants. *Proceedings of the 33rd Cognitive Science Society*, pages 3011–3015, 2011.

S. Denison, E. Bonawitz, A. Gopnik, and T. Griffiths. Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, 126:285–300, 2013.

A.-M. Di Sciullo and E. Williams. *On the definition of word*, volume 14. Springer, 1987.

B. Dillon, E. Dunbar, and W. Idsardi. A single-stage approach to learning phonological categories: Insights from inuktitut. *Cognitive Science*, 37(2):344–377, 2013.

G. Dogil and B. Williams. The phonetic manifestation of word stress. *Word prosodic systems in the languages of Europe*, page 273, 1999.

G. Doyle and R. Levy. Combining multiple information types in bayesian word segmentation. In *HLT-NAACL*, pages 117–126, 2013.

P. Eimas. Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America*, 105(3):1901–1911, 1999.

N. Feldman, T. Griffiths, and J. Morgan. Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st annual conference of the cognitive science society*, pages 2208–2213, 2009.

N. Feldman, T. Griffiths, S. Goldwater, and J. Morgan. A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 120(4):751–778, 2013.

T. Ferguson. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1 (2):209–230, 1973.

A. Fernald. Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, 8:181–195, 1985.

A. Fernald and H. Morikawa. Common themes and cultural variations in japanese and american mothers' speech to infants. *Child development*, 64(3):637–656, 1993.

M. Fleck. Lexicalized phonotactic word segmentation. In *Proceedings of ACL-08: HLT*, pages 130–138, 2008.

C. Fougeron and P. Keating. Variations in velic and lingual articulation depending on prosodic position: Results for two french speakers. *UCLA Working Papers in Phonetics*, 92:88–96, 1996.

A. Fourtassi, B. Börschinger, M. Johnson, and E. Dupoux. Whyisenglishsoeasytosegment. In *Cognitive Modeling and Computational Linguistics 2013*, pages 1–10, 2013.

M. Frank, N. Goodman, and J. Tenenbaum. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20:579–585, 2009.

M. Frank, S. Goldwater, T. Griffiths, and J. Tenenbaum. Modeling human performance in statistical word segmentation. *Cognition*, 117(2):107–125, 2010.

O. Fujimura. Methods and goals of speech production research. *Language and Speech*, 33: 195–258, 1990.

J. Galliers and K. S. Jones. Evaluating natural language processing systems. Technical Report 291, Computer Laboratory, University of Cambridge, 1993.

T. Gambell and C. Yang. Word segmentation: Quick but not dirty. 2006.

A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, New York, 2004.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6: 721–741, 1984.

L. Gleitman, E. Newport, and H. Gleitman. The current status of the motherese hypothesis. *Journal of Child Language*, 11:43–79, 1984.

S. D. Goldinger and T. Azuma. Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, 31(3):305–320, 2003.

S. Goldwater, T. Griffiths, and M. Johnson. A bayesian framework for word segmentation. *Cognition*, 112(1):21–54, 2009.

T. L. Griffiths, F. Lieder, and N. D. Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7 (2):217–229, 2015.

P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *Systems Science and Cybernetics, IEEE Transactions on*, 4(2): 100–107, 1968.

M. Hauser, E. Newport, and R. Aslin. Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition*, 78:B53–B64, 2001.

J. Hay, B. Pelucchi, K. G. Estes, and J. Saffran. Linking sounds to meaning: Infant statistical learning in a natural language. *Cognitive Psychology*, 63:93–106, 2011.

E. Hoff. *Language Development*. Wadsworth, Belmont, CA, 2008.

E. Hohne and P. Jusczyk. Two-month-old infants' sensitivity to allophonic differences. *Perception & Psychophysics*, 56(6):613–623, 1994.

L. Hyman. Do all languages have word accent? or: What's so great about being universal? In *Proceedings of the 1st Annual Workshop on Stress and Accent*, University of Conneticut, 2010.

E. Johnson and P. Jusczyk. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44:548–567, 2001.

E. Johnson and M. Tyler. Testing the limits of statistical learning for word segmentation. *Developmental Science*, 13(2):339–345, 2010.

M. Johnson. Unsupervised word segmentation for sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, 2008.

M. Johnson and K. Demuth. Unsupervised phonemic chinese word segmentation using adaptor grammars. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 528–536. Association for Computational Linguistics, 2010.

M. Johnson and S. Goldwater. Improving nonparametric bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325, 2009.

M. Johnson, T. Griffiths, and S. Goldwater. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in Neural Information Processing Systems*, 19:641–648, 2007.

M. Johnson, K. Demuth, B. Jones, and M. J. Black. Synergies in learning words and their referents. In *Advances in neural information processing systems*, pages 1018–1026, 2010.

B. K. Jones, M. Johnson, and M. C. Frank. Learning words and their meanings from unsegmented child-directed speech. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 501–509. Association for Computational Linguistics, 2010.

P. Jusczyk. *The Discovery of Spoken Language*. MIT Press, Cambridge, MA, 1997.

P. Jusczyk and C. Derrah. Representation of speech sounds by young infants. *Developmental Psychology*, 23(5):648–654, 1987.

P. Jusczyk, A. Cutler, and N. Redanz. Infants' preference for the predominant stress pattern of english words. *Child Development*, 64(3):675–687, 1993.

P. Jusczyk, A. Jusczyk, L. Kennedy, T. Schomberg, and N. Koenig. Young infants' retention of information about bisyllabic utterances. *Journal of Experimental Psychology: Human Perception and Performance*, 21(4):822–836, 1995.

P. Jusczyk, E. Hohne, and A. Baumann. Infants' sensitivity to allphonic cues for word segmentation. *Perception and Psychophysics*, 61:1465–1476, 1999a.

P. Jusczyk, D. Houston, and M. Newsome. The beginnings of word segmentation in english-learning infants. *Cognitive Psychology*, 39:159–207, 1999b.

D. Kahneman, P. Slovic, and A. Tversky, editors. *Judgment under uncertainty: Hueristics and biases*. Cambridge University Press, 1982.

C. H. Kam and A. Chang. Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory and Cognition*, 35(3):815–821, 2009.

C. H. Kam and E. Newport. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development*, 1 (2):151–195, 2005.

C. L. H. Kam and E. L. Newport. Getting it right by getting it wrong: When learners change languages. *Cognitive psychology*, 59(1):30–66, 2009.

P. Kingsbury, S. Strassel, C. McLemore, and R. MacIntyre. *CALLHOME American English lexicon (PRONLEX)*. Linguistic Data Consortium, 1997.

C. Kitamura and D. Burnham. Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy*, 4(1):85–110, 2003.

O. Kolodny, A. Lotem, and S. Edelman. Learning a generative probabilistic grammar of experience: A process-level model of language acquisition. *Cognitive Science*, 39:227–267, 2015.

K.-M. Köpcke. The acquisition of plural marking in english and german revisited: schemata versus rules. *Journal of child language*, 25(02):293–319, 1998.

M. Korman. Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First language*, 5:44–45, 1984.

P. Kuhl, K. Williams, F. Lacerda, K. Stevens, and B. Lindblom. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255:606–608, 1992.

P. Ladefoged. *A course in phonetics*. Harcourt Brace, Fort Worth, TX, 1993.

B. Landau, L. Smith, and S. Jones. The importance of shape in early lexical learning. *Cognitive Development*, 3:299–321, 1988.

I. Lehiste. Suprasegmental features of speech. *Contemporary issues in experimental phonetics*, 225:239, 1976.

P. Liang and D. Klein. Online em for unsupervised models. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages 611–619, 2009.

A. Liberman, F. Cooper, D. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. *Psychological Review*, 74(6):431–461, 1967.

C. Lignos. Modeling infant word segmentation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 29–38. Association for Computational Linguistics, 2011.

C. Lignos. Infant word segmentation: An incremental, integrated model. In *Proceedings of the 30th West Coast Conference on Formal Linguistics*, pages 237–247, 2012.

C. Lignos and C. Yang. Recession segmentation: Simpler online word segmentation using limited resources. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 88–97, 2010.

B. MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edition, 2000.

E. Markman. *Categorization and Naming in Children: Problems of Induction*. MIT Press, Cambridge, MA, 1989.

E. Markman and G. Wachtel. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20:121–157, 1988.

E. Markman, J. Wasow, and M. Hansen. Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, 47:241–275, 2003.

L. Markson and P. Bloom. Evidence against a dedicated system for word learning in children. *Nature*, 385:813–815, 1997.

D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Co. Inc., New York, N.Y., 1982.

B. Marthi, H. Pasula, S. Russell, and Y. Peres. Decayed mcmc filtering. In *Proceedings of 18th UAI*, pages 319–326, 2002.

D. W. Massaro. Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, 79(2):124, 1972.

D. W. Massaro. Perceptual units in speech recognition. *Journal of experimental Psychology*, 102(2):199, 1974.

D. W. Massaro. *Understanding Language: An Information Processing Analysis of Speech Perception, Reading and Psycholinguistics*. Academic Press, New York, 1975.

S. Mattys, P. Jusczyk, and P. Luce. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38:465–494, 1999.

J. Maye, J. Werker, and L. Gerken. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82:B101–B111, 2002.

J. Maye, D. Weiss, and R. Aslin. Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, 11(1):122–134, 2008.

T. McRae. *The impact of computers on accounting*. Wiley, 1964.

J. Mehler, J. Y. Dommergues, U. Frauenfelder, and J. Segui. The syllable's role in speech segmentation. *Journal of verbal learning and verbal behavior*, 20(3):298–305, 1981.

G. A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

B. B. Murdock Jr. The serial position effect of free recall. *Journal of experimental psychology*, 64(5):482, 1962.

A. Newell and H. A. Simon. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3):113–126, 1976.

E. Newport. Maturational constraints on language learning. *Cognitive Science*, 14:11–28, 1990.

S. J. N.Z. Kirkham, J.A. Slemmer. Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, 83:B35–B42, 2002.

L. Pearl. Evaluating learning strategy components: Being fair. *Language*, 90(3):e107–e114, 2014.

L. Pearl and J. Sprouse. Computational modeling for language acquisition: A tutorial with syntactic islands. *Journal of Speech, Language, and Hearing Research*, 58:740–753, 2015.

L. Pearl, S. Goldwater, and M. Steyvers. Online learning mechanisms for bayesian models of word segmentation. *Research on Language and Computation*, 8(2):107–132, 2011. special issue on computational models of language acquisition.

F. Pellegrino, C. Coupé, and E. Marsico. Across-language perspective on speech information rate. *Language*, 87(3):539–558, 2011.

B. Pelucchi, J. Hay, and J. Saffran. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, 113:244–247, 2009.

A. Perfors. Memory limitations alone do not lead to over-regularization: An experimental and computational investigation. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 3274–3279, 2011.

A. Perfors, J. Tenenbaum, T. Griffiths, and F. Xu. A tutorial introduction to bayesian models of cognitive development. *Cognition*, 120(3):302–321, 2011.

A. Perfors, K. Ransom, and D. J. Navarro. People ignore token frequency when deciding how widely to generalize. In *36th Annual Meeting of the Cognitive Science Society (CogSci 2014)(23 Jul 2014-26 Jul 2014: Quebec City, Canada)*, 2014.

A. Peters. *The Units of Language Acquisition*. Monographs in Applied Psycholinguistics. Cambridge University Press, New York, 1983.

L. Phillips and L. Pearl. The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, pages 1–31, 2015.

J. Pierrehumbert and D. Talkin. *Lenition of /h/ and glottal stop*, pages 90–117. Cambridge University Press, Cambridge, UK, 1992.

D. Pisoni. Perceptual processing time for consonants and vowels. Status Report on Speech Research 31/32, Haskins Laboratories, 1972.

L. Polka and J. Werker. Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20(2):421–435, 1994.

L. Polka, C. Colantonio, and M. Sundara. A cross-language comparison of/d/–/ð/perception: evidence for a new developmental pattern. *The Journal of the Acoustical Society of America*, 109(5):2190–2201, 2001.

P. Rollins. Caregiver contingent comments and subsequent vocabulary comprehension. *Applied Psycholinguistics*, 24:221–234, 2003.

A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, volume 7, pages 410–420, 2007.

D. Roy, R. Patel, P. DeCamp, R. Kubat, M. Fleischman, B. Roy, N. Mavridis, S. Tellex, A. Salata, J. Guinness, et al. The human speechome project. In *Symbol Grounding and Beyond*, pages 192–196. Springer, 2006.

S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Always learning. Pearson, 2013. ISBN 9781292024202. URL `https://books.google.com/books?id=DFJtngEACAAJ`.

J. Saffran, R. Aslin, and E. Newport. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928, 1996.

A. Sanborn, T. Griffiths, and D. Navarro. Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4):1144–1167, 2010.

A. Seidl and E. K. Johnson. Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental science*, 9(6):565–573, 2006.

E. Selkirk. *English Compounding and the Theory of Word structure*, pages 229–277. Foris, Dordrecht, 1981.

L. Shi, T. Griffiths, N. Feldman, and A. Sanborn. Exemplar models as a mechanism for performing bayesian inference. *Psychonomic Bulletin & Review*, 17(4):443–464, 2010.

L. Smith and C. Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008.

D. Swingley. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50:86–132, 2005.

Y. Teh, M. Jordan, M. Beal, and D. Blei. Heirarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

T. Teinonen, V. Fellman, R. Näätänen, P. Alku, and M. Huotilainen. Statistical language learning in neonates revealed by event-related brain potentials. *BMC neuroscience*, 10(1): 21, 2009.

E. Thiessan and J. Saffran. Learning to learn: Infant's acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, 3(1):73–100, 2007.

E. Thiessen and J. Saffran. 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, 2003.

R. Tincoff and P. W. Jusczyk. Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, 10(2):172–175, 1999.

R. Tincoff and P. W. Jusczyk. Six-month-olds comprehend words that refer to parts of the body. *Infancy*, 17(4):432–444, 2012.

A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(415):1124–1131, 1974.

G. Vallabha, J. McClelland, F. Pons, J. Werker, and S. Amano. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33):13273–13278, 2007.

A. Venkataraman. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27(3):351–372, 2001.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

H. Wang and T. Mintz. A dynamic learning model for categorizing words using frames. In *Proceedings of BUCLD*, volume 32, pages 525–536, 2008.

R. Weide. The cmu pronunciation dictionary, release 0.6, 1998.

A. Weisleder and A. Fernald. Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, 24(11):2143–2152, 2013.

J. Werker and C. Lalonde. Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, 24(5):672–683, 1988.

J. Werker and R. Tees. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*, 7:49–63, 1984.

M. Wilson. The mrc psycholinguistic database machine readable dictionary. *Behavioral Research Methods, Instruments and Computers*, 20:6–11, 1988.

F. Xu and S. Denison. Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, 112(1):97–104, 2009.

F. Xu and V. Garcia. Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, 105(13):5012–5015, 2009.

F. Xu and J. Tenenbaum. Word learning as bayesian inference. *Psychological Review*, 114 (2):245–272, 2007.

C. Yang. On productivity. *Linguistic variation yearbook*, 5(1):265–302, 2005.

C. Yu and L. B. Smith. Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5):414–420, 2007.

C. Yu and L. B. Smith. What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science*, 14(2):165–180, 2011.

Appendix A

Phoneme-Based Segmentation

Although experimental evidence suggests that children do not begin segmentation using phonemes (see Section 2.1.4), it can still be useful to understand how models behave when placing boundaries between phonemes rather than syllables. Following the same methodology as in Chapter 4, I evaluate the DPSEG, SubtrSeg, TPminima, and RandOracle learners on phoneme-based corpora across seven languages. Descriptive statistics for the phoneme-based corpora are given in Table A.1

A.1 Model Parameters

Table A.2 presents the parameter settings for the DPSEG learners on each language. The parameter values for each language vary somewhat more than for the syllable-based corpora. That being said, results are not entirely dependent on identifying the optimal free parameters. In the case of the unigram phoneme-based English learner, as α is increased from 1 to 500, the resulting word token F-score ranges only from 54.2 to 55.0. Parameter values for the bigram model (both phoneme- and syllable-based) are somewhat more important. In general, low

Language	Corpus	Ages	Utt	Phon types	Phons/Utt	B Prob
English	Brent	0;6-0;9	28391	41	10.37	25.69
German	Caroline	0;10-4;3	9378	45	13.70	23.22
Spanish	JacksonThal	0;10-1;8	16924	38	9.71	23.56
Italian	Gervain	1;0-3;4	10473	29	19.35	21.17
Farsi	Family & Samadi	1;8-5;2	31657	37	15.82	17.67
Hungarian	Gervain	1;11-2;11	15208	46	15.13	19.19
Japanese	Noji, Miyata, & Ishii	0;2-1;8	12246	33	8.63	18.50

Table A.1: Summary of the phoneme-based child-directed language corpora, including the CHILDES database corpora they are drawn from (Corpus), the age ranges of the children they are directed at (Ages), the number of utterances (Utt), the number of unique phonemes (Phon types), the average number of phonemes per utterance (Phons/Utt), and the probability of a word boundary appearing between phonemes (B Prob).

	α	β	γ
English	200	10	200
German	100	200	200
Spanish	50	50	100
Italian	500	500	500
Farsi	500	500	3000
Hungarian	500	500	500
Japanese	200	500	500

Table A.2: Parameter values for all phoneme-based unigram and bigram Bayesian models across each language. Values were selected based on previous research in the following ranges α: [1, 500], β: [1, 500], γ: [1, 3000].

values of γ produce poor word token performance. As reported by Goldwater et al. (2009), changes to β and γ generally have modest impacts on word token F-scores. However, very low values of γ ($<$100) tend to reduce token F-score performance significantly. For English, at the optimal value of γ at 200 changes in β produce only minor changes in performance (token F-score ranging from 70.7 to 73.1).

A.2 Gold Standard Results

A.2.1 Word Token Results

	English	German	Spanish	Italian	Farsi	Hungarian	Japanese
DPSEG-1	54.9	62.7	53.1	58.5	57.2	58.9	**64.8**
DPSEG-2	**67.2**	**68.0**	**66.4**	**61.9**	**57.7**	**60.1**	64.4
SubtrSeg	6.4	13.5	9.5	4.2	5.1	2.8	5.0
TPminima	53.4	40.6	30.5	21.2	23.0	22.9	28.1
RandOracle	5.5	4.1	9.3	6.4	1.6	5.7	10.0

Table A.3: Word token F-score results for the Bayesian BatchOpt and non-Bayesian learners on the phoneme-based corpora.

Word token F-score results for each language are presented in Table A.3. The first finding from these results is that the Bayesian bigram model tends to outperform the unigram model (although see Japanese as a counter-example). In some cases, the benefit of the bigram assumption is quite modest (e.g. Farsi, Italian, and Hungarian) and in others the benefit is quite large (e.g. English, German). The crucial benefit of the bigram model is that it is able to account for frequently co-occuring words. For example, if *what's* and *that* frequently appear next to one another, the bigram model can account for this fact by treating them as a bigram. The unigram model, on the other hand, can only account for their co-occurrence by treating them as a single word. This means that unigram learners will tend to *undersegment* these items, treating shorter words as if they jointly made up a larger word. This generally occurs for items that show up frequently, disproportionately affecting the word token scores.

As opposed to previous cross-linguistic work, I find that both the unigram and bigram Bayesian DPSEG model performs quite well on all languages. The DPSEG model has not been evaluated previously on many of these languages, making comparison with previous results somewhat more difficult. On English, the results are generally consistent with Goldwater et al. (2009) and Pearl et al. (2011). Likewise, the results on Spanish are comparable

with previous results on Spanish adult-directed speech (Fleck, 2008). In that same study, Fleck (2008) report very poor DPSEG results on an adult-directed corpus of Arabic. It may well be the case that the DPSEG model performs poorly on phoneme-based corpora of other languages, the poor performance on Arabic must also take into account that the corpus used was of transcribed telephone recordings. In cases where a phonetic pronunciation could not be found (e.g. foreign words, partial words), Fleck (2008) used orthographic forms instead, possibly compounding the poor performance of the model, although it is unclear how often this occurred.

Looking at the non-Bayesian learners, performance is much reduced. Neither the SubtrSeg or RandOracle learners are able to successfully segment any of the languages, their best performance being German for the SubtrSeg (token F-score: 13.5) and Japanese for the RandOracle (token F-score: 10.0). The TPminima learner does much better, but its performance ranges from very poor (e.g. Italian: 21.2) to nearing the Bayesian performance (e.g. English: 53.4). All of these non-Bayesian learners were designed for a syllable-based corpus, and therefore it might be no surprise that they do not perform well. However, this indicates that if children do actually segment using phonemic input, then they must not be using any of the non-Bayesian learning strategies evaluated here.

A.2.2 Lexicon Results

	English	**German**	**Spanish**	**Italian**	**Farsi**	**Hungarian**	**Japanese**
DPSEG-1	**55.6**	**60.0**	54.2	**47.8**	**50.6**	**44.1**	**54.8**
DPSEG-2	55.2	57.4	**57.2**	45.3	47.7	38.6	51.2
SubtrSeg	5.4	8.0	11.4	1.3	0.7	1.5	6.7
TPminima	33.2	26.8	25.6	13.0	15.5	15.2	26.6
RandOracle	16.2	11.3	18.6	8.8	10.1	9.6	16.4

Table A.4: Word type F-score results for the Bayesian BatchOpt and non-Bayesian learners on the phoneme-based corpora.

Another way to measure the performance of these models is to examine word types[1], rather than word tokens. Based on Table A.4, one can see that word type scores are in general much worse than word token scores. Because there are so many more tokens than types, an error which occurs just once will have a much larger impact on the word type scores. For instance, imagine a model which encounters *the doggie* ten times. A model which correctly segments this in nine cases will achieve a better token F-score than a model which correctly segments it in only five cases. If each model incorrectly segments the phrase as *thedoggie* at least once, then each model will have the same lexicon F-score, because their lexicons are the same (i.e. *the*, *doggie*, *thedoggie*). Therefore, making an incorrect segmentation, even just once, can have a large impact on lexicon performance.

In comparing unigram and bigram learners, I find a general increase in word token F-score performance for most learners when moving to the bigram language model. In terms of lexicon F-scores, the bigram assumption is only useful for Spanish. The decrease in lexicon performance might be attributable to increased segmentation of affixes (e.g. plural *-s* or the progressive ending *-ing*).

Generall, the lexicon results confirm the findings of the word token scores. The unigram and bigram Bayesian learners are capable of producing good segmentations on all of the languages evaluated here. The non-Bayesian learners, however, are unable to do so, with the SubtrSeg and RandOracle learners unable to perform well on any of the languages and the TPminima learner performing well only on English and German. It would seem then, that if infants do in fact use phonemes as their basic unit of representation, then the Bayesian DPSEG model might be a good candidate for the very beginnings of infant speech segmentation.

[1]Because the SubtrSeg learner sometimes segments words without adding them to its lexicon, it should be noted that the word type results take into account *all* of the words a model outputs. This assumes the SubtrSeg learner is keeping track of all words segmented, even if some of these words do not impact later segmentation.

Based on the evidence presented in Chapter 2, this seems unlikely to be the case. Certainly, infants do not appear to have learned the full phonology of their language, making the use of phonemes questionable. Since the corpora explored here do not have detailed phonetic transcriptions available, it cannot be said whether or not the model is successful when using phones rather than phonemes. Future work might explore this avenue of research, especially if any evidence were to accumulate that infants do in fact use phones as the basis for speech segmentation.