– STROLL –

A combined model of word segmentation
and word learning

Muriël Irene Eerkens
3216942

Master thesis Linguistics
Utrecht University
December 20, 2013

Supervisor: prof. dr. R.W.J. Kager
Second reader: dr. T.O. Lentz

“By words we learn thoughts –
and by thoughts we learn life. ”

-Jean Baptiste Girard

Contents

1 Introduction

An interesting field of research is focused on two related processes of early infant language acquisition, occurring mainly in the second half of the first year of the infant's life. This research is aimed at the process of word segmentation and the related process of word acquisition. The aim of this research is to create an accurate representation of the two processes and the way they interact with each other. The research field consists mainly of two parts, an experimental and computational part. The former is focused on defining the skills and knowledge infants have at their disposal for this task and the trajectory of how these abilities develop. The latter combines and applies the skills in a computational simulation of the segmentation and acquisition process, in an effort to reproduce the learning trajectory. Chapter 2 presents an overview of the experimental research focused on this field of study.

An auditorily presented sentence is perceived as a continuous string of sounds. To transform this into a meaningful string of words, the sound string needs to be split up. Chapter 3 presents an overview of the computational models which have simulated the segmentation process. As we shall see in section 3.2, in adult speech perception this task is mainly performed through word recognition (McClelland and Elman, 1986; Norris, 1994; Norris and McQueen, 2008), by locating words in the input. Perceived sounds are matched with forms from the lexicon. However, in the process of language acquisition a language learner does not yet possess a lexicon of the language and needs to segment the speech first in order to build one. In this situation, the learner starts with bottom-up information which is directly present in the input to locate word boundaries. This segmentation task has been modeled multiple times, either only using bottom-up information and no information from the acquired lexicon (Christiansen et al., 1998; Adriaans and Kager, 2010) or using both segmentation and word recognition strategies (Brent and Cartwright, 1996; Goldwater et al., 2009; Daland and Pierrehumbert, 2010; Apoussidou and Kager, 2013).

When a sentence has been segmented either by locating words or by locating word boundaries, the separate parts of the sentence are added to the lexicon. This task of word learning influences the processing of subsequent sentences: it adds forms to the lexicon which is the source of the word recognition process. There are a few models that pose restrictions on which words may enter the lexicon (Blanchard et al., 2010; Apoussidou and Kager, 2013), but the relation between word learning and subsequent word segmentation has not been studied extensively (Daland, 2009, chap.5). Experimental research has shown the importance of certain factors in the word learning process (Swingley, 2009). However, there are no segmentation models that include these factors in the simulation of the word acquisition process. In order to create a complete simulation of the word segmentation process, word acquisition needs to be included. In order to create an accurate simulation, the simulated word acquisition needs to reflect aspects that are determined relevant by the experimental research.

This thesis presents a segmentation and word learning model that integrates two relevant aspects of the infant's word learning process. The first factor is a property of the words that the infant learns; the second is a property of the environment in which an infant acquires word forms. Among the aspects in which word forms can differ from each other, is the frequency with which it is perceived by an infant. The more often it is perceived, the earlier it is learned (Roy et al., 2009). The second factor is related to the specific language that adults use when talking to children. There are several special characteristics of infant-directed speech, which help the infant in the language acquisition and word learning process (Brent and Siskind, 2001). One of the characteristics of the way words are presented to infants is that they often occur with a frequency peak. They appear a high number of times within a short amount of time, for example in the environment of a story that is read to the child. Section 2.3 presents the experimental research supporting the relevance of these two factors. In order to create a complete and accurate representation of the word segmentation and acquisition process, the factors that are relevant in the infant word acquisition process should be included. This thesis starts with including the aspects of frequency and frequency peaks.

Chapter 2 presents an overview of the experimental literature relevant for this thesis. Chapter 3 presents an overview of the computational literature; segmentation and word learning models that have been proposed, with different applications of the lexicon. The subsequent chapter presents the model proposed in this thesis, called Segmentation Through Restrictions On Lexical Learning (STROLL), with the two factors of word learning integrated. Chapter 5 presents the results of the STROLL model, followed by a general discussion.

2 Experimental research on word segmentation and word learning

Every listener has to face the task of parsing continuous input speech into meaningful units – words. This task consists of three parts: word segmentation, word acquisition and word recognition (Daland, 2009). When a lexicon of the words that make up the continuous speech is available to the listener, top-down word recognition takes on the biggest role in segmenting the string into words. However, an infant does not possess a lexicon but needs to create one from the speech input it receives. This chapter presents an overview of experimental studies on the infant word segmentation and word learning processes. Section 2.2 is focused on the concept of a word form lexicon that functions as a stepping stone to the adult lexicon containing phonological, semantic and syntactic information. Section 2.3 presents an overview of the experimental research on infant word learning, focusing on aspects of words and the environments in which words are presented to infants that are relevant for the learning process.

In order to learn words from the input speech, they need to be extracted from it first. This process of segmentation is an interesting field of both computational and experimental research. There are several types of cues available in the speech input that can be used to locate boundaries between words. The computational segmentation research focuses on modeling the process of applying these cues to the input an infant receives to see whether it leads to enough information to correctly locate word boundaries and build a lexicon. The experimental segmentation research is focused on determining whether infants are sensitive to these cues and at what age, to see whether they can apply them in the segmentation process. The next chapter will present a subset of the computational segmentation models presented in the literature. The next section gives an overview of the cues available in the speech infants receive and infants' sensitivity to these cues.

2.1 Segmentation cues

The field of segmentation research aims to reproduce the infant segmentation process while limiting the set of assumptions necessary in the reproduction. One of the main areas in which the assumptions should be as minimal as possible, is the area of the segmentation cues available to the infant. A segmentation cue is a specific type of information which is available in the speech input infants receive and which can be used to locate word boundaries. This section presents an overview of segmentation cues and experimental research supporting them as usable cues in the segmentation process. For this, the cues need to be present and informative about the location of word boundaries in the input speech. This is tested by the computational models presented in chapter 3. Primarily, infants need to be sensitive to the cues and they need to be shown to be able to apply each cue in the segmentation process. Segmentation cues for which infants are sensitive and which they can apply in the process of locating word boundaries can be used in

a bottom-up segmentation strategy. A bottom-up strategy relies on the input signal, which is available to infants, and not on (lexical) knowledge, which is not available. It thus requires less assumptions than a top-down strategy since all the information is available in the input and no pre-existing knowledge is required. This section presents several types of segmentation cues and their availability to infants in the language acquisition process.

Prosody

A phonological cue is that of prosody. One way in which prosodic cues occur is at utterance boundaries. A pause is often inserted at an utterance boundary, signaling the presence of the utterance boundary which is also a word boundary. Besides the pause, the utterance final prosody is distinctive and usable as an extra utterance/word boundary cue. Seidl and Johnson (2006) have tested 8 months old infants on their ability to recognize words in isolation after having heard them in utterances in the training phase. Words occurred either utterance initial, medial or final in the training phase. The utterance medial words were significantly less recognized in the test phase than the other two types of words.

Infants' sensitivity to metrical cues of lexical stress has been tested by Jusczyk et al. (1999b), who compared the segmentation abilities of 7.5 months olds with words following the dominant strong-weak versus words with a weak-strong stress pattern. Strong syllables were seen as word initial. In situations where a weak syllable followed a weak-strong word, the word boundaries were incorrectly placed: from 'guiTAR is', the word 'TARis' was segmented. Both at 6 and 9 months infants are sensitive to the patterns of stressed and unstressed syllables in their native language (Morgan and Saffran, 1995), and from 7.5 months these metrical patterns are available in the segmentation process (Jusczyk et al., 1999b).

Transitional probabilities

In a very influential study, 8 months old infants were presented with two minutes of continuous artificial language in which the only available cue to word boundaries were the transitional probabilities between syllables (Saffran et al., 1996). Some syllables – word internal ones – were followed by another syllable with 100% probability. For word final syllables, the probability of the following syllable was only 33%. After the two minute training phase, infants were tested on novel words (combinations of syllables that did not occur in the training data) and on part words (syllable strings that occurred across word boundaries in the training phase). In both cases, infants showed a significant preference for the words they had not heard in the training phase, called a novelty preference. This study showed that infants are sensitive to probability distributions of syllables. Among many other replications, Johnson and Jusczyk (2001) found the same results. The only changes from the Saffran et al. (1996) study were using natural speech instead of synthesized speech and only part-words were tested in the testing phase.

Phonotactics

Another use of the transitional probabilities cue is the probability of phoneme sequences. Each language

distinguishes between grammatical and ungrammatical phoneme sequences, or more precisely makes a gradient distinction between more and less well-formed phoneme combinations (Chomsky and Halle, 1965). This is the phonotactic grammar of the language. Phonotactics is a useful cue for segmentation, since certain combinations of phonemes are not grammatical within a word and can only occur across a word boundary. Infants can learn about the phonotactics of a language by learning the probability distribution of phoneme sequences, or the transitional probabilities between phonemes. The probability distribution of phoneme sequences does not provide all the information that is contained in the phonotactic grammar, but a distinctly low transitional probability between phonemes can signal a word boundary.

Jusczyk and Luce (1994) have studied whether 6 and 9 months old infants are sensitive to the likelihood of phonotactic patterns in word forms. The infants were presented with two lists of monosyllables. One consisted of words with a phonetic pattern with a high likelihood, the other had phonetic patterns with a low likelihood. Likelihood was determined by the frequency with which the pattern occurred in child-directed speech in the language the infants were acquiring (in this case, American English). Only the older infants listened significantly longer to the high likelihood list. The younger infants showed no discrimination of the two lists, showing that infants develop a sensitivity to the likelihood of phonotactic patterns between their 6th and 9th months of age.

This sensitivity to phonotactic probabilities has been shown to be useful for infants in the segmentation task as well (Mattys et al., 1999; Mattys and Jusczyk, 2000). Mattys et al. (1999) presented 9 months old infants with two lists of strong-weak stressed bisyllabic novel words of CVC-CVC form. In one list, the coda-onset consonant cluster inside the word had a high probability of occurring within a word in the natural language, in the other list the cluster had a high between-words probability. Because both lists contained strong-weak stressed forms, they were perceived as lists of word forms. The list with between-word consonant clusters inside the forms would therefore be seen as conflicting, and the infants showed a significant preference for the other list. Mattys and Jusczyk (2000) tested whether 9 months old infants showed discrimination in their attention for a word which was previously heard in a context with or without useful phonotactic cues. The context had useful cues when the offset of the word preceding the target had a high probability of being word final, and the onset following the target had a high probability of being word initial. The infants showed a significant difference in listening time, leading to the conclusion that they use probability differences in phoneme sequences to segment speech.

Acoustic cues

The combination of different sounds in the language input does not only provide information on the categorical phonotactic grammar, it also provides acoustic cues. Two such cues are discussed here. The first phenomenon is still phonological, while the second is the result of phonetics. The allophonic phonological cue, which signals the location of a phoneme in the syllable by the phonetic form the

phoneme takes, is useful for segmentation for 10.5 months old infants, although 9 months old infants require other information as well (Jusczyk et al., 1999a). The first but not the latter group was able to correctly identify a word boundary in 'night rate' versus 'nitrate'. Johnson and Jusczyk (2001) showed that 8 months old infants are sensitive to co-articulation cues, which are phonetic effects of two phonemes occurring together within a word. This cue is also useful for segmentation.

Familiar words

When a familiar word in the continuous speech stream is recognized, word boundaries can be placed around the form. This process inserts a boundary with a high probability inside the speech stream. As discussed above, infants are more likely to recognize a word that occurs utterance initially (Seidl and Johnson, 2006). A familiar utterance medial word can give the same certainty to the word following it as an utterance boundary would. Bortfeld et al. (2005) have tested 6 months old infants on their ability to recognize words which were previously presented following either a familiar word (their name, or 'mommy') or a non-familiar word (but phonologically similar to familiar ones, such as 'tommy'). The infants showed a significant difference in recognition of words which they previously heard following a familiar word, showing their ability to identify a word boundary with a new word starting after the known form.

For adults who have a full lexicon available, it can occur that the input matches with multiple entries in the lexicon. In this case, the different lexical forms need to compete with each other. In a word recognition with priming task, Goldinger et al. (1989) tested the Neighborhood Activation Model (NAM) on adults' speech processing. This model states that when a form is activated, other forms which are phonetically close are inhibited. This especially has an effect with low-frequency primes. The authors found that when a form was primed, a phonetically similar target form was less well recognized by the participants. This effect was not present when the prime and the target were presented to the test subjects with more time in between. The results of this study show the presence of lexical competition effects in adult speech recognition: when multiple familiar forms fit with the input data, they inhibit each other.

Possible words

When a continuous input string is segmented into parts, the form of these resulting parts can be a cue to whether a boundary should be placed. When the string /dɔgz/ is heard and the familiar word /dɔg/ is in the lexicon which leads to placing a boundary after the /g/, this could leave the /z/ as a single segment form. In an adult word recognition task, Norris et al. (1997) presented subjects with nonsense words containing familiar words, and filler nonsense words which did not contain a known form. The familiar word was either preceded or followed by extra material, and this extra material was either syllabic or not. The latter distinction was the main independent factor of the study. The authors predicted a longer response time to recognize a word when the added material was not syllabic, and could thus not form a word on its own.

To recognize the word /æpəl/ in /fæpəl/, the single consonant /f/ residue is not a possible word, in contrast to /vʌfæpəl/, leaving /vʌf/ as a segmented part and possible word (Norris et al., 1997, p202). There was a significantly longer reaction time (67 ms, $p<0.001$) for forms with non-syllabic added material compared to syllabic material.

It has not been tested whether infants use the same segmentation filter. There are studies that show that infants perceive speech input in (holistic) syllables (Bertoncini and Mehler, 1981), rather than as a string of separate phonemes. This would entail that a word boundary cannot be placed in /dɔgz/ because the string is already perceived as only one unit. In the computational segmentation literature, it is typically assumed that infants perceive the input as a string of segments rather than syllables. Mostly because word boundaries occur on syllable boundaries, and providing the latter unfairly simplifies the segmentation task. However, many models incorporate a vowel-constraint, stating that a form which does not contain a syllabic element cannot be a licit word form and may not be the result of a segmentation.

Cue integration

This section discusses a multitude of cues which are available to infants in the segmentation process. None of these cues is informative enough to solve the segmentation problem by itself, several cues need to be combined to detect all word boundaries correctly. For example, the metrical stress cue is useful in English since most word forms have a strong-weak stress pattern. But other cues need to be applied to correctly segment the weak-strong patterned forms such as *guitar* and *belong*. It has been suggested that the integration of multiple segmentation cues starts around 9 months (Lalonde and Werker, 1995).

A study by Mattys et al. (1999) looked at phonotactic and prosodic segmentation cues and their interaction. They presented infants of 9 months old with two lists of conflicting word forms to see which group of forms would be preferred. The first list contained forms which were prosodically seen as whole words with a strong-weak pattern, the second list contained forms that appeared as two forms because of the weak-strong stress pattern. The phonotactic cue was applied the opposite way. The first list contained a word internal consonant cluster that had a high probability of being found between words in the natural language the infants were acquiring. Based on the phonotactics, the infants would not perceive this list as containing well-formed bisyllabic word forms but rather containing pairs of monosyllabic forms. The list with the weak-strong pattern contained consonant clusters with a high likelihood of occurring word internally. The infants significantly preferred the list with the words with prosodic marking, even though the phonotactic cue in that list was conflicting. This experiment showed that 9 months old infants have a stronger preference for the prosodic cue than a phonotactic cue.

Johnson and Jusczyk (2001) have presented 8 months old infants with three types of familiarization stimuli: containing only statistical information (a repetition of the Saffran et al. (1996) experiment), containing conflicting statistical and prosodic information and containing conflicting statistical and coar-

ticulatory information. All types were naturally spoken strings of CV-syllables, with either 100% or 33% transitional probability between the syllables. For the conflicting prosodic cue, the syllable preceding the 33% probable transition was stressed. For the conflicting coarticulation cue, the syllable following the less probable transition was coarticulated with the previous syllable. In the experiment with only the statistical cue, infants showed a novelty effect in the testing phase. This denotes that infants listen longer to something that appears as new and interesting to them. Extending this result to the other two experiments, the infants are expected to listen longest to the forms that appear novel to them in the sense that they had not segmented them from the speech as word forms themselves. In both conflicting experiments, the infants listened significantly longer to the word forms that were normal words when segmented with the statistical cue, but part-words when segmented with the conflicting cue.

In an extensive study, Mattys et al. (2005) have looked at the interaction of segmentation cues in adult speech perception. They distinguished three tiers of segmentation cues: lexical (familiar words) and sublexical, the latter divided in segmental (phonotactics, coarticulation) and metrical prosody (word stress). These three tiers are hierarchically ordered in that people most heavily rely on the lexical tier. When lexical information is lacking or sparse, segmental information is a useful cue to locate word boundaries. In the case of poor segmental information, such as in a noisy environment, the prosodic information is more useful. In the language acquisition process, this hierarchy is turned upside down. The lexical tier is obviously not yet reliably available for the segmentation process, at least not in the function of providing the optimal segmentation strategy. Of the subsegmental tiers, the stress based one seems to be available earlier. Morgan and Saffran (1995) found that 6 months olds use only stress as a segmentation cue when presented with conflicting prosodic and segmental cues, while 9 months old infants were influenced by both cues. This suggests that prosodic information is more salient, but segmental cues provide more precise information which is preferred at later stages of development.

Trajectory

The cues which are useful for the speech segmentation process all become available to infants in the second half of their first year. Starting with the recognition of familiar words at 6 months of age followed by the metrical cue at 7.5 months, the transitional probability, utterance boundary and co-articulation cues (8 months), probabilistic phonotactic cues (9 months) and allophonic cues (10.5 months) become available for the segmentation process. Jusczyk and Aslin (1995) have tested groups of 6 and 7.5 months old infants on their ability to recognize a monosyllabic word from speech after they had been presented with it in continuous speech. The task required the infants to be able to segment the target word, of which only the older group of infants was capable. The ability to segment words from continuous speech seems to develop in infants from the beginning of the second half of the first year of their life.

The different cues are used complementarily in the segmentation process, although the strength which

is given to each cue can change over time as other cues become available. Thiessen and Saffran (2003) have compared the use of statistical versus metrical cues in 7 and 9 months old infants. The two cues led to alternative segmentations. Thiessen and Saffran (2003) found that the younger infants used the statistical cue more than the metrical cue, and vice versa for the older infants. The authors concluded that infants use the statistical cue (which is purely bottom-up) to generate an initial lexicon, from which they generalize the language-specific stress pattern which is used in later segmentation tasks. The second cue is available later in the trajectory, as the infant develops.

As discussed in the previous section on the integration of cues, the type of segmentation cue that is available to infants changes over time together with the weight that is given to each type. Mattys et al. (2005) present a hierarchy for the adult speech segmentation process, with the lexical tier on top. In the developmental trajectory of the infant, the hierarchy is developed from the bottom upwards. The highly salient prosodic information, which is already available for newborns (Nazzi et al., 1998) is the cue that is applied first. The next step for infants is the more detailed segmental information necessary for phonotactic and coarticulatory cues. These lead to the start of a lexicon, which then takes on an increasingly powerful role in the segmentation process.

2.2 Building a lexicon

Once forms are segmented from speech using a combination of the cues discussed in the previous section, the forms can be stored in memory. The adult word storage, the lexicon, is a complex network of word forms, their meaning and their syntactic properties. The words are linked to each other based on different criteria and common characteristics (Ernestus and Baayen, 2003; Elman, 2004). The infant task of acquiring this complex system is simplified with the assumption that a separate word form lexicon exists. Acquisition of word forms could start separately, not requiring every form to be associated with a meaning and syntactic properties before it can be added to the lexicon. This would disentangle the acquisition of phonological, semantic and syntactic parts of the lexicon. To test this hypothesized learning procedure, it is useful to see whether infants can memorize forms that have no meaning (yet) and whether they can associate meanings to already known forms. The hypothesis would receive even more support if infants are better able to acquire a form-meaning association for forms they have learned before than for novel forms.

Graf Estes et al. (2007) have studied whether infants are more likely to learn a word-meaning association when the infants had segmented the word in a previous task than when they had not segmented the form before. Infants were subjected to the same segmentation task as performed by Saffran et al. (1996), in which they heard a 2.5 minute stream of fluent artificial speech consisting of bisyllabic words. The only segmentation cue in the input was the distribution of transitional probabilities between the syllables.

Immediately after the training phase, the 17 months old infants performed an object labeling task. They were habituated with an object while hearing either a word from the previous segmentation task, a non-word (syllable combination which did not occur in training phase) or a part-word (syllable combination which occurred with low transitional probability). After the habituation, the infant was either presented the same object-label pair or a different one. Switching to a different label only caused an effect in the infants which were habituated with words segmented in the segmentation task. Infants habituated on non-words or part-words showed no effect of whether the label presented with the object matched with the label in the habituation phase or not.

The Graf Estes et al. (2007) study shows that having the word form available simplifies the task of learning a word-meaning pair. It also shows that being subjected to speech which only contains a transitional probability cue is sufficient for infants to create word forms. Rather than only being able to distinguish between transitional probabilities of syllables in the training phase, infants are able to store the results of this statistical segmentation task in (short-term) memory. The resulting segmentations are used in the concept of an actual word form, as viable labels for an object. This seems to suggest that a form which has been segmented from speech is available for later use, at least for a short amount of time after segmentation. The results in this study support the notion of a word form lexicon. Swingley (2009) discusses some values of a word form lexicon in the lexical acquisition process, such as the possibility of generalizing over previously encountered forms to find a dominant metrical pattern.

Thus far, this chapter has presented several cues which are available to infants in the segmentation process and a trajectory of when which cue becomes available. It has also provided argumentation for the concept of a separate word form lexicon, which infants start to acquire before they start learning word meanings. This word form lexicon would be built up using the results of the segmentation process. In the next chapter, computational models of the segmentation process are discussed which apply a subset of the segmentation cues discussed in this chapter to model the process of segmenting speech into word forms. Many of these models use a form of lexicon, storing the parts which have been segmented. Often, no specification is given on the requirements or format of forms added to this protolexicon. As soon as a word is segmented by the model, it is added to the protolexicon. As Swingley (2009) has mentioned, the forms in the protolexicon can be used to gather knowledge on general word forms in the language, and used in the segmentation process as discussed in the section on familiar words above. It seems reasonable that forms are not added to the lexicon freely but rather that infants want to reach a level of certainty before they store the form in their memory. The next section discusses several factors which are relevant in the word form learning process.

2.3 Factors enabling word learning

Segmenting speech, learning the segmented parts by adding them to the lexicon and using the lexicon in the segmentation process are among the first tasks of infant language acquisition. This section is focused on experimental research about the middle part, learning words. This is the bridge between the final adult word segmentation process, which is largely done through recognition of words from the lexicon in the input speech stream, and the starting point of infants who use bottom-up segmentation strategies to build up this lexicon. The segmentation models discussed in chapter 3 are experimentally well supported in the segmentation cues they apply, but the simulation of the word acquisition process is less validated. This chapter identifies several factors that are relevant in the lexical acquisition process and that should be included in a representative computational simulation of infant word segmentation, acquisition and recognition.

2.3.1 Intrinsic factors

This section looks at intrinsic factors of word learning: which lexical characteristics determine ease of acquisition. The relevance of different factors is determined by testing their correlation with the Age of Acquisition (AoA) of a word form. The AoA of a word type is the moment in the vocabulary development that the word is acquired by an infant. AoA's of word types are estimated and averaged to get a list of which words are acquired when. Such a list can be used to determine relevant factors of word learning: what are the factors that predict the age of acquisition of a word?

Word form

The form of a word is a highly relevant factor for predicting whether it will be acquired relatively soon or late. The factor 'form of a word' is divided in two distinct factors: word length and neighborhood density. Neighborhood density is the number of neighbors a form has in the phonological space. Commonly, this is measured in all forms that differ one phoneme from the target form. For words in dense neighborhoods, there are many other word types with an added, deleted or substituted phoneme. Neighborhood density is suggested to be a highly relevant factor in the organization of the mental lexicon (Luce and Pisoni, 1998), making it an interesting factor to look at in the word learning process. The final organization of the mental lexicon could be a determining factor in which forms are acquired, and integrated into that organization, first.

Storkel (2004) has used two databases for information on the age of acquisition of words, and correlated these two lists with several factors of the acquired words. The main factor of interest was the neighborhood density, but other factors were included to test their confounding effects. The overall finding was that lexical characteristics were interdependent in their prediction of AoA: the acquisition of

low frequency words was predicted by neighborhood density, as was the AoA of short words. For high frequency words, the high number of exposures was suggested to neutralize the effect of neighborhood density. In long words there was less variety in neighborhood density, making it a less clearly distinguishing factor. The acquisition of short words was also predicted by word frequency, a factor which will be further discussed below.

High frequency words had their AoA predicted mostly by word length: the longer the word, the higher its AoA. Neighborhood density and frequency were negatively correlated with age of acquisition. More dense and more frequent forms had a lower AoA, whereas longer words had a higher AoA. The effect of neighborhood density, as a precursor of the organization of forms in the mental lexicon, seems to be present in the lexical acquisition process, albeit overturnable by word length and word frequency.

Word type

The distinction between open and closed class words is a relevant factor for word learning, given the distinct difference between the meaning or function, frequency and length of the words in the two groups. Functors occur with a high frequency, relative to content words. However, they are often reduced or cliticized in speech production making them harder to perceive (Shi et al., 2006). Experimental evidence suggests that infants make a distinction between function and content words early on in the lexical acquisition process, at 14 months of age (Booth and Waxman, 2009). Shi et al. (1999) found that even newborns were able to discriminate between lists of grammatical and lexical word forms extracted from infant-directed speech. Recognition and accurate perception of functors starts to develop around 11 months and is available to infants of 13 months old (Shi et al., 2006).

Function words are useful in the lexical acquisition process, since they signal content words and provide information on the syntactic type of the word that follows. A function word signaling the occurrence of a noun versus an adjective informs the infant to focus on category versus properties of an object respectively (Waxman and Booth, 2001). 14 months old infants were able to distinguish novel forms in a noun and an adjective setting and respectively focus on the object categories and object properties (Waxman and Booth, 2001). It seems that from the start of the second year of life, functors are categorically distinguished from content words. Their grammatical function is applied in the lexical acquisition process but their production seems to be held off compared to the production of content words (Swingley, 2009).

Word frequency

An important characteristic of a word form in determining its age of acquisition is the frequency with which it is presented to the infant. Its relevance was attested in Storkel (2004), and there are many more studies supporting the importance of word frequency in the acquisition process. Roy et al. (2009) found a significant effect of frequency as a predictive factor for age of acquisition, with a higher correlation for content than for function words. Goodman et al. (2008) found slightly different results that were

less homogeneous. A cause for the difference could be the age of the infants that were part of the used corpora. Goodman et al. (2008) discovered a changing effect over time for the frequency factor, and given the higher age of the studied infants in their research compared to Roy et al. (2009), this could be an explanation for the different results. Another interesting factor related to the effect of frequency is that if an infant starts to produce a specific word form, it is probable that the parents start to use that form more frequently in the interaction with their child. To take this possibility into account, Roy et al. (2009) have looked at the frequency of a form up to the moment the form was first produced by the infant.

The role of frequency has a positive effect on the age of acquisition: the higher the input frequency, the earlier it is acquired. Related to this factor is the requirement of exposure. An infant needs to encounter a word form often enough to be able to learn it. The factors determining what is 'often enough' are abundant. The prosodic and articulate quality of the exposure, the length of the utterance it occurs in, the amount of attention the infant and the adult have at the moment of exposure are all relevant besides merely quantity of exposure.

2.3.2 Extrinsic factors

The previous section discussed qualities of word forms that enabled fast learning. This section looks at circumstances in which words are presented to infants, and which factors in the presentation enable word form learning. The input infants receive differs from adult language. Child directed speech consists of shorter utterances, exaggerated prosody and more repetition than adult directed speech. This section looks at some of the factors of child directed speech to see whether they are predictors for which word forms are acquired by infants.

Available word forms under different conditions

In a corpus-analysis study, Swingley (2009) discusses the issue of how many word forms an infant could learn, based on a few requirements. The author has plotted the number of word types which occur with certain frequencies in two distinct environments: presented in isolation or occuring utterance initially or finally. Assuming that some or all of these factors – frequency of occurrence, being perceived in isolation or at an utterance boundary – are relevant for when an infant acquires a form, the graphs show a rough estimation of the number of word types available under those conditions. Using the Brent corpus of infant directed speech (Brent and Siskind, 2001), Swingley plotted the number of word types occurring in two circumstances against the number of perceived tokens. The two plots are presented in figure 1.

The different lines in both graphs represent the frequency factor. Each line stands for a postulated number of times a word type needs to be encountered before it is learned. The frequency thresholds represented by the lines have the values of 1, 5, 20 and 50. The two circumstances in which words can be presented to enable learning are presented respectively in the two graphs. The first graph plots the

number of word types presented in isolation, the second graph plots the number of word types presented utterance initially or utterance finally. In short, the graphs show the development of the number of word types available to an infant over amount of speech input, when words are acquired after they have been encountered a certain number of times (the different lines) in a certain situation (in isolation for the left graph, utterance initially or finally in the right graph).

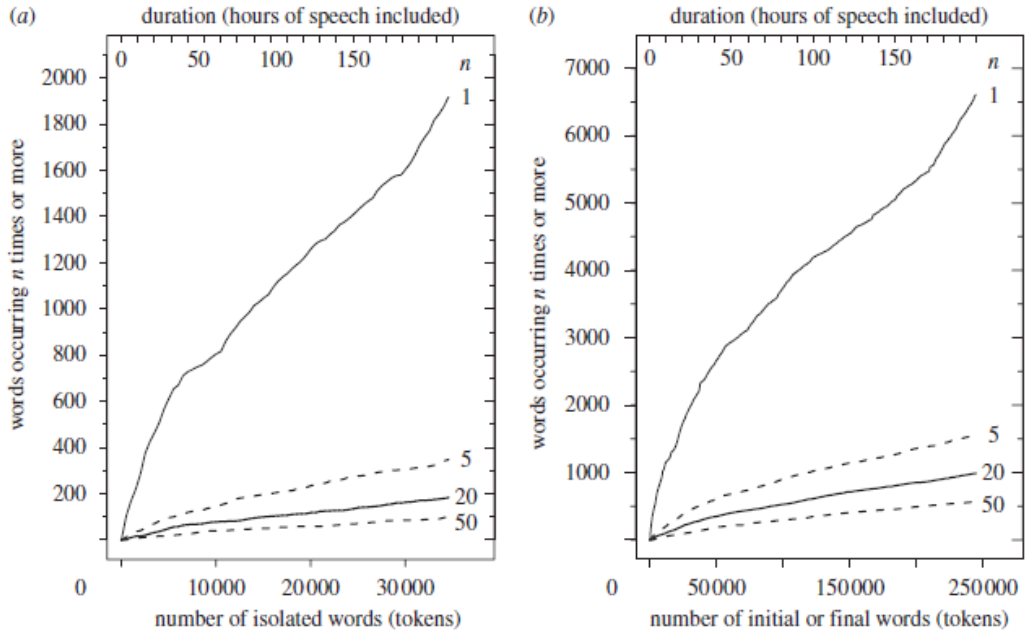


Figure 1: Estimation of number of word types available to infants plotted against both the number of perceived word tokens and hours of perceived speech, data gathered from the Brent infant directed speech corpus (Brent and Siskind, 2001). The lines in both graphs represent the growth of the number of available word types for different frequency thresholds n with values 1, 5, 20 and 50. The graph in a) plots the number of known word types with different frequency thresholds for words in isolation. The graph in b) plots the number of known word types with different frequency thresholds for words occurring utterance initially or finally.

The graphs show that there are three to five times as many words available in the environment of words occurring utterance initially or finally than words occurring in isolation. From this it can be concluded that it is useful to pay specific attention to forms occurring at utterance boundaries. Another interesting fact resulting from figure 1 is the drop in available word types when the frequency threshold increases from 1. Apparently only 10-15 % of word types occur 5 times or more, suggesting that the requirement of needing to hear a form multiple times is a strong restriction on the available data.

It is important to clarify that 'available words' is not the same as words known by infants. It is unknown how many times a form needs to be perceived before it is part of the infant's lexicon. The isolated words need to be recognized as such in order to be correctly acquired; the words at utterance boundaries need to be partly segmented first before they can be added to the lexicon. On the other hand, experimental studies have shown that infants are able to moreover segment words from utterance medial positions, making the source of words available to infants even larger. What can be concluded is that

when infants start to utter their first words, they already have at least dozens of words at their disposal.

The next part discusses the environment of words occurring in isolation in more detail. It is clear from figure 1 that words in isolation are a smaller source for learning word forms than words occurring at an utterance boundary. But after 200 hours of speech, which is estimated to be reached in around six weeks, infants have encountered 100 to 1900 distinct word types.

Isolated words

Brent and Siskind (2001) present a corpus study focused on the role of isolated words – utterances consisting of only one word – and its influence on word segmentation and word learning. The authors were interested in the usability of isolated words in the word segmentation process, whether they are frequent and diverse enough to be used for boot strapping other words from continuous speech. A corpus was compiled by visiting eight mothers for 14 separate recording sessions of 90-120 minutes each and transcribing the resulting recordings. Using this corpus, the authors were able to draw several conclusions. Among the results that Brent and Siskind found, was that of the words produced by the infants – production data were also collected by the authors – around 75% had been presented in isolation. This high proportion of words being presented in isolation recurring in the infant's vocabulary is an indication of the effectiveness of the isolation environment discussed above.

In the previous section, frequency was presented as a relevant predictor of how fast a word would be acquired. Brent and Siskind (2001) found contradictory results. They performed a linear regression analysis for each form that had been produced at least once by one of the infants. The predictor variables were both overall frequency of the word produced by the mother and frequency with which the form was produced in isolation. The authors found that for all measures of the infant's productive vocabulary, only the frequency with which words were perceived in isolation was a predictive factor for whether the infant produced the form. The overall frequency with which a form was uttered by the mothers was not predictive for whether the form was part of the infant's vocabulary. Given the fact that other studies did find a significant correlation of age of acquisition and frequency with which a form was produced by the care givers (Huttenlocher et al., 1991; Goodman et al., 2008; Roy et al., 2009), these results will be considered as only specific to this study.

Isolated words seem to present an advantage to infants; they are more easily learned. By presenting single word utterances, parents enable their infants in the word learning process. In the next part, the role of the parents is discussed in the language they present to their infants.

Parental tuning

The Human Speechome Project (HSP, Roy et al. 2009) was set up to construct a complete picture of the language input an infant receives in the earliest years of life. 14 microphones and 11 cameras were placed in the home of the researcher, to track the linguistic and contextual input of a new born for the

first three years of its life. This resulted in 230,000 hours of recordings, among which could be heard the first words uttered by the infant and all the input it had received to get to that point. An analysis of the speech uttered in 72 days spread out over the period the infant was 9-24 months old, led to a corpus of 10.2 million words of infant directed input and child produced data. From this corpus, Roy et al. (2009) and Roy (2009) gathered interesting information on the word learning process of the recorded infant.

This period in the collected data was selected specifically to track the development of the infant's vocabulary. The authors were interested in the 'birth of a word' and what led up to the moment a word type was produced for the first time. One of the main factors predicting the age of acquisition of a word form was the frequency with which it was uttered in the child directed speech data (Roy et al., 2009). All word types, from closed and open classes, had a correlation between their age of acquisition and frequency of utterance. Another interesting factor was the influence of prosody, measured in the length of the vowel denoting emphasis. The age of acquisition had a higher correlation with a combined factor of frequency and prosody than with either one separately (Roy, 2009).

One of the most interesting discoveries of this subset of the collected data, was the way the care givers tuned their language to the level of the infant's linguistic abilities. Roy et al. (2009) distinguished two types of tuning: coarse and fine. The former is a general adjustment to match the infant's development, for example visible in the mean length of utterances, with extra focus on the number of isolated words, and the lexical diversity measured in type-token ratio. The fine tuning of parental speech was particularly interesting. The mean length of utterances containing a specific word was shortened over a long period, up to the point the infant first produced the word. For all three care givers, for hundreds of word forms acquired by the infant in a period spanning 15 months, the lengths of utterances containing a specific word type dropped significantly over several months until that form was correctly produced by the infant for the first time. These findings suggest that care givers are distinctively and explicitly, though probably subconsciously, presenting word types to their infant.

Frequency peaks

The previous part presented an aspect of child directed speech that suggested that parents fine tune their child directed language output to facilitate word learning. Word forms are produced in increasingly shorter utterances, up until the point the form is correctly uttered by the infant. Word forms are provided in a way that makes them easier to acquire: with a high frequency and in a context from which they are easily segmented. This part discusses a form of word type presentation that also enables word learning, namely a peak in form frequency.

Jusczyk and Hohne (1997) exposed 8 months old infants at home to 30 minutes of speech for 10 days in a 2 week period. The prerecorded speech contained short stories and a list of words, half of which were forms occurring frequently in the stories and half of which did not occur in the stories. Two weeks

after the last exposure session, the infants were tested in a lab for retention of the forms. The infants showed a significant longer listening time for words which were both in the stories and the list compared to the words in the list which were not part of the stories. The infants' preference for the story words also resulted in a longer listening time than that of a control group of infants who did not take part in the exposure sessions but only took part in the retention test. These results suggest that infants can segment frequent forms from fluent speech and store them in memory for at least a period of two weeks.

In this study, the forms were presented with a peak in frequency. The stories contained specific word forms that were not part of the infant's every-day input, they only perceived them in high concentration for 30 minutes a day. This repetition of word forms within a short period of time was beneficial for learning the tested word forms. Situations of short periods of high concentration of certain word forms are normal for infants. They occur in story telling or when infants and care givers interact using an object that is frequently named by the adult. Only looking at isolated words, Brent and Siskind (2001) determined in their corpus that 27.2% of word types occurred more than once in a 30 second period. Predicting these words to be the topic of the conversation, their peak in frequency can be a useful source for word learning.

2.4 Preliminary summary

This chapter has presented a review of the experimental research on infant word segmentation, the infant lexicon and infant lexical acquisition. Several segmentation cues have been presented that are useful in the segmentation process and that are available to infants: they are sensitive to the cues and able to apply them for segmentation in a laboratory setting. The concept of a word form lexicon has been presented, a lexicon only containing phonological information as a simplified version of the adult lexicon. This protolexicon as a simplifying step in the lexical acquisition process seems to be available to infants, who are shown to be able to more easily recognize and acquire the meaning of a form that has been encountered before in a semantically empty context (Graf Estes et al., 2007).

The prefinal section of the chapter presented experimental research on word learning. Factors inherent to the word forms that are relevant to infants in the acquisition process and factors inherent to the environments in which word forms are presented to infants were discussed. Two of the factors, one of each type, are included in the segmentation and word learning model presented in section 4. These are the frequency of word forms and the peak in frequency with which forms are presented to infants. This chapter has presented an overview of the available evidence on the infant word segmentation and word learning process, the cues and factors that infants have at their disposal.

3 Modeling infant speech perception

This chapter presents an overview of computational literature on word segmentation. The chapter is organized by the function of the lexicon in the segmentation models. Two adult speech perception models are presented, these have a full lexicon available and rely mostly on word recognition to locate word boundaries. Because the creation of the lexicon as a result of the segmentation process is of central interest in this thesis, the adult segmentation process is included in order to understand the final and optimal situation. As discussed in the previous chapter, the preferred segmentation strategy for adults is the opposite of the infant's developmental trajectory (Mattys et al., 2005). Understanding the final stage of the trajectory is informative for the development.

There is one other major distinction between the segmentation models presented in this chapter in the role of the lexicon. Two models will be presented that do not include a lexicon in the word boundary location process at all. These models are focused on predicting the location of word boundaries from the signal only, without including the knowledge which becomes available from a growing lexicon. The final sections present models that combine both types of cues: lexical cues from a lexicon that is being built as a result of the segmentation process, and sublexical cues to create this lexicon. For the latter type of segmentation model, the chapter discusses where feasible the procedure with which forms are added to the lexicon. As discussed in the introduction, most segmentation models do not include a representative lexical acquisition process. This encompasses that models that built up a lexicon in the segmentation process acquire a word form as soon as it has been segmented from the input. The model presented in the next chapter has two factors that are part of the infant's reality of lexical acquisition.

The next section presents an introductory overview of how the segmentation models function, together with an overview of the used input corpora and the evaluation methods of the produced outputs of the models. This information is relevant to understand and compare the performance of the models presented in the other sections of this chapter. Section 3.2 presents two adult segmentation models which use a lexicon to segment continuous speech into word forms. Section 3.3 presents the infant's first task of locating word boundaries without having a lexicon available. Using information which is directly available in the input, the models presented in this section identify word boundaries with a purely bottom-up strategy. Section 3.4 presents models that combine a bottom-up segmentation strategy with a top-down word recognition strategy. The models discussed in this section start with an empty lexicon, which receives input from the sentences that are segmented with a strategy based on sublexical information. The parts in the segmented sentences are 'learned'. They are added to the lexicon, from which they are used in the segmentation of following sentences. The final section discusses a subset of segmentation models which combine lexical and sublexical strategies with a Bayesian approach. In determining the best

segmentation of a sentence, Bayes' rule is used to determine the probability of separate segmentations.

3.1 Input corpora and evaluation measures

In the following sections several segmentation models will be presented. This section presents an overview of the input and the output of all the different models, while the following sections focus on how the segmentation process is modeled in different ways. The models presented in this chapter turn a text in which no word boundaries are indicated into one including word boundaries. In order to perform this process, each model uses a different set of cues and a different procedure for both segmenting the input and acquiring the necessary information for this segmentation process. In this section, the different corpora which are used as input text are presented first, with an overview in table 1. Subsequently, the types of output of the segmentation models are presented together with the ways in which they are evaluated.

Model input: corpora

An overview is given in table 1 of all the corpora that are used in the models presented in this chapter. The corpora differ from each other on a few factors. One factor is the type of speech, whether it is child-directed speech or adult conversation. As mentioned before, the stage of adult speech processing with an extensive lexicon available is of interest as the final stage towards which infants develop. The adult segmentation models discussed in this chapter work differently than the infant models in that they do not segment continuous corpus texts into strings of words. The segmentation tasks they perform are based on the forms in the lexicon with which they have been equipped. There is one corpus, used by two of the presented models, that contains adult-directed speech. This corpus is selected in those cases because it is phonetically transcribed. Most corpora are phonemically transcribed, with each word transcribed to its canonical phonemic form. The CGN corpus (see table 1), contains broad phonetic transcription, including the variation with which the words are pronounced in speech.

Besides the phonemic information which is present in all the input texts, some corpora contain stress information on the word level. All corpora contain utterance boundaries, which are considered salient in speech. One corpus also contains syllable boundaries. The language of the corpus is another differing factor; the corpora in table 1 are either in English or Dutch. The corpora vary in size, presented in the table in number of tokens. Each corpus in table 1 is given a code, used to denote which corpus was used by a model.

Code	Corpus	Description
Ch	CHILDES MacWhinney and Snow (1985)	Collection of corpora of transcribed spontaneous interactions between children and care-givers in multiple languages. The English part of the corpus contains around 750,000 tokens directed at children (of any age).
CBr	CHILDES Brown (1973)	One of the English CHILDES corpora, which is phonemically transcribed including stress and syllabification through onset-maximization by Gambell and Yang (2005). Contains 37,500 tokens.
CBR1	CHILDES Bernstein-Ratner (1987)	One of the English CHILDES corpora, containing speech from nine mothers to their infants with an average age of 18 months. Orthographic corpus is phonemically transcribed by Brent and Cartwright (1996) and contains 33,399 tokens.
CBR2	CHILDES Bernstein-Ratner (1987)	Modification by Blanchard and Heinz (2008) of the Bernstein-Ratner corpus as transcribed by Brent and Cartwright (1996).
CK	CHILDES Korman (1984)	One of the English CHILDES corpora, containing speech directed at preverbal infants of 6-16 weeks old and consisting of 37,549 tokens. Phonemically transcribed and provided with lexical stress information by Christiansen et al. (1998).
CGN	Corpus Gesproken Nederlands Oostdijk (1999)	Corpus of Spoken Dutch, corpus of adult speech in broad phonetic transcription (Goddijn and Binnenpoorte, 2003) containing pronunciation variation. The StaGe segmentation model (Adriaans and Kager, 2010) used a subset of the corpus containing 660,424 tokens.
VW	Van de Weijer Van de Weijer (1998)	Dutch corpus of speech recorded in the presence of a child, phonemically transcribed.

Table 1: Corpora

The input texts are presented to the models without indication of the word boundary locations. Each model focuses on finding either the word boundary locations, the words in the input text (tokens), the word types that will build up the lexicon, or a combination of the three. The output is a segmented version of the input text and in some cases a lexicon of word types, which often includes word frequencies. This output is evaluated quantitatively using one of two methods discussed next.

Model output: evaluation measures

There are two ways to calculate the quantitative performance of segmentation models: information retrieval metrics (precision, recall, F0) or using the Signal Detection Theory (Peterson et al., 1954). In either case, it is necessary to measure two performance scores. Both methods calculate the percentage of correct word boundaries detected by the model out of all the correct boundaries, this is called *recall* or *hit rate*. Since placing a word boundary after every phoneme also leads to a perfect recall or hit rate, a second measure is required. This measure either calculates how many of all the boundaries that were placed were correct (*precision*) or the percentage of non-word boundary locations that were predicted to be a word boundary by the model (*false alarm rate*). The precision, recall and hit rate measurements should be as close to 100% as possible for a model to perform well, the false alarm rate should be as

low as possible. The precision measure is sometimes called accuracy, in which case the recall is called completeness. In the presentation of the models' performance, the terms precision and recall are used rather than accuracy and completeness.

Besides how the performance is measured, there is a difference in what is measured. The scores can be given for tokens, types or boundaries. When the scores for the tokens are given, the measurements for the segmented parts in the test text are: correctly segmented words (hits, a boundary before and after the word and none in between), missed words (words in the text that are not found by the model) and false alarms (words in the text segmented by the model that are not in the actual text). The word types are measured in the lexicon, whether they are correctly identified, whether they are segmented but no real words or an actual word type in the text is not present in the output lexicon.

In the case of measuring word boundary performance, any correct boundary is tested to see whether it is found by the model and any boundary found by the model is tested for correctness. Signal Detection theory is only applicable to evaluate the location of word boundaries, not to evaluate word types or tokens. In the calculation of the false alarm rate (see (1c.)) a value for *True Negatives* needs to be given, which is the number of possible boundary locations that were correctly identified as not being a boundary. This measure is difficult to calculate for the correctness of word tokens or types since it requires to calculate the total number of possible parts in the corpus that are not actual words nor defined by the model as such. In the tables presenting the models' performance, the abbreviations presented in table 2 will be used. Remember that hit rate and false alarm rate can only be calculated for the boundaries.

	Precision	Recall	F0	Hit rate	False alarm rate	d'	A'
Boundary	BP	BR	BF	BH	BFA	Bd'	BA'
Token (Word)	WP	WR	WF	-	-	-	-
Type (Lexicon)	LP	LR	LF	-	-	-	-

Table 2: Abbreviations

Because two measures are necessary to give a complete evaluation of the performance results, each evaluation method has a third measure which combines the two measures and gives an overall evaluation of the performance. The precision and recall measures can be combined into a measure termed the F0-score. Its calculation is presented in (1). The hit rate and false alarm rate can be combined in two distinct ways, the d-prime, which stands for discriminability index, or A-prime (also written as d' and A'). D' is called the discriminability index because it measures the distance between the hit and false alarm rates. Because the former needs to be as close to 1 and the latter as close to 0 as possible, the larger the difference between the two the better the model's overall performance. D-prime can rise up to its maximum score of 6.93, although values up to 2.0 are usual. A d' score of 1.0 denotes 69% correct placement of boundaries and correct identification of absent word boundaries (Keating, 2005). The d-prime measure is the most

commonly used combinatory measure of hit rate and false alarm rate. Another combination of the two rates is presented in the A'-score (A-prime), which requires fewer assumptions about the distribution of the performance of the model, such as the assumption that the two rates have an equal variance (Pollack and Norman, 1964; Lignos, 2012).

All formulas for the discussed measures are presented in (1). TP stands for *True Positives*, the number of boundaries correctly located by the model (hits); FP stands for *False Positives*, when the model identified a boundary or word which does not occur as such in the actual text (false alarms); FN, *False Negatives* are missed boundaries or words that are not but should have been identified by the model (misses); *True Negatives* (TN) are locations that could have been boundaries but are not in the actual text and that are not identified as such by the model. H stands for Hit rate, F for False alarm rate.

(1) a.

$$Precision = \frac{TP}{TP + FP}$$

b.

$$Recall/Hitrate = \frac{TP}{TP + FN}$$

c.

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

d.

$$F0 = \frac{2 \times precision \times recall}{precision + recall}$$

e.

$$d' = z(H) - z(F)$$

f.

$$A' = \frac{1}{2} + \frac{(H - F)(1 + H - F)}{4H(1 - F)}$$

3.2 Word recognition

When modeling adult speech perception, a lexicon is provided to a model which then proceeds to match (recognize) the discrete forms in the lexicon to the continuous input string. Two successful models of adult speech perception are the TRACE model (McClelland and Elman, 1986) and Shortlist(-B) (Norris, 1994; Norris and McQueen, 2008). Of course, infants do not possess a lexicon yet, this has to be acquired during the segmentation process. To understand the parsing process when a lexicon is available, the TRACE and the Shortlist model are presented in this section. These models provide insight in the word recognition process, which is applied alongside sublexical word boundary cues in the segmentation models that include a growing lexicon. The TRACE model combines bottom-up and top-down strategies to recognize words, whereas Shortlist only relies on bottom-up information from the phoneme level.

3.2.1 The TRACE model

Almost three decades ago the TRACE model of adult speech perception was presented (McClelland and Elman, 1986). This model is both capable of correctly inferring words from a string of feature patterns and modeling several psychological phenomena of the speech perception process such as recoverability from mispronunciation of part of a word. The former is the focus of this section, how the model is able to parse a string of words from a string of feature patterns through word recognition. Using a pre-defined lexicon, the model is able to activate words and influence the perception process top-down from the lexical level.

The TRACE model is an interactive-activation model that successively activates nodes on three levels: the feature level, the phoneme level and the word level. The model looks at patterns of activation. A specific pattern of activated features leads to the activation of a specific phoneme, and a pattern of several activated phonemes leads to activation at the word level. The model is interactive in that there is a continuous flow of information between the phonemic and the lexical level. In the activation process, the phonemic level influences bottom-up activation of lexical forms and the lexical level uses top-down influence to (help) activate certain phonemes.

The time dimension of speech processing is represented by consecutive copies of the units representing each feature, phoneme and word. Every feature pattern that is perceived activates a pattern on a new set of features, and similar for the phonemes and words. This enables the model to keep previously received information available for later processing, and to let current activations influence the processing of following input, thus enabling the representation of phonetic context. However, this is an implausible representation of the word recognition process, and limits the model to a small lexicon only containing a small set of words. The latter is also a consequence of the fact that at the lexical level activation of a form inhibits activation of other forms. Lexical competition is represented with inhibitory connections between all forms, making the model complex and sensitive to the number of forms in the lexicon.

The TRACE model receives feature activation patterns, representing different phonemes. No explicit segmentation information is given, such as syllable boundaries or stress. One specific activation pattern represents silence. This pattern is used at utterance boundaries or to denote word boundaries in non-segmentation tasks. The word level, representing the lexicon, consists of a list of 211 words in their phonemic transcription. There are no effects of word frequency incorporated at the word level; each word is considered to be a priori equally probable to be produced.

The TRACE model is an identification model. It receives information and identifies it as a pattern representing a form at a higher level of representation: feature nodes are activated and they activate nodes on the phoneme level. In a segmentation task, the authors presented the model with two consecutive

words, not separated by a pause. For 189 out of the 211 presented pairs, the words that received the highest activation on the word level were the ones that had been presented to the model in the form of feature patterns. In the other cases, the model either activated two words with another location of the segmentation boundary, two words with overlapping phonemes (e.g. /bus/ and /top/ were presented, but /bus/ and /stop/ received the highest activation), or a word was activated which did not cover all the activated phonemes. The segmentation task was performed through competition of the lexical forms. Goldinger et al. (1989) have shown that in adult speech recognition, phonetically similar forms are indeed inhibited by other forms as a result of the competition process.

The TRACE model has been tested on segmentation with novel forms. A known word, 'target', was either following another known word, 'possible', or a novel form 'pagusle'. In the latter case, the recognition of the /t/ as word initial phoneme occurred only after the entire form 'target' was processed. The word 'possible' was activated with such a high level of activation that even before the /t/ was processed it was clear that it would be word initial. The recognition of words in the input string shortens the segmentation process and lowers the uncertainty through a lower number of possible segmentations. Placing one boundary in a string of two words is not equal to the actual segmentation task of receiving a long string of sounds. As stated by the authors (McClelland and Elman, 1986, p.64), more cues than the top-down influence from the word level are needed in a harder segmentation task.

The TRACE model is able to recover the words and phonemes in situations where the input is noisy and imperfect, it can identify sounds even with partial or ambiguous information making use of its interactive architecture. As a segmentation model, the model is able to place a word boundary on the correct location between two forms in 90% of cases, using a lexicon in which all forms are available. A segmentation task involving a novel form significantly slows down the model and lowers its certainty level with which forms are activated. In situations with multiple novel forms, more novel than known words, or no information from the lexicon, other cues are needed to parse a string of phonemes into a string of words.

3.2.2 Shortlist and Shortlist-B

Norris and colleagues have proposed a speech perception model similar to TRACE but based on more plausible assumptions (Norris, 1994; Norris et al., 1995). Contrary to TRACE, the basic architecture of the Shortlist model is two-staged. First, the model selects a set of word forms based on the phonemic input. These forms form the 'shortlist' that are sent to the second stage. There, the forms compete with each other to determine the final winner. In TRACE, this process occurs on the same level, requiring connections between all word forms in order for them to compete with each other. The Shortlist model only requires inhibitory connections between a small set of words simultaneously, namely only the ones that compete with each other because they are in the shortlist. This enables the model to be computationally

less complex and handle a significantly larger number of word forms in the lexicon.

A second large distinction between Shortlist and TRACE is the interaction aspect. The TRACE model is highly interactive in that information flows top-down and bottom-up between the levels simultaneously and actively. The Shortlist model functions purely in a bottom-up fashion. The phoneme level selects the shortlist, which is then sent to the lexical level. Lexical constraints are only active on the lexical level, they have no influence on a lower level. According to Norris (1994), this is a more realistic representation of adult speech perception. Experimental studies show that top-down feedback does not play such an active role as in the TRACE model.

One of the biggest problems with the TRACE model's architecture is its need for a copy of the lexical framework at every moment a new word can begin. These copies are required in order to enable lexical competition between words which are simultaneously activated. In Shortlist, a recurrent network is used to replace part of the architecture as a more plausible alternative. Instead of a copy of the full set of nodes for each moment in time, the Shortlist model just needs one set. At the output of the model each lexical form has its own node the activation of which is conditioned on the activation of nodes it is connected with. To represent the influence of preceding and following sounds, every node has a hidden unit which contains information on the activation of the node in a previous cycle, and the cycles before that. This 'memory' of each node is capable of identifying a word as soon as it is the sole candidate left that matches with the input.

This single set of lexical nodes creates a shortlist which is sent to the second stage. On this second stage, lexical competition takes place by having each form inhibitorily connected to the other shortlisted forms. Instead of requiring each node in the lexicon to be connected to each other node at every moment in time, as in TRACE, the Shortlist model only requires this interconnectedness of the shortlisted set which maximally contains 30 forms. The two stages enable the model to split up the representation of the time dimension in recognizing the input forms from the lexical competition between these forms.

The Shortlist model has been subjected to a similar segmentation task as the TRACE model. A set of two words was presented to the model, containing other possible words which the model had to discard. In the case of 'holiday weekend' (/hɒlədeɪ wi:kɛnd/), the model activates several competing words (/hɒt/, /hɒloʊ/, /wi:k/) but the activation levels become significantly higher for the actual words as soon as the competing forms no longer conform with the input. The model also comes to the correct output in the case of 'ship inquiry' (/ʃɪp ɪŋkwaɪri/), although it takes longer to activate the correct forms. This is because the form /ɪŋkwaɪri/ needs to be unambiguously activated first in order to let /ʃɪp/ be higher activated than /ʃɪpɪŋ/. Because /ʃɪpɪŋ/ receives a higher amount of support from the input since it has more phonemes activated, and there are forms that can comply with the subsequent phonemes (/kwaɪt/, /kwaɪət/), the entire form needs to be processed in order for /ɪŋkwaɪri/ to win, which is required to let /ʃɪp/ win from

/ʃɪpɪŋ/. In the latter case, locating the word boundary is not a straightforward task given the input, but the Shortlist model is able to identify the correct forms.

The segmentation process of locating a word boundary as performed by the Shortlist model (Norris, 1994) is based on competition between lexical forms (Goldinger et al., 1989). In a follow-up study, the model was extended with a separate segmentation strategy to help select a word from a shortlist of possible forms (Norris et al., 1995). Norris et al. (1995) integrate a Metrical Segmentation Strategy (MSS) into Shortlist. This is done through two procedures: the Boost procedure leads to an increase of activation for words which start at onsets defined as belonging to a strong syllable. The Penalty procedure subtracts activation of words in the shortlist which do not contain a strong onset when there is a strong onset in the input. For the integration of these two procedures, the assumption is made that the model identifies a syllable as strong from the first phoneme in its onset. This is not very plausible, usually the vowel signals a strong syllable or the weight of the rhyme does. The combined integration of the two procedures led to a correct modeling of the results found in an experimental study in which subjects had to perform a segmentation task based on both lexical competition and metrical information.

A second extension to the Shortlist model is presented in Norris et al. (1997). The extension is a Possible Word Constraint (PWC). As discussed in section 2.1, the segmentation of a string which leads to forms which are not syllabic is dispreferred. People have more trouble recognizing /æpəl/ in /fæpəl/ than in /vʌfæpəl/ because in the former, the consonant /f/ remains as a residue but cannot form a word on its own. Norris et al. (1997) have incorporated the PWC by penalizing the activation of a word which would lead to a non-syllabic residue in the parsed string: if there are only consonants between a possible boundary and a known boundary such as an utterance boundary, the possibility of the boundary is reduced. This implementation of the PWC correctly simulates the results of the word recognition task.

A new version of the Shortlist model, called Shortlist-B, is presented in Norris and McQueen (2008). The basic principles are kept the same but the architecture and input to the model are changed drastically. One difference is the input the model receives. This is more realistic than the input to the Shortlist model in that it is more continuous. Instead of discrete categories of phonemes which are presented one at a time, with 100% certainty of the identity of the phoneme that is presented, the Shortlist-B model receives phoneme probabilities. This adjustment makes Shortlist-B a more realistic representation of the speech perception process. The second distinction between Shortlist-B and its predecessor is more interesting to the segmentation procedure, mostly because the new model is built in a framework which has been applied more often in computational segmentation models (see section 3.5).

The Shortlist-B model is built on Bayesian principles, rather than as a connectionist network. This entails that instead of a network of nodes and connections in which words directly compete through their activation level and by inhibiting each other, the model functions on finding the most probable path

through a network of the word forms. Each path forms a hypothesis of the input in the form of a sequence of words. Similar to the Bayesian models that will be discussed later on, Shortlist-B has a stepwise process of calculating the probability of a specific segmentation. The probability of a specific segmentation is conditioned on the 'evidence', or the perceived input. A shortlist of words is created and each is given a conditional probability based on the amount of overlap of the phonemes perceived and those in the word. A path through the generated words is created, with the aim of having each phoneme represented in a word exactly once. The probability of different paths is calculated based on the words they contain. The segmentation with the highest combined probability is the definite one. This stepwise calculation of different probabilities will be seen in different models of the infant segmentation process as well. The Bayesian version is an improvement on the connectionist Shortlist model because it makes less assumptions about the architecture of the model and less assumptions are needed to calculate experimentally attested response time differences from the computational data, which is not as straightforward in the connectionist version as it is in the Bayesian model.

3.3 Word segmentation

The previous section presented the adult segmentation process, which relies highly on word recognition and the knowledge provided by the lexicon. This section is focused on the opposite situation, when no lexicon is available and all information needed to locate word boundaries has to come from the signal. In this section, two models are discussed that do not use a lexicon in their segmentation process. The models are focused on correctly locating word boundaries only using the information directly available in the input. Unlike the models discussed in the next section, there is no specific interest in the words resulting from the segmentation process. Rather, the purely bottom-up segmentation models are a test of the usefulness of segmentation cues. As discussed in section 2.1, finding that an infant is sensitive to a segmentation cue is not the whole picture. This cue should also be determined as useful to locate word boundaries in the segmentation process. The two models presented here model different segmentation cues to discover their usefulness in the segmentation process. Both models use a bottom-up strategy based on cues that are present in the input string to get a probabilistic estimate on where to place word boundaries. Whereas in word recognition using a lexicon the task is to locate words, in word segmentation the task is focused on locating word boundaries.

3.3.1 Christiansen, Allen, and Seidenberg (1998)

The model created by Christiansen et al. (1998) is a simple recurrent network (SRN). This is similar to the interactive-activation model of the TRACE model, because it is a set of nodes connected with each other through connections that have different strengths. Whereas the interactive-activation model contains both

facilitatory and inhibitory connections – simultaneously activated words or phonemes inhibit each other's activation – the connections in the simple recurrent network only have positive weights. The network is trained on a set of training data, and subsequently tested on its ability to predict the next phoneme, or a word boundary, in a sequence.

The focus of the article by Christiansen et al. is on the combination of segmentation cues. Johnson and Jusczyk (2001) have shown that 8 months old infants are able to integrate different segmentation cues. Three cues in the SRN are presented in different combinations to see the usefulness of each for the prediction of word boundaries. The three cues are: the location of utterance boundaries, probabilities of phoneme sequences and lexical stress. When the model is trained on all three cues, the training data contains utterance boundaries, the phonemes are specified for their phonetic features and the syllables are marked for either no, intermediate or strong stress. When all three cues are present in the training data, the input the model receives is either the activation of a special node for the utterance boundary, 'Ubm', or phonetic feature patterns combining into phonemes together with lexical stress information. As output, the model activates either the 'Ubm' node in a separate set of output nodes, or one of the phonemes together with a stress identification. If not all three cues are presented to the model, both the input and the output show no activation for the corresponding cues.

The model performs best when trained on all three cues. All performance measures show a significant difference between the results of the best performing condition and the second-best, which is the model trained on phoneme and utterance boundary information but no lexical stress information. The difference in performance between the model trained on all three cues and on phonemes and utterance boundaries alone is larger when measuring the set of words resulting from the segmentation than when measuring the placement of word boundaries. It seems that the lexical stress cue has a more derived effect of resulting in well segmented words than the immediate effect of correctly predicting word boundaries. Overall, the model performs better in locating boundaries than in locating word tokens.

For this and every following model in this chapter, the quantitative results of the best performance of the model are presented. The same results as presented in the relative article are reported in order to facilitate model comparison. All test results reported in this thesis are taken from the literature. In the case of the simple recurrent network of Christiansen et al. the best performance is achieved with all three cues available to the network:

Model	Corpus	WP	WR	WF	BP	BR	BF
SRN, 3 cues	CK	42.71%	44.87%	43.76%	70.16%	73.71%	71.89%

Table 3: Results Christiansen, Allen, and Seidenberg (1998)

The SRN was trained on a set of data and tested on its ability to generalize from the training data to un-

seen forms: for example, generalizing the presence of a word boundary following a phoneme sequences after seeing this sequence in front of an utterance boundary in the training set. The model presented in the next section used generalization as one of its two learning strategies to predict the locations of word boundaries.

3.3.2 Adriaans and Kager (2010)

Similarly to the simple recurrent network of Christiansen et al. (1998), the segmentation model of Adriaans and Kager (2010) only uses information directly available in the input to predict the location of word boundaries. Contrary to the previously discussed model, however, the Adriaans and Kager model is a constraint-based model. It creates phonotactic constraints based on the training data and uses these to evaluate a set of test data. The result of the model is, besides the evaluated test data, a phonotactic grammar of the learned language. The constraints are combined in an Optimality Theoretic framework (Prince and Smolenksy, 1993). Each constraint is given a weight, the weights determine the ranking of the constraints. A 'heavier' weight means a higher ranked constraint, which has more influence on deciding whether a boundary should be placed than a constraint with a lower weight. The constraints are strictly dominated: in a competition between two candidates – in this model always two phonemes with or without an intervening word boundary – the candidate that violates the highest ranked constraint loses directly.

How are the constraints created that make up the grammar? There are two strategies that subtract information from the input data and transform this into constraints. The first is statistical learning, the second generalization (together: StaGe, the name of the model). The model looks at biphones – two subsequent phonemes – calculates the probability of the two phonemes co-occurring (statistical learning) and makes generalizations between different phoneme pairs based on shared features underlying the phonemes. To be able to perform the generalization task, an infant needs to have a fully specified representation of each phoneme on the feature level available. This assumption is experimentally supported. Jusczyk and Aslin (1995) have found that 7.5 months old infants did distinguish non-words which differed in only one or two phonetic features of their initial segment. Furthermore, Davis (2010) has shown that the StaGe model works similarly well if some feature contrasts are neutralized.

The co-occurrence probability of a phoneme pair is calculated based on the observed/expected ratio (O/E), the ratio of observed probability of the biphone to the expected probability of the biphone. The intuition behind this measure is as follows: if two phonemes are expected to occur together frequently, because both phonemes separately have a high probability of occurring, but they are not observed to occur often together, there is a high chance that the two should be separated by a word boundary. This is because the observed probability of two phonemes that can occur together inside a word is higher than if

they can't occur together. In the latter case the two phonemes will only form a pair when the first is the end of a word, followed by a word starting with the second phoneme of the pair. This probability is a lot lower than when the pair occurs within single words as well.

The observed probability is the probability of finding this phoneme pair given all phoneme pairs in the training set. The expected probability is calculated by taking together the total probability of the first phoneme being the first phoneme and the total probability of the second being the second. The expected probability is the probability of these two phonemes in these two position, independent of each other. The formula in (2) shows the calculation of the O/E value, the Y and X stand for any possible segment respectively following x or preceding y:

(2)
$$\frac{O(xy)}{E(xy)} = \frac{Pr(xy)}{\sum Pr(xY) * \sum Pr(Xy)}$$

The O/E ratio is useful for statistical learning, but there are more ways to calculate the probability of a phoneme pair. One other type is mutual information, which takes the $\log_2$ of the O/E ratio. A more diverging measure used in statistical learning is transitional probability, which is discussed further on in this section as well. Transitional probability calculates the probability of a certain phoneme following another phoneme, rather than two phonemes co-occurring. The formula for transitional probability is the same as the one presented in (2), minus the $\sum Pr(Xy)$ argument in the denominator. This measure contains less information, since it is only concerned with the relation between two phonemes in one direction: the probability of the second phoneme given the first. The O/E ratio also includes the probability of the first phoneme given the second. Transitional probability is tested in infants as an available cue in many experimental studies (Saffran et al., 1996; Jusczyk and Luce, 1994).

The calculation in 2 results in two types of statistical learning constraints, corresponding to the two types of constraints in the Optimality Theory (OT) framework. The first type, a sort of Faithfulness constraint, is aimed at not changing the input. It has the format CONTIG-IO([PHONEME][PHONEME]), and militates against the insertion of a word boundary between the specified phonemes. The second type is a sort of Markedness constraint, which wants to change the input to make it comply better with the grammar. It has the format *[PHONEME][PHONEME], and raises a violation if the two phonemes occur together (without an intervening word boundary). For every phoneme pair in the training data, if the O/E ratio is 2 or higher, a Faithfulness constraint of the phoneme pair is added to the grammar. If the ratio is lower than 0.5, the phoneme pair is used in a Markedness constraint. A ratio that is between 0.5 and 2 leads to no constraints, for these phoneme pairs the O/E ratio is not conclusive enough to directly create a statistical learning constraint. In order to let the StaGe model be able to make a decision on these pairs, the generalization constraints are used.

The generalization constraints are made by combining constraints from the same markedness or faithfulness category that are minimally different on the account of the features of their phonemes. 'Minimally different' is defined as differing in value for only one feature, all the rest being equal: CONTIG-IO([P][L]) and CONTIG-IO([B][L]) are minimally different, with only the value of the [voice] feature of the first phoneme differing. These constraints will be generalized to the constraint: CONTIG-IO([PB][L]).

The ranking of the constraints, defined by their weight, differs for the statistical learning and the generalization constraints. The former receive a weight based on the expected probability of the pair, equal to the calculation of the denominator in the O/E ratio. For generalization constraints, the weights of all the constraints that are combined into one are added, and then divided by the total number of phoneme pairs captured by the constraint. In the case of the CONTIG-IO([PB][L]) constraint, the weights of both contiguity constraints (which are their expected probabilities) are summed and then divided by two. In some cases, combining two constraints based on only one differing value leads to the constraint containing more phoneme pairs than the original constraints, if the generalized constraint contains bigger natural classes. Then, the weight is determined by dividing the summed weights by a larger number than the elements in the numerator. This leads to a lower ranking, because the constraint is not directly supported in all its statements. In such a situation, it is possible that the model becomes able to state something about a phoneme pair which did not receive an O/E ratio high or low enough to create a statistical learning constraint.

Four simulations with the StaGe model are presented in Adriaans and Kager (2010), two of which are of interest to us. The first is the default model, trained on a set of data from the corpus of spoken Dutch (CGN, Oostdijk 1999), the second is the inclusion of context information surrounding each biphone. For every biphone *xy*, the O/E ratio of its neighbors *wx* and *yz* are also calculated. The biphone with the lowest O/E score receives a word boundary. Adriaans and Kager (2010) presents the following results:

Model	Corpus	BH	BFA	Bd'
StaGe biphones	CGN	0.4454	0.1324	0.9785
StaGe biphones+context	CGN	0.4135	0.0913	1.1142

Table 4: Results Adriaans and Kager (2010)

3.4 Word learning

The models presented in this section are of the most interest to the topic of this thesis. The models combine word segmentation with word learning, in order to create a simulation of the infant developmental procedure. The discussion of the models in this section includes the learning process in the models for the

cases where there is something interesting to say about it. In many cases, the lexicon is simply a storage for all the parts of the segmented sentences and the frequency with which they have been segmented from the input, limiting the 'learning process' to memory storage.

This section presents models that combine sublexical segmentation strategies and lexical word recognition from a growing lexicon. The lexicon is empty at the beginning of training, cues in the input stream lead to segmentation. The parts resulting from the segmentation are then added to the lexicon, which becomes more involved in the segmentation process the more it contains. The process of 'adding parts to the lexicon' is called lexical learning. This section discusses several models that build a lexicon during the segmentation process, and increasingly use the recognition of forms from the lexicon for the segmentation process. During development, infant word learning necessarily takes place after segmentation and before recognition. Forms need to be segmented from continuous speech in order to be learned, and they need to be learned before they can be recognized in speech.

A word that is added to the lexicon subsequently takes part in the segmentation process. In some cases, this can cause a series of bad decisions leading to bad results. One problematic situation would be when the word /dɒg/ is in the lexicon and the word /dɒgz/ is encountered. If the model places a word boundary after /dɒg/, adding /z/ to the lexicon as a new entry, this can lead to a snowball effect: /ɪz/ becomes /ɪ/ and /z/, /sɪt/ becomes /s/, /ɪ/, /t/. Models that build and use a lexicon during the segmentation process often insert a restriction on word learning, in order to prevent this snowball effect. Two types of restrictions are presented: a phonological and a frequency related one. The first type is a restriction on the form of the part that is added to the lexicon. The form needs to follow certain phonological rules in order to be learned. A frequency constraint states that a form needs to be encountered a set number of times before it is added to the lexicon.

3.4.1 Gambell and Yang (2005)

A computational model that combines bottom-up word segmentation strategies with top-down word recognition is Gambell and Yang (2005). The model uses a subtraction process, in which a known word is subtracted from the to-be-segmented string. It starts from a string of syllables, rather than a string of segments. The authors state that infants perceive speech as holistic syllables rather than perceiving distinct phonemes. Neither hypothesis is incontrovertibly supported by the experimental literature. What is clear, is that providing the segmentation model with syllable boundaries significantly simplifies the segmentation task. Word boundaries always co-occur with syllable boundaries; providing them limits the error range.

Gambell and Yang (2005) have looked at language-independent strategies of segmentation. The experimentally heavily tested strategy of using transitional probabilities (Saffran et al., 1996) – the least

probable syllable pair of a set of pairs is the location of a word boundary – is compared to a combined segmentation strategy. The combined strategy extends the transitional probability approach with a segmentation strategy based on stress. The authors propose the Unique Stress Constraint (USC), the assumption that infants have innate knowledge that a word contains at most one main stress. This constraint states that a word boundary needs to be placed between two strong syllables (syllables carrying primary stress). When multiple weak syllables occur between two strong ones, and there are multiple places to place the boundary, the lowest transitional probability is chosen as the location for the word boundary.

Morgan and Saffran (1995) have attested that 6 to 9 months old infants are sensitive to the native stress patterns and thus able to distinguish strong and weak syllables. Though actual use of the stress pattern in a segmentation task is dispreferred over the use of a simple statistical cue in 7 months old infants (Thiessen and Saffran, 2003). The presence of the 'innate knowledge' that words can have maximally one main stress is an assumption that is hard to verify. Given the many tone languages in the world, the concept of innate knowledge on the stress system of a language would seem strange. In any case, the USC simplifies the segmentation task significantly, providing the model with information that is not directly present in the input but highly relevant for the segmentation process.

The results from the transitional probability strategy are significantly worse than those of the combined strategies. This is not too surprising, given that a model that knows that it needs to put a boundary around every strong syllable is similar to a model pre-specified with the average length of words. Furthermore, function words are represented in the corpus as strong syllables. The many monosyllabic (function) words in the English language are all pre-specified as strong, directly informing the model of word boundary locations.

Another combination with the USC is made by extending the model with a subtraction mechanism. This mechanism is activated only after the USC fails to place a boundary with absolute certainty, so in cases of multiple weak syllables occurring between two strong ones. When both the first strong plus subsequent weak syllable, and the final strong plus weak syllable are recognized as (part of) a word, the intervening syllables are added to the lexicon as a new word. Whether it is safe to say that two syllables which could form part of a familiar, longer, word are actually a signifier of this word, is not tested by the authors. In the case that neither or only one of the two syllable pairs is recognized as familiar, two different strategies are tested, one in which a boundary is placed randomly (*random*) and one in which no boundary is placed (*ignoring*). In both strategies, the resulting strings are not added to the lexicon as newly segmented words. The results of all four models as presented in the article are given in table 5. Keep in mind that the task performed by these models is simpler than that for models looking at phoneme sequences.

Model	Corpus	WP	WR	WF
TP	CBr	41.6%	23.3%	29.9%
TP + USC	CBr	73.5%	71.2%	72.3%
USC + subtraction (*random*)	CBr	85.9%	89.9%	87.9%
USC + subtraction (*ignoring*)	CBr	95.9%	93.4%	94.6%

Table 5: Results Gambell and Yang (2005)

3.4.2 Lignos (2012)

Another model that uses the subtraction strategy to combine bottom-up with a top-down segmentation strategy is proposed by Lignos (2012). Similarly to the Gambell and Yang (2005) model, the assumption is made for the Lignos model that syllable boundaries are salient enough in the input. The model is focused on the algorithmic level of Marr (1982), on *how* infants segment speech rather than only *what* (which cues) they use for the task. It is an incremental and online algorithm, which looks at one input sentence at a time. When it has segmented the sentence, only the lexicon is available for processing the next sentence. The algorithm is a simple model using the subtraction strategy. For every syllable, the lexicon is checked for matches. The match with the highest frequency 'wins' and subtracts the matching part from the sentence that is being segmented. The parts of the sentence that cannot be matched are added to the lexicon with a frequency of 1.

This baseline model is extended twice. The first extension is with a *Trust* feature: only segmented parts that touch an utterance boundary and thus have one of the two boundaries given, are added to the lexicon. The *Trust* feature is an implementation of the experimental results found by Seidl and Johnson (2006) that infants are more sensitive to words occurring at utterance boundaries. The second extension is that when a syllable is found in the lexicon, not only the entry with the highest frequency is subtracted, but a second segmentation is created using the second-highest frequency. For both segmentations, a calculation is made by multiplying the probabilities of each part. The segmentation with the total highest probability wins. Table 6 presents the results from Lignos (2012).

Model	Corpus	WP	WR	WF	BH	BFA	BA'
Subtraction	CBr	74.6%	85.4%	84.9%	0.992	0.960	0.795
+Trust	CBr	81.7%	86.6%	84.1%	0.960	0.469	0.860
+Multiple Hypotheses	CBr	83.2%	86.7%	84.9%	0.953	0.401	0.875

Table 6: Results Lignos (2012)

Both the Gambell and Yang and the Lignos model have restrictions on what is added to the lexicon. The fact that the models are presented with syllabified input ensures that only syllabic elements are learned. In other models presented in this section, there is a separate requirement on segmented parts

that they need to be syllabic before they can enter the lexicon. Only words with one primary stress are 'learned' in the Gambell and Yang (2005) model, or if they don't have primary stress a word needs to occur between two known words in order to make it into the lexicon. For the model proposed by Lignos (2012), a word needs to have one certified boundary before being learned and only parts that are encountered frequently enough are subtracted in later environments. These strong pre-specifications and requirements lead to good results for the models.

3.4.3 Blanchard, Heinz, and Golinkoff (2010)

The requirement that parts segmented from the speech stream are syllabic in order for them to be added to the lexicon, is used in the model proposed by Blanchard et al. (2010). The authors present a simple algorithm called PHOCUS – PHOnotactic CUe Segmenter. The model combines a language-specific and a language-universal phonotactic cue with the recognition of familiar words in the speech stream. The language-universal phonotactic cue is the requirement that every word has a syllabic element, the language-specific phonotactic cue is the distinction between phoneme combinations that are or are not grammatical in the language of the data the model is trained on.

Every segmented part in every segmentation of an input sentence is given a score. The segmentation with the highest score wins and its parts are added to the lexicon, either as an update of the form's frequency count or as a novel form with a frequency of 1. The score for a segmentation is the product of the probabilities of all of its parts. The probability of a familiar word is the frequency with which it has been encountered, divided by the total number of tokens in the lexicon. The probability of a novel word is based on its phonotactics. If the part contains no syllabic element, its probability is zero. The language-specific probability of the novel part is calculated from the probabilities of the n-grams in the part. These probabilities are determined from the forms in the lexicon, not those in the input corpus. The forms in the lexicon are delimited by word boundaries, the probabilities of a (pair of) phoneme(s) following or preceding a word boundary are learned by the model as well.

Jusczyk and Luce (1994) have attested infants' sensitivity to the likelihood of phonotactic patterns; and Mattys et al. (1999) and Mattys and Jusczyk (2000) have found that infants use this cues in segmentation tasks as well. The PHOCUS model is a selection model that creates all possible segmentations of a sentence and from that selects the one with the highest probability, given the lexicon and n-grams probabilities. This procedure is not necessarily plausible as the actual method applied by the infant, since the number of possible segmentations grows exponentially with the number of phonemes in the sentence (number of segmentations = 2^{n-1} with n = number of phonemes). However, in this format the model considers every possible segmentation and does not rule out any possible segmentations a priori.

Three models were compared in the Blanchard et al. (2010) article, in which three different n-grams

are used. The first model only looks at single phonemes, the probability of a novel word is the product of the probability of each of its phonemes. The second model is based on phoneme pairs, the third model uses phoneme triplets and multiplies the probabilities of each sequence of three phonemes in a part to calculate the probability of a novel form. The separate effect of the language-universal phonotactic constraint was tested by the authors: it is particularly effective against over-segmentation since consonants cannot be cut off without an accompanying vowel. The results of the three models with different n-grams as presented in the article, all including the syllabic requirement on the segmentations, are presented in table 7. The model discussed here forms the baseline of the model presented in chapter 4.

Model	Corpus	WP	WR	WF	BP	BR	BF	LP	LR	LF
PHOCUS-1s	CBR2	76.8%	69.7%	73.1%	91.0%	79.0%	84.6%	45.6%	64.1%	53.3%
PHOCUS-2s	CBR2	75.2%	64.2%	69.3%	93.7%	74.5%	83.0%	43.0%	63.7%	51.4%
PHOCUS-3s	CBR2	77.7%	74.0%	75.8%	89.7%	83.8%	86.5%	47.3%	64.0%	54.4%

Table 7: Results Blanchard, Heinz, and Golinkoff (2010)

3.4.4 Apoussidou and Kager (2013)

A study that focused on the effects of top-down and bottom-up segmentation strategies and a combination of the two is Apoussidou and Kager (2013). Apoussidou and Kager present a top-down word recognition model which is compared to and combined with the previously presented bottom-up segmentation model StaGe (Adriaans and Kager, 2010). The top-down mechanism uses forms in a protolexicon, which contains forms but no syntactic or semantic information. The forms that are added to and taken from the lexicon are called 'protowords', leading to the name of the model 'Use Protowords'(UP). The combination of bottom-up and top-down strategies leads to the best performance, both in creating the lexicon and in locating word boundaries. The model performs best when extended with a possible word filter. This filter is somewhat phonological: it states that single-segment words are not allowed. The result is similar to the requirement of a syllabic element in a word, though less phonologically grounded.

The top-down mechanism proposed by Apoussidou and Kager (2013) inserts per input sentence one boundary at all possible locations. The differently segmented sentences form the segmentation candidates that are evaluated using the protolexicon. Of all the candidates, including the faithful candidate which is the unsegmented sentence, the frequencies of the parts that make up the segmentation are looked up in the lexicon. The sum of the frequencies forms the score of the word. If a part does not have an entry in the lexicon, its frequency is 0. The winning segmentation will update the frequencies in the lexicon, the frequency of known words will be augmented by 1 and a novel form in the segmentation will be added to the lexicon with frequency 1. In situations where the faithful candidate has the highest score, or all scores of all segmentations are 0 (as is the case at the start of the model), the unsegmented sentence will

be added to the lexicon. The addition of only one boundary per sentence makes the model conservative, preventing snowball-effect errors.

All the segmentation parts of the sentences are added to the lexicon, and many candidates are generated simultaneously for the segmentation of a new sentence. The algorithmic level of this model (Marr, 1982) requires a lot of memory and computational power. Pearl et al. (2010) have investigated the influence of limitations on the memory of their Bayesian segmentation model in order to test a more realistic representation of a human learner. Besides the high demands on memory resources required for this model, the framework used as a selection mechanism for the segmentation process is a complex system. A strict algorithmic-level-implementation of this framework would require a lot of assumptions, such as the generation and evaluation of every possible candidate for every new sentence that is processed.

The top-down segmentation model is fitted within the Optimality Theoretic framework (OT, Prince and Smolenksy 1993). However, in order to fit the model a strange interpretation of the framework is required. The constraints are formed by the words in the lexicon, receiving a ranking score equal to their frequency. The concept of different types of constraints directly militating against each other – Markedness and Faithfulness in standard OT – is not present in the set of constraints created from the lexical forms. Furthermore, a segmentation candidate that matches with a lexical form constraint is satisfying this constraint rather than violating it. In standard OT, the concept of constraint satisfaction is not tenable (Kimper, 2010). Finally, the proposed model contains the non-OT concept of weighted constraints and the commensurate concept of constraint ganging. Two lower-ranked constraints can outweigh a higher ranked constraint if their combined frequencies are higher than the higher-ranked constraint's frequency. The ganging of weighted constraints is a property of Harmonic Grammar (Legendre et al., 1990), not Optimality Theory. The OT framework seems to be chosen only in its functionality as a selection process, for which the Harmonic Grammar framework seems to have been a more suitable choice.

The top-down segmentation mechanism alone leads to a lexicon with almost half of the twenty most frequent words consisting of only one segment. The possible-word filter successfully counters this. The best performance is produced when the model including the possible-word filter is recursively applied, or when it is combined with the phonotactic segmentation model StaGe (Adriaans and Kager, 2010). A recursive application of the model feeds the outputs of the first cycle, including the placed boundaries, as input to the second cycle, which adds at most one other boundary. In the combined model, 90% of the text is segmented by StaGe, the remaining 10% is given to UP directly. The by StaGe segmented output is concatenated into a lexicon with frequencies, which is presented to UP to use for the recognition strategy. Including the possible-word filter is a necessary improvement in the recursive model. For the combined model, it mostly improves the word boundary placement, the lexicon does not improve significantly with the inclusion of the filter.

All results in table 8 are produced by the model including the possible-word filter. The recursive model has processed five cycles. Apoussidou and Kager (2013) tested the model on two corpora: the adult and variation-including CGN (Oostdijk, 1999) and the child-directed speech of the van de Weijer corpus (Van de Weijer, 1998). All four combinations of these corpora were tested, with the model trained and tested on the same corpus or trained on the one and tested on the other. The results with the van de Weijer corpus led to the best results, and the results with the CGN are included to enable direct comparison with the performance of the StaGe model (see table 4, Adriaans and Kager 2010). All results are taken from the article.

Model	Corpus	BH	BFA	Bd'
UP Recursive	VW	0.572	0.119	1.362
UP + StaGe	VW	0.625	0.155	1.335
UP Recursive	CGN	0.186	0.093	0.429
UP + StaGe	CGN	0.589	0.180	1.163

Table 8: Results Apoussidou and Kager (2013)

3.4.5 Daland (2009); Daland and Pierrehumbert (2010)

A different type of word learning restriction than the previously discussed phonological ones, is a frequency constraint. A form is added to the lexicon if it is encountered a set number of times. Multiple occurrences of a segmented part provide the necessary certainty to the model. This concept has been examined in Daland (2009) and Daland and Pierrehumbert (2010). A bottom-up phonotactic segmentation strategy is presented, which looks at biphones, comparable to Adriaans and Kager (2010) and the bigram version of Blanchard et al. (2010). The Diphone-Based Segmentation (DiBS) uses Bayes' theorem, similarly to other segmentation models presented in the next section. Two versions of the DiBS model are presented: Phrasal-DiBS acquires phonotactic information from utterance boundaries, Lexical-DiBS acquires the information from the lexicon. The latter model will be the main focus in this section. In order for a word to be added to the lexicon and let its diphones become available for future segmentation, a form needs to be encountered (segmented) in the corpus a set number of times.

The DiBS model requires a set of parameters which are assumed to be available to the infants. Some of these assumptions are common for segmentation models: phonemes are independent of each other across word boundaries; the infant can distinguish phonetic categories; it keeps track of the frequencies of learned words; and knows the probability distribution of diphones. The latter of these assumptions is tested and confirmed by Mattys and Jusczyk (2000). The assumption that infants know the distribution of phonemes at phrase edges is computationally tested by Brent and Cartwright (1996). In their model, the results were not worse when the phrase edge distributions were learned in the segmentation process versus when they were presented to the model a priori (see section 3.5).

The final assumption of the model is that it knows the probability of a word boundary. This assumption is the least likely to be available for infants. It can be estimated from average word length, but this information is only available after a certain amount of words has been segmented and learned. The authors state that even before a lexicon is available to generalize over, the average word length can be estimated from the knowledge that content words are minimally bi-moraic. How this knowledge would become available for infants is not discussed, it is assumed to be a 'cross-linguistic generalization' (Daland and Pierrehumbert, 2010, p129). Because the plausibility of this assumption is doubtful, its robustness was checked. The oversegmentation rate is particularly robust against variation in the probability value for word boundaries. The undersegmentation rate shows variation, although there is a reasonable range in which the variation stays minimal.

Phrasal-DiBS is an unsupervised model that approximates the distribution of diphones utterance medially – and the probability of finding a word boundary given a pair of phonemes – from the distribution of diphones at utterance boundaries. Utterance boundaries are either preceded or followed by a pause and often prosodically marked, making them a salient cue which is available to infants without requiring previously acquired knowledge (Seidl and Johnson, 2006). Lexical-DiBS is semi-supervised. The distribution of diphones at word boundaries is learned from actual diphones at word boundaries: those in the lexicon. The model starts with a small set of words, from which the diphone distribution is measured and whose acquisition is not modeled.

In his dissertation, Daland (2009) has experimented with a combination of the two DiBS models. First, the Phrasal-DiBS poses some boundaries. When the frequency thresholds are passed and words start entering the lexicon, Lexical-DiBS comes into play as well. The combined DiBS model is not discussed in Daland and Pierrehumbert (2010). The Mixture-DiBS model severely oversegments in its baseline version, requiring Daland to insert a word learning restriction: a word needs to possess a vowel in order for it to be added to the lexicon. The Mixture-DiBS model with the vowel constraint leads to a surprisingly high number of words in the lexicon. This is probably possible because the vowel restriction requires longer words, which leads to more possible word forms than the set of one or two phoneme forms resulting from over-segmentation.

Daland and Pierrehumbert (2010) have compared three values of a word learning restriction on Lexical-DiBS. Words had to be encountered either 10, 100 or 1000 times before they were added to the lexicon. 100 was the actual estimate of the researchers and the one that is presented in table 9, while the 10 and 1000 frequency thresholds were a lower and an upper bound. The variation in the three versions of the Lexical-DiBS model was not very large. The probability of finding an undersegmentation error per word was around 0.1 for every version. The variation in oversegmentation errors was larger, ranging between 0.3 and 0.5. The frequency threshold of 100 encounters performed best overall. The performance

as reported in Daland and Pierrehumbert (2010) of the Lexical-DiBS with a frequency threshold of 100, of the Phrasal-DiBS model and of a baseline model which was given the actual diphone distributions in the corpus in order to show the ceiling performance of the DiBS model, are presented in table 9.

Model	Corpus	WP	WR	WF	BP	BR	BF	LP	LR	LF
Baseline-DiBS	Ch	73.7%	69.6%	71.6%	88.3%	82.1%	85.1%	14.6%	53.6%	23.0%
Phrasal-DiBS	Ch	53.4%	35.2%	42.5%	87.4%	48.9%	62.7%	5.6%	50.8%	10.1%
Lexical-DiBS	Ch	44.8%	44.8%	33.3%	89.7%	39.0%	53.1%	4.5%	47.3%	8.2%

Table 9: Results Daland and Pierrehumbert (2010)

In the model presented in chapter 4, a similar frequency restriction on word learning is presented. Two types of frequency thresholds are compared. The first is equal to the one built into Lexical-DiBS. The second type of frequency threshold contains a time element: a form needs to be encountered a set number of times within a set amount of time before it is added to the lexicon.

3.5 Bayesian models

The segmentation models presented in this section are all Bayesian models, a type of model that has inspired many following studies. This type of model has an interesting and active role for the lexicon in the segmentation process. Brent and colleagues created a series of influential segmentation models (Brent and Cartwright, 1996; Brent, 1997, 1999), the last of which was extended into other models (Blanchard and Heinz, 2008; Goldwater et al., 2009). The first model in the series used a type of word learning restriction, a language-universal phonotactic constraint requiring every word to contain a syllabic element. Brent and Cartwright (1996) were the first to propose this restriction. Later versions were built within a generative model-based Bayesian framework. A generative model probabilistically generates the observed data, the input corpus, given a set of hypotheses which are different segmentations of the corpus and the resulting lexica. The Bayesian framework determines which hypothesis wins. It performs better in this task than maximum-likelihood estimations, which are not well-equipped to deal with varying complexity levels of the hypotheses (Goldwater et al., 2009). All Bayesian models presented in this section are in some way or another focused on finding a segmentation of a corpus which leads to minimal 'cost' to the lexicon, and thus the least complexity of the segmentation or the lexicon.

3.5.1 Brent and Cartwright (1996); Brent (1997, 1999); Blanchard and Heinz (2008)

In the Brent and Cartwright (1996) model, limiting the lexical cost is implemented as Minimal Description Length (MDL, Rissanen, 1989). The MDL method defines a function that minimizes the description of the data, or minimizes the rules and memory storage necessary to produce the input data. Such a

function is called a distributional regularity function (DR function). The description length of the data in Brent and Cartwright (1996) is measured in the following way:

1. All unique items in the segmentation are added to the lexicon.
2. All items in the lexicon receive an index.
3. All words in the segmented sentence are replaced by their index.
4. All phonemes and index numbers in the lexicon are counted.
5. The sum is added to the sum of the number of words (indices) in the segmented sentence.
6. The total sum is the description length of the evaluated segmentation of the sentence.

For a segmentation (*s*) consisting of words (*w*) whose length is counted in phonemes, the function stepwise described above looks as in (3), in which an argument between | | denotes the number of items in the argument. In the actual computational experiment, an extra argument is added that is left out here. The extra argument covers the entropy of the relative frequencies of the words, in which the relative frequency of a word is its probability: how many times it has appeared divided by the total number of encountered tokens. The entropy of this measure is the distribution of probabilities over the words in a segmentation, this number is high when the words have comparable relative frequencies. When there are big differences between the relative frequencies of forms, the entropy goes down, also lowering the MDL score. The DR function looks as follows:

(3)

$$f(s) = |tokens(s)| + |types(s)| + \sum_{w \in types(s)} length(w)$$

Together with the MDL score, there are two phonotactic constraints that determine which segmentation is the optimal one: a) every word must have a vowel; b) only a limited set of consonant clusters is allowed to occur word initially or word finally. The second of these constraints can be learned from the phoneme sequences at utterance boundaries, as is tested by the authors. The model even performs slightly worse using a pre-specified list of word-initial and word-final consonant clusters than when this list is learned from utterance boundaries (Brent and Cartwright, 1996, p.114).

The algorithm implementing the MDL works in batch mode: it first reads in all the input sentences before it starts segmenting them. This is of course not how an infant goes about the problem. In a follow-up article, Brent (1997) proposes an incremental implementation of the MDL model. This processes each sentence as it comes in before moving on to the next. Furthermore, the extension includes a focus on the online processing of partial utterances, how the model functions when only part of the utterance is received. This sheds light on the process of lexical access, how forms are retrieved from the lexicon.

The new model, INCremental DR Optimization, or INCDROP, militates specifically against novel forms. This is understandable given the arguments in the DR function: a novel form increases the number of types in the lexicon and the sum of the forms of the types. Furthermore, a novel form has not been seen before, its relative frequency is calculated with a special formula, and is always lower than that of any previously encountered word. Since the product of the relative frequencies is to be maximized, any distinctively different relative frequency is unwanted. In order to minimize the length and number of novel words, the model adds not previously encountered parts in the input to the lexicon in their unsegmented form. The product of the relative frequencies will be higher if only one low relative frequency is part of the equation.

The incremental implementation enables the model to process each sentence on its own. Besides the incremental adaptation of the MDL model, its online extension enables the model to process each new phoneme as soon as the model receives it. Every new phoneme causes a re-calculation of all possible segmentations of the previously perceived data. Simultaneously, the rest of the word that is currently being perceived is predicted from all available forms in the lexicon. This double task makes the model computationally complex. The INCDROP described in Brent (1997) is only a theoretical model with no computational implementation. Consequently, there are also no simulations or results available.

An actual implementation (and further extension) of the INCDROP model is presented in Brent (1999). This (newest) model is called Model Based Dynamic Programming (MBDP-1). It uses an abstract, non-linguistic architecture as basis, to which several linguistic modules are fitted. The abstract model generates from a pre-defined set of phonemes a text with word boundaries in four steps, followed by a fifth step that removes the word boundaries. In the segmented text at step four, the probability of the text can be calculated by recursively calculating the relative probability of each word based on the words that have been encountered in the text before the current word. The text that receives the highest probability through the recursive relative probability calculation forms the best segmentation. The segmentation that receives the highest probability is the best one.

This recursive relative probability calculation has two forms: one for if the word of which the probability is being calculated is encountered in the text up to that word and one for if the word is novel. The model proposed in Brent (1999) is abstract, the authors propose some ideas on how to fill in the details of making this model function more precisely as a segmentation model and they invite researchers to implement them. The calculation of the relative probability is only based on the parameters used in the generation of the text in the four steps, and not on extra, more linguistically supported arguments. In order to facilitate model comparison, performance of the MBDP-1 model is presented together with performance of the Goldwater et al. (2009) model, discussed in the next section. Goldwater et al. (2009) directly compared their model with Brent's MBDP-1 by letting the models segment the same text.

Two researchers that have responded to the request of Brent to extend his model are Blanchard and Heinz (2008). They extended the function calculating the relative probability of novel forms with a more sophisticated phonotactic model and named the extended model MBDP-Phon. In the Brent (1999) model, the phoneme combination in the novel form is evaluated with a probability score based on the probabilities of the phonemes in the form, but not on the order of those phonemes. The extension proposed by Blanchard and Heinz is to calculate the maximum likelihood estimate (MLE) of the form, instead of the product of the phonemes' probabilities. This is done by calculating the likelihood of each n-gram in the novel form: how many times that n-gram has occurred divided by the total number of n-grams of that length in the lexicon. The extension is based on the idea that it is not so much phoneme probabilities as much as phoneme combinations that make up a well-formed word.

The results of all the Bayesian segmentation models are presented in table 10.

3.5.2 Goldwater, Griffiths, and Johnson (2009)

Another extension of the MBDP-1 model (Brent, 1999) is proposed by Goldwater, Griffiths, and Johnson (2009). The ten years younger model is very similar to the MBDP-1, but makes use of faster and more advanced statistics that have become available. Especially the search procedure for determining the winning segmentation out of all the options is improved, using a method called Gibbs sampling. This improvement enables the model to work without interference and produce the optimal results according to its own performance. The two models proposed by Goldwater et al. are named DP and HDP, which is short for Dirichlet process and hierarchical Dirichlet process. This is a specific kind of statistical model that functions as a non-parametric prior in Bayesian models used to cluster data. The general task of the model is to find the segmentation of a corpus that has the highest likelihood given the lexicon. Similar to the Minimal Description Length idea of Brent and Cartwright (1996), the Dirichlet process results in an unequal distribution over word frequencies, with few items occurring often and many items occurring only a few times.

The focus of Goldwater et al. was on the assumptions made in the segmentation process. This led to a specifically computational level algorithm, contrary to the algorithmic model presented by Lignos (2012). The authors were interested in the usefulness of assuming that words in speech are predictive rather than randomly occurring independent units. Many segmentation models focus on the relations between sub-lexical units, such as syllables (Lignos, 2012), phonemes (Blanchard et al., 2010) or features (Adriaans and Kager, 2010). The relation between lexical units is often ignored.

The first model presented by Goldwater et al. (DP) does not take account of the relations across words, their second model (HDP) does. The performance of the DP model leads to a high number of undersegmentation, especially in high frequency words and collocations of function words such as 'that's

a' or 'look at the'. The DP model selects the next word based on the segmentation of the corpus thus far. The probability of the next word, if it is known, is equal to its token probability. In the HDP model, words are seen as part of a bigram. In determining the probability of a familiar word occurring, the occurrence of the bigram given the word preceding the upcoming form is calculated.

The performance of the HDP model is better than that of the DP model, see table 10. The bigram model has also performed better than the unigram MBDP-1 model. The authors present precision, recall and F-scores for word tokens, word boundaries (utterance boundaries are excluded because they are known beforehand and would unfairly influence the results) and word types in the lexicon. These results are repeated in table 10.

The concept of dependence across word forms is an interesting assumption on word learning. Taking the direct context into account enables a segmenter to more successfully segment speech. Especially speech which is difficult to parse using other cues, such as function words which are often unstressed and reduced. The predictivity of words given their predecessor is a useful assumption in learning words. It is a relevant notion to keep in mind when designing a computational model for word learning.

Model	Corpus	WP	WR	WF	BP	BR	BF	LP	LR	LF
MBDP-Phon	CBR2	73.21%	76.05%	-	-	-	-	54.34%	-	-
MBDP-1	CBR1	67.0%	69.4%	68.2%	80.3%	84.3%	82.3%	53.6%	51.3%	52.4%
DP	CBR1	61.9%	47.6%	53.8%	92.4%	62.2%	74.3%	57.0%	57.5%	57.2%
HDP	CBR1	75.2%	69.6%	68.3%	90.3%	80.0%	85.2%	63.5%	63.5%	59.1%

Table 10: Results Bayesian segmentation models

3.6 Preliminary summary

This chapter has presented several computational segmentation models. All models are able to locate word boundaries in a continuous string of input sounds. Three types of segmentation models were distinguished based on the role of the lexicon in the models. The models in the section on word segmentation contain no lexicon, they are focused on locating word boundaries. The section on word learning contains models starting with an (almost) empty lexicon which was extended and applied during the segmentation process. The section which discusses adult speech processing models and word recognition covers segmentation models using a pre-specified lexicon such as adults have available in the speech recognition process.

All types of models represent an important part of the segmentation process. However, the models which do not involve the lexicon at all in the process leave out an important interaction, one that is a focus in this thesis. Already at six months of age, infants apply familiar forms to recognize the start of a novel form (Bortfeld et al., 2005). Leaving out the role of the lexicon entirely gives an incomplete perspective

on the segmentation task. The word recognition models are focused on the adult segmentation process in which a lexicon is available, contrary to the infant segmentation process. In the next chapter, a new model of segmentation is proposed. This model starts with an empty lexicon which plays an active role in the segmentation process as soon as it contains information on word forms similar to the models in the section on word learning. Contrary to Lexical-DiBS in Daland and Pierrehumbert (2010), the lexicon starts out completely empty to get the most realistic representation of the task infants face.

Besides a contrast in the role of the lexicon, the models differ in their architecture and in the assumptions on the available input and segmentation cues. The assumptions on the available cues are all in some degree supported by the experimental literature. Every cue applied in the models has been attested to be available for segmentation in infants. This does not mean that every model is equally probable based on its input assumptions. Both Gambell and Yang (2005) and Lignos (2012) have oversimplified the segmentation task for their models. Not only is the number of possible word boundary locations lowered because the syllable boundaries are provided, but the presence of stress information and the knowledge that a word can have at most one primary stress makes the segmentation task simpler than what infants actually have to do. The fact that infants can distinguish words which differ only in a few phonetic features of the onset phoneme (Jusczyk and Aslin, 1995) is used as argumentation for the choice to give the model presented in chapter 4 segmental rather than syllabic input. The finding that infants prefer syllabic over non-syllabic stimuli (Bertoncini and Mehler, 1981), which is used as argument for the syllabic input in Gambell and Yang (2005) and Lignos (2012), is seen as argumentation for the vowel-filter in the model presented in chapter 4.

Several frameworks have been used for the segmentation models. The main ones are connectionist (McClelland and Elman, 1986; Christiansen et al., 1998), Optimality Theory (Adriaans and Kager, 2010; Apoussidou and Kager, 2013) and Bayesian (Brent, 1999; Norris and McQueen, 2008; Blanchard and Heinz, 2008; Goldwater et al., 2009). The connectionist framework is built as a representation of the human brain. Nodes are connected with each other with different strengths, similar to the neurons in our brain. However, in order to build this structure and to translate its output into measures that are comparable to the results of experimental studies, assumptions need to be made which cannot be based on the structure of the human brain. This makes the framework less representative than the Bayesian one.

A Bayesian model is based on a statistical calculation system, parts of which have been attested in infants (Jusczyk and Aslin, 1995; Saffran et al., 1996). How this system is implemented in the human brain is not a principle factor and does not require as many assumptions as in the connectionist framework. But given its straightforward statistic framework, confined to requiring only the ability to perform statistical calculations, transforming to a neurologically valid implementation does not seem implausible. Furthermore, the statistical aspect is domain-general and not language-specific. It does not require

infants to have innate knowledge of (a specific) language. The Bayesian framework makes a limited set of assumptions on the computational abilities of the infant, assumptions which can both be tested on the computational and algorithmic level (Marr, 1982).

The Bayesian framework is a computational framework, using statistics and calculations to derive optimal outcomes. The Optimality Theory framework (Prince and Smolenksy, 1993) is a grammatical framework. The output of an OT analysis is a grammar, for a Bayesian analysis it is a function. Both frameworks are applicable to many languages, the Bayesian because statistics are language-independent and the OT framework is designed to be language-independent. The latter framework has already been used to analyze a wide range of languages. For a computational simulation of the segmentation and word learning process, which is the aim of this article, the Bayesian framework is more straightforward. Because it is in essence already computational, less translational assumptions need to be made. For example, the OT framework has constraints as its main building blocks. In a computational version of the OT framework, these constraints need to be translated to a usable function first in order to be able to use the framework. These extra translations are not required in the Bayesian framework. Since the benefits of the OT framework – such as the output of a grammar – are not relevant for the aim of this thesis, the segmentation model presented in chapter 4 is built in the Bayesian framework.

A final distinction between the computational segmentation models is whether they function in batch mode or incrementally. A batch model (Brent and Cartwright, 1996; Goldwater et al., 2009) receives the entire set of training data at once and recursively analyzes the data until it has subtracted enough information from it. This is not an accurate representation of the infants' procedure. An infant receives each sentence as it is uttered, and it is not a plausible assumption that every perceived sentence is kept active in memory for continuing analysis. Incremental segmentation models (Blanchard et al., 2010; Lignos, 2012) are constructed to analyze every sentence as it comes in. When the analysis is finished, the next sentence is processed. Between sentences, only abstract information is transferred such as the lexicon and probability distributions. Incremental models are more plausible and a better representation of the task infants have to perform.

The model presented in chapter 4 is modeled after the PHOCUS model (Blanchard et al., 2010). This is an incremental model in the Bayesian framework which receives segmental information without indicated syllable boundaries as input. The cues the model uses are transitional probabilities between phoneme sequences of different lengths and recognition of words from the lexicon. The assumptions the model makes are supported by experimental research and its structure is not complex. The model is proposed with an integrated set of factors from the infant word learning process, in order to make a representative simulation of the infant word acquisition process. The models discussed in this section were experimentally well supported on the word segmentation element, since only segmentation cues

that have been experimentally attested in infants have been included. However, the word learning portion of the models is often simplified: forms are added to the lexicon as soon as they have been segmented for the first time from the input. This is a simplified representation of the lexical acquisition process, which ignores factors that are relevant to the actual infant word learning process. The next section presents a segmentation and word learning model that has taken the first steps to a representative simulation of the lexical acquisition process, as well as of the word segmentation process.

4 The STROLL model

The previous chapter discussed segmentation models, a subset of which include a word learning component. Most of the models have no restrictions on word learning, some of them have a phonological filter such as requiring every segmented part to contain a syllabic element. Only the Lexical-DiBS model of Daland and Pierrehumbert (2010) contains an extra filter, one that is also present in the infant word learning process. The model contains a coarse-grained frequency filter, only adding forms to the lexicon after they had been encountered a set number of times. Three values with a large range were compared, but the values did not cause an equally dispersed effect in the output results. There is a deficiency of representative simulations of the infant word learning process in combination with the word segmentation process. Computational models that contain both processes simplify the word learning process, which does not contain factors that are relevant to infants in their word learning reality.

This chapter presents a segmentation model with restrictions on word learning, simulating the infant segmentation and lexical acquisition process. The model is called STROLL, Segmentation Through Restrictions On Lexical Learning. STROLL is designed in a way that the assumptions the model makes are experimentally plausible. All the assumptions on the input the model receives and the knowledge that is applied in the process have been experimentally attested to be available to infants. These are assumptions on the computational level of the model. The computational level is concerned with *what* the model does, the algorithmic level is focused on *how* the model functions (Marr, 1982). Ideally, a computational model makes experimentally supported assumptions on both these levels in simulating the segmentation process of an infant. In the current model, the algorithmic level is less plausible. This is necessary in order to be able to consider all segmentation options and avoid a priori selections. This will be discussed in more detail below. Section 4.1 discusses the segmentation cues used by the model, section 4.2 presents the structure of the model and how it decides on an optimal segmentation of a sentence. Section 4.3 focuses on the restrictions the model imposes on word learning. The final section of this chapter presents the methodology of the simulations run on STROLL. The results of these experiments are presented in chapter 5.

The STROLL model is similar to the PHOCUS model of Blanchard et al. (2010). The STROLL model contains the same segmentation cues and procedure, and many of the assumptions made by the PHOCUS model are present in the STROLL model as well. However, there are some significant distinctions between the models that will be discussed in the relevant sections.

4.1 Segmentation cues in STROLL

The STROLL model applies two segmentation cues simultaneously, each of which applies a different segmentation strategy. The first cue is part of a bottom-up segmentation process. The information used to segment the sentence is directly available in the utterance, and only the information in the utterance is determining where word boundaries will be placed. This cue can be applied early in the segmentation process because with every processed sentence, more information becomes available. In contrast, the top-down segmentation strategy of the other segmentation cue focuses on previously acquired information in the lexicon. This information needs to be acquired and stored before it can be applied in the segmentation process. The two cues discussed in this section are the same as those in the PHOCUS model. Their implementation in the two models differs, since in the PHOCUS model they are not implemented as a bottom-up and a top-down cue. The distinction in the implementation of the two cues between the STROLL and PHOCUS model is discussed in this section.

4.1.1 Bottom-up segmentation based on phonotactic probabilities

The bottom-up, signal-based segmentation strategy which is directly learned from and applied to the input is the phonotactic transitional probability cue. The probabilities of different phoneme combinations are acquired from the input, by calculating their conditional probability based on their frequency of occurrence. As discussed in section 2.1, the probability distribution of phoneme sequences denoted by their transitional probabilities are a useful source of information about the location of word boundaries in the speech input. The phonotactic grammar prescribes that certain phoneme combinations cannot occur within a word but only across word boundaries. This makes the probability of encountering that sequence distinctly lower. This distinction is noticeable by infants, and can be used in the segmentation process (Mattys et al., 1999; Mattys and Jusczyk, 2000).

The phoneme combinations are called n-grams. For every utterance that is perceived, the frequency counts of the phoneme combinations are updated with the ones that occur in the utterance. The n in n-grams denotes the number of phonemes per sequence. In the PHOCUS model, the n stands for 1, 2 or 3. The authors have compared performance of the model for when it was calculating probabilities of single phonemes, phoneme pairs or triplets. The STROLL model presents a comparison of bigrams and trigrams, but most of the results are outputs of the model calculating bigram probabilities. Similar to the StaGe model (Adriaans and Kager, 2010), STROLL uses bigrams to gain information about phonotactics. Contrary to that model, the probabilities of bigrams in the STROLL model are calculated by their transitional probability, not their Observed/Expected ratio (see section 3.3.2). The transitional probability of an n-gram is its conditional probability: the frequency of the phoneme sequence given the

first part of the sequence excluding the final segment. The O/E ratio takes into account both directions of a phoneme pair: the probability of the first phoneme given the second, and the probability of the second given the first phoneme. This measure contains more information than the transitional probability, which only looks at the probability of the second phoneme given the first. The STROLL model has used the transitional probabilities equal to the PHOCUS model after which it is designed. This cue is supported by many experimental studies, which attested the sensitivity of infants for transitional probabilities (Saffran et al., 1996; Mattys and Jusczyk, 2000; Johnson and Jusczyk, 2001).

The probability distribution of the n-grams can be calculated from three different sources of input: types, tokens or utterances. The PHOCUS model calculates the frequencies from the types in the lexicon (Blanchard et al., 2010). They refer for discussion of the choice of types over tokens to Venkataraman (2001). Both these two sources of information on n-gram frequencies take into account the word boundaries inserted by the model. The list of n-grams over which the probability distribution is calculated, includes phoneme sequences in which one of the phonemes is a word boundary. In the sequence /kæt/, the following bigrams occur: [#k],[kæ],[æt],[t#]. If the word /kæt/ has been segmented by the model, the location of the word boundaries (denoted by #) and their occurrence in phoneme combinations is a result created by the model. This result can be false, leading to skewed transitional probabilities for certain phoneme combinations.

There is a third information source besides types and tokens from which to calculate the frequencies of phoneme combinations – the input utterances – e.g. used in the StaGE model (Adriaans and Kager, 2010). The unsegmented utterances only contain the word boundaries which coincide with the utterance boundaries. The salience of utterance boundaries and their usefulness in the segmentation process has been demonstrated by Seidl and Johnson (2006), who showed that infants are significantly better at segmenting words from utterance initial or final positions than from an utterance medial position. Words falling at the beginning or ending of an utterance have one boundary given, making the word finding task only half as hard. Furthermore, Lignos (2012) has included a *Trust* feature, giving more probability to words which have one boundary falling at an utterance boundary. Inclusion of this feature improved the performance of Lignos' model. The STROLL model calculates the probability distribution of phoneme combinations from unsegmented input utterances.

The distinction between calculating n-gram probabilities from the utterances or from the lexicon is also made by the DiBS model (Daland and Pierrehumbert, 2010). Phrasal-DiBS acquires bigram probabilities from utterance edges, Lexical-DiBs from the forms in the lexicon. The latter model is semi-supervised, it receives a small set of word forms at the start in order to calculate bigram probabilities before adding forms to the lexicon. Although it is not an unrealistic assumption that infants learn a small set of words before really starting the segmentation process (Bortfeld et al., 2005), the Phrasal-DiBS

model is more plausible in that it requires less assumptions such as which and how many word forms are already available. Another assumption the Lexical-DiBS model makes in providing a small set of words, is that these forms are correctly segmented. As discussed, calculating the n-gram distribution over utterances including utterance boundaries provides more certainty than when calculated over types or tokens including the by the model inserted word boundaries. Using the unsegmented utterances for calculating the phoneme sequence probabilities in the STROLL model ensures that the segmentation cue is entirely bottom-up.

4.1.2 Top-down recognition of familiar words

The second segmentation cue is top-down: forms that are stored in the lexicon are used for the segmentation of new utterances. This lexical cue is derived from the knowledge the infant has available, rather than from the input signal. The usability for infants of familiar forms in the segmentation process has been shown by Bortfeld et al. (2005). A word following a familiar word was significantly better segmented and recognized in a later presentation than a word following an unfamiliar form. When a form does not occur at an utterance boundary, it is harder to segment for infants (Seidl and Johnson, 2006). If that form receives support from the lexicon, it can be segmented from the utterance with a higher level of certainty. The certainty level of segmenting familiar forms from the input is discussed in the next section. The word boundaries surrounding the form function as utterance boundaries giving certainty to the preceding and following forms, albeit with a smaller level of certainty than the actual utterance boundaries.

Several other models have used the recognition of familiar forms as a segmentation cue. Lignos (2012) presents a model which only uses familiar forms in the segmentation process in the baseline version. The model receives utterances in which syllable boundaries are indicated. Starting at the first syllable, the model looks up the forms that match it in the lexicon. The matching form with the highest frequency is segmented out of the utterance. The same process is applied to the remaining part of the utterance. The application of the familiar form segmentation cue in the STROLL model functions differently than in the Lignos (2012) model, which is discussed in the next section.

The use of lexical forms in the model presented in Brent and Cartwright (1996) differs from their application in PHOCUS and STROLL. The Minimal Description Length model aims specifically at limiting the size of the lexicon. A familiar form is always preferred over a novel form because the latter increases the size of the lexicon. The PHOCUS and STROLL models focus on the probability of the segmentation of the entire utterance. A segmentation which introduces a novel form can receive a higher probability than one that doesn't, if the n-grams are supporting the novel form and the familiar forms have a low frequency. Irrespective of the implementation of the familiar forms cue, all models have in common that they prefer high frequency forms. A form with a high frequency has a higher probability

of being segmented from a new utterance, further increasing the frequency of the form. This discrepancy between frequent and infrequent forms is a good representation of how words occur in natural speech (Zipf, 1949), and can be seen as a bias which infants apply in the segmentation process (Goldwater et al., 2009).

4.2 The structure of STROLL

This section discusses the structure of the STROLL model. The general architecture of the model will be presented first, followed by a stepwise presentation of the way the model processes a sentence and applies the segmentation cues. The final part of this section discusses the mathematics involved in the process. As discussed in section 3.6, besides the choice of segmentation cues as a distinguishing factor between segmentation models, the framework is also relevant together with whether the model is incremental or functions in batch mode. The segmentation cues were discussed in the previous section, this part focuses on the framework and how the model functions.

4.2.1 Model's architecture

The STROLL model works incrementally, similar to Blanchard et al. (2010) and Lignos (2012). It processes sentences one by one, updating the information that is necessary for the segmentation cues. After a sentence is processed the updated information is used for the next sentence, but the processed sentence is not stored in memory and will not be re-analyzed. This is contrary to the batch approach of Brent and Cartwright (1996) and Goldwater et al. (2009). An incremental model is a more realistic representation of the infant's segmentation procedure. An infant perceives a sentence, analyzes it and moves on to the next. A batch model analyzes the learning data several times before it moves on to a testing phase in which the model's performance is tested. The incremental model is learning and performing at the same time. It segments a sentence and uses the segmentation of that sentence to update the information needed for the segmentation cues: in the STROLL model's case the lexicon and the n-gram probability distribution. There is no clear distinction between a learning and testing phase in an incremental model.

As discussed in section 3.6, there are three major frameworks for segmentation models: connectionist, Bayesian and Optimality Theory. It had been concluded that the Bayesian framework requires less assumptions than the connectionist framework in predicting reaction times or other measures of language processing. The Bayesian framework is more computational in nature than Optimality Theory, consequently it requires less adaptation to build a computational model in a Bayesian than an OT framework. The STROLL model is built in the Bayesian framework.

The diagram in figure 2 presents an overview of the STROLL model. The chart is taken from Blanchard et al. (2010), the functioning of the PHOCUS model as outlined in figure 2 is the same as it is

for the STROLL model. The differences between the models are in the presence of the restrictions on word learning and where the n-grams distribution is calculated, which is not indicated in the chart. After the best segmentation is determined, the lexicon is updated before the next utterance is processed. In the PHOCUS model all new forms are added to the lexicon. In STROLL only the forms that pass the threshold are added. After the lexicon is updated, the n-gram distribution is updated as well. For the PHOCUS model this is done by calculating the n-gram distribution over the types in the updated lexicon. For STROLL, the n-grams occurring in the freshly processed utterance are added to the n-gram frequency distribution. The output of the model is the lexicon after being updated by the last segmentation and the winning segmentation of each processed utterance.

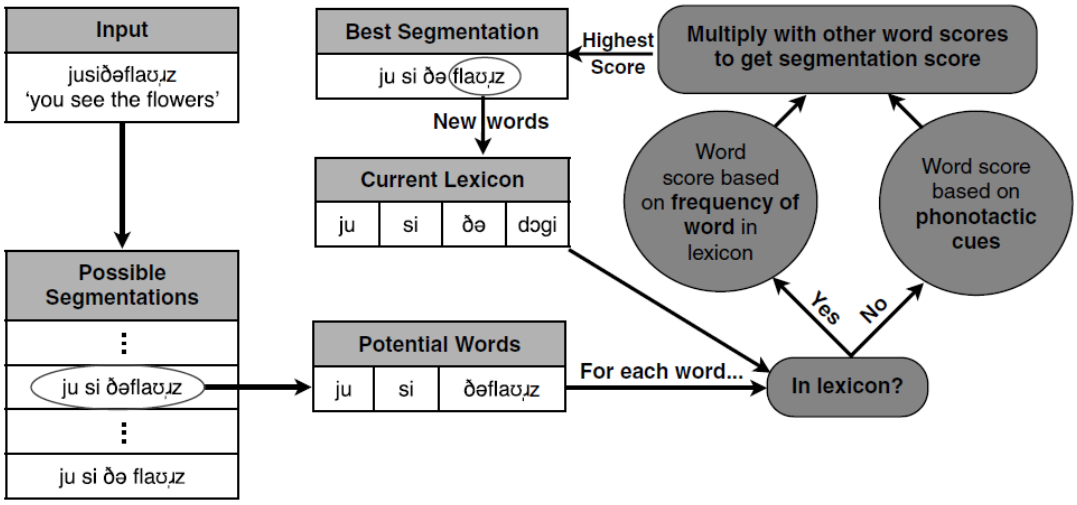


Figure 2: Outline of the functioning of the PHOCUS and the STROLL model.

4.2.2 Stepwise procedure

Equal to the PHOCUS model, STROLL creates a set of all the possible segmentations of the input sentence that is to be analyzed. For all these segmentation options, STROLL calculates the probability of the sentence given the previously created segmentations. The probability of a sentence is calculated by scoring the forms in it. The details of this scoring procedure are discussed in section 4.2.3. The STROLL model distinguishes between forms it has seen before and forms which are novel, both types of form have a scoring strategy using a segmentation cue. The segmentation with the highest likelihood is selected as the correct one and used to update the lexicon. The next part of this section discusses the mathematical specifications of the calculation of the most probable segmentation. This part presents the procedure step by step.

A filter is applied in the process of creating all possible segmentations of an input sentence. It filters out segmentations that include a form which does not contain a vowel. Experimental evidence supporting

the syllabic requirement of forms comes from Bertoncini and Mehler (1981), who have found that already before they are 2 months old, infants prefer syllabic over non-syllabic sequences. In the PHOCUS model, a universal phonotactic constraint with a similar effect was optionally included. This constraint states that a segmentation with a non-syllabic part receives a probability of zero. The universal phonotactic constraint was optionally included. It improves the results for PHOCUS the model when tested on English data but it did not improve results for a corpus in another language, Sesotho, which contains more words consisting of just one syllable (Blanchard et al., 2010, p505). In the STROLL model, the filter requiring a syllabic element in each segmented form is not optional. Any segmentation including a non-syllabic form is a priori excluded.

For every sentence in the input text, the model follows these steps:

1. Read in the input sentence, the lexicon and the n-gram probability distribution.
2. Create all possible segmentations of the input sentence, except for segmentations containing vowelless parts.
3. For every segmentation, calculate the probability based on the probability of the known and the novel forms in the segmentation.
4. Select the segmentation with the highest probability as the winner.
5. Update the lexicon by adding the forms from the winning segmentation after applying the frequency or peak filter.
6. Update the n-gram probability distribution by including the n-grams of the processed sentence.
7. Go to the next sentence.

4.2.3 Mathematics

The probability of a segmentation is calculated by multiplying the probabilities of each of the forms in the segmentation. In determining the probability of a form, it depends whether this form is known or novel. The probability of a known form is related to the frequency with which it has been segmented from the input corpus so far. The probability of a novel form is related to the n-grams which constitute the form and their probabilities. A known form receives a probability equal to the likelihood of its occurrence, given all forms that could occur. Not all forms that could occur are in the lexicon, since not all of them have been encountered or segmented. This causes the requirement that part of the total probability mass, which sums up to 1, needs to be reserved for novel forms. First is discussed how much mass needs to be reserved, how the mass is reserved is discussed next. The amount of probability mass for novel forms is calculated by dividing the number of novel forms in the lexicon, defined as forms with a frequency of 1, by the total number of tokens in the lexicon. This is an indication of how likely it is to encounter a novel form. A novel form in a possible segmentation receives a score equal to the product of the probabilities

of its n-grams, multiplied with the probability of encountering a novel form.

The probabilities of the known forms are their frequencies divided by the total number of encountered forms. However, in order for the total probability mass to sum to 1, the frequency of a known form cannot only be divided by the sum of the frequencies in the lexicon. This does not save the probability mass for novel forms. To achieve this, all the probabilities of known forms need to be scaled down. This is done via Good-Turing smoothing (Good, 1953), as implemented in Daland et al. (2011). Good-Turing smoothing calculates the exact size of the probability mass and scales down the probabilities of the known forms in the lexicon to reserve this mass. In the STROLL model forms are added to the lexicon only after they pass certain restrictions. When they are in the lexicon, their frequency counts are increasing with every new time they are encountered. There is a threshold for forms to enter the lexicon, but when they have passed it their frequency increases normally. This leads to a relatively low number of forms in the lexicon with a probability of 1. It even occurs that there are no forms in the lexicon that are defined as novel when calculating the probability of a novel form. This leads to a probability of encountering a novel form as 0, making it impossible to add novel forms to the lexicon. To counter this, the lexicon contains a form that always has a frequency of 1 that is not raised. The inclusion of this form ensures that a novel form can always receive a probability higher than 0. It lowers the threshold for entering the lexicon for all novel forms.

The formula for calculating the probability of a segmentation, Pr(segmentation), is presented in a) and b). 'familiar' and 'novel' are lists of the familiar and the novel forms in the segmentation. The probability of a novel form is calculated as in b), which is contained in the formula in a). Pr(novel) in b) denotes the probability of finding a novel form as discussed above:

a)

$$Pr(segmentation) = \prod_{x \in familiar} Pr(x) * \prod_{y \in novel} score(y)$$

b)

$$score(z) = Pr(novel) * \prod_{ngram \in z} Pr(ngram)$$

4.3 Restrictions on lexical learning

The STROLL model is an implementation of the PHOCUS model extended with restrictions on lexical learning. These restrictions are added to simulate the infant word learning process more realistically. Following from the discussion on relevant factors in the lexical acquisition process in chapter 2.3, two distinct restrictions are formulated. Besides these two restrictions, there is the vowel filter discussed in the previous section. This is a restriction on which forms are possible words and which are not. The restrictions discussed in this section are specifically relevant to the word learning process. The first

restriction is a frequency threshold, requiring a form to be encountered a set number of times before it is added to the lexicon. The second restriction is a peak filter. This requires a form to be encountered a set number of times within a set number of sentences before it is added to the lexicon.

4.3.1 Frequency threshold

The frequency threshold is implemented as a representation of the stipulation that infants (as well as adults) need to perceive a form multiple times before it is adequately stored in memory. Roy et al. (2009) have presented evidence that input frequency is a significant predictor of when a word is uttered for the first time. The implementation of the frequency threshold is similar to the implementation of the frequency factor by Swingley (2009) in postulating the number of words available for acquisition under several conditions (see figure 1). It is not known how many times an infant needs to be exposed to a word form before it acquires it. As is discussed in section 2.3, there are many factors which can be of overruling influence over the frequency factor. The frequency restriction is tested with many different threshold values, to adequately simulate the variance in – and account for the lack of information on – the required number of exposures of a form before it is acquired.

4.3.2 Peak filter

The peak filter simulates a type of environment in which infants often perceive words, for example in a story telling situation (Jusczyk and Hohne, 1997). The peak filter restricts lexical acquisition to forms which occur a set number of times within a set amount of time; forms which show a frequency peak. To represent the time element of the peak filter, a number of input sentences is defined within which the form needs to be segmented enough times. The number of sentences is set at 5, 10 or 20. Looking only at isolated words, Brent and Siskind (2001) found a repetition of 27.2% of forms within 30 seconds. The peak filter is applied to test the usefulness of this source of information in the word learning process.

4.3.3 Testing procedure

The model is subjected to a set of simulations, the results of which are presented in the next chapter. The STROLL model is tested using the Bernstein-Ratner corpus (Bernstein-Ratner, 1987), in the transcribed version of Blanchard and Heinz (2008) which does not contain word boundaries. This corpus of child-directed speech consists of 9790 utterances, containing 33,399 tokens in phonemic transcription. Each utterance is presented and processed incrementally. Because the model is incremental, there is no distinction between a training and a testing phase. However, the model starts with an empty lexicon and no information on the n-gram probability distribution. Consequently, the first 1000 utterances are removed from the evaluation procedure. This ensures that the model is not scored on its initial learning

phase, but only on the utterances that it can segment with a higher level of certainty. The same procedure of excluding the first 1000 output sentences from the evaluation is performed in Blanchard et al. (2010). For the evaluation procedure, the model uses the output lexicon and the winning segmentations of the processed sentences. The evaluation procedure is performed with the notions discussed in section 3.1.

The simulations presented in the next chapter test the STROLL model on different settings of a few parameters. The dimension that is of central interest is the one related to the restrictions on word learning. Results are presented for the model run without a restrictive filter, with a frequency threshold and with a peak filter. For the filters, different values of restrictiveness are compared. The values are different heights of the frequency filter and different time frames plus frequency thresholds for the peak filter. Another parameter that is tested is the n-gram cue, the model run with bigrams is compared to one working with trigrams. All runs of the model apply the lexical cue and the n-gram cue, albeit the latter with two different values.

5 Results

This chapter presents the results of the STROLL model. The first two parts of the sections on the frequency threshold and the peak filter are focused on the primary results of the segmentation model: the lexicon it has produced and the produced segmentation of the input text. The model is run separately for the two filters, and for different value-settings for the two filters. The third part of the two sections focuses on the development of the influence of the lexicon on the segmentation results.

To check for the confounding factor of utterance order inside the corpus, the STROLL model with the two word learning factors is tested on a rearranged version of the corpus as well. The corpus is rearranged in two ways, one in which all sentences individually are put in a different order, one in which they are moved around in blocks. The distinction is predicted to be especially relevant for the model including the peak filter, since rearranging utterances individually could break up frequency peaks that are present in the corpus. If the produced lexicon and segmentation performance are significantly worse in the individual utterance rearrangement but not in the block rearrangement, this shows the presence of frequency peaks in the natural infant-directed speech collected in the corpus. The results of the STROLL model with the rearranged input corpora are presented for the frequency threshold and the peak filter in section 5.1.4 and 5.2.4 respectively.

A second adjustment of the corpus is presented in 5.1.5 and 5.2.5 for the frequency and peak filters respectively. This adjustment is an extension of the corpus. The Bernstein-Ratner corpus contains an amount of input approximately equal to what an infant receives in a day (Van de Weijer, 1998; Daland and Pierrehumbert, 2010). The corpus is copied five times. Each copy is differently rearranged per block. This section tests whether a larger corpus makes the restrictive influence of the filters smaller, because there is more time to pass each threshold and make use of the forms that have passed it.

Section 5.3 compares the STROLL model using trigrams as the bottom-up segmentation cue with the bigrams which are used in the other simulations. The PHOCUS model (Blanchard et al., 2010) includes different versions, of which the one including the vowel-constraint – which is a standard part of the STROLL model – and using trigrams is the best performing. To see whether the segmentation results for the STROLL model also improve when switching to trigrams instead of bigrams, this test is performed. Section 5.4 presents a more detailed comparison of the STROLL model with the parameter settings in which it performance best. It also qualitatively compares the 50 most frequently produced lexical types for the two word learning restrictions.

5.1 Frequency threshold

This section presents the results of the STROLL model including a frequency threshold. As discussed before, more frequent words are learned faster. To simulate the learning procedure more representatively, including this frequency bias, a threshold is added which adds forms to the lexicon only after they have been segmented a set number of times. This section presents an evaluation of the lexicon generated by the model first, followed by an evaluation of the boundaries the model has located. Both sections present the evaluation of the STROLL model which is run with different settings for the frequency threshold parameter. This section further analyzes the influence of the lexical cue – recognition of familiar forms – on the segmentation performance, and how the frequency threshold affects this influence. The last two parts of this section test a small set of frequency threshold settings on modifications of the input corpus, with the utterances rearranged and with a larger corpus to provide more learning data.

The model is run with different values for the frequency threshold. All integer numbers between 1 and 15 are used as a value, between 15 and 40 steps of 5 are taken, and from 40 to 100 every tenth number is a value. As a transition between the one- and the five-sized steps, 17 is included as an extra value. These 27 values give an overview of the development of the model's performance as the frequency restriction becomes more strict. The frequency threshold with value 1 is included in the tests as a measure for the model without a frequency restriction on word learning.

5.1.1 Lexicon generation

The STROLL model is run with 27 different values for the frequency threshold. For each run, the generated lexicon is tested on the types it contains. These types are tested for precision, recall and F0 scores (see section 3.1). As can be seen in figure 3, the higher the frequency threshold, the lower the recall. This is expected, since a higher frequency entails that less forms are added to the lexicon. The drop in recall occurs fast: at a threshold of 9 encounters or higher, less than 10% of all word types have been acquired by the model. The F0 score, which is a combination of the precision and recall score, is highly influenced by the low recall and drops below 10% at a frequency threshold of 12. Both the recall and F0 scores require the frequency threshold to be as low as possible, and preferably even absent. Contrary to the recall and F0 score, the precision of the model stays high throughout the threshold increase. The only threshold values resulting in a score of less than 60% of all detected forms being correct, are 1 and 90. There does not seem to be a reliable pattern in the precision scores for the different threshold values. The highest precision is reached for a threshold of 19, with 82% precision. The lowest precision score of 47.3% is reached by the frequency threshold of 1 (equal to no threshold). This suggests that it is beneficial for the model's precision scores to include a frequency threshold. Between no frequency filter and

the least restrictive one the precision and recall scores shift, causing no change in the F0. The absence of a filter causes more word forms to be added to the lexicon. A small frequency threshold causes a drop in number of word forms, but an equally sized rise in quality of the forms that are found.

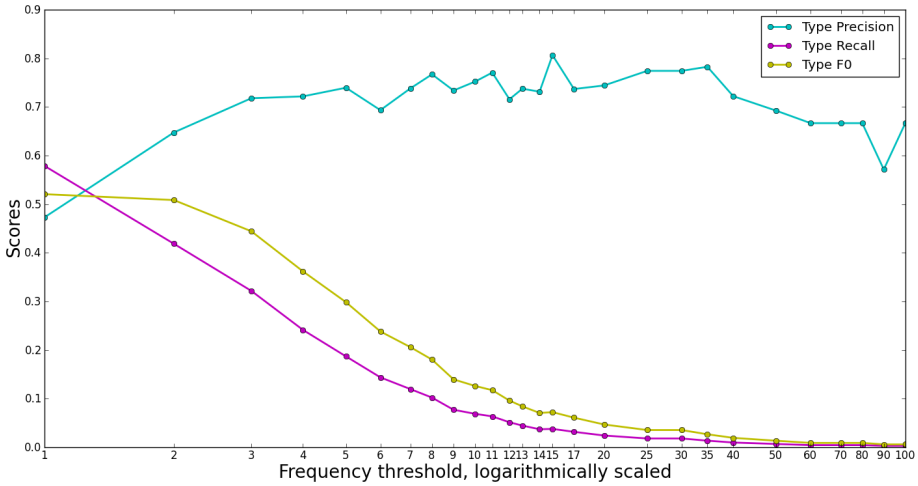


Figure 3: Overview of the lexical output evaluation for 27 different values of the frequency threshold. The lexical output are the types that are found in the lexicon. They are scored using precision (of the found types, how many are correct), recall (of all correct types, how many are found) and F0 (a combinatory score of precision and recall).

5.1.2 Segmentation performance

This section presents two distinct ways in which the segmentation performance of the STROLL model is evaluated. The segmentation performance is the ability of the model to locate word boundaries in the input. As presented in section 3.1, this performance can be measured using the same measures as used to evaluate the types (precision, recall and F0), or using the Signal Detection Theory. Graphs are presented for both measures (figure 4 and 5). The general slope of the graph showing the precision, recall and F0 scores for the located word boundaries (figure 4) is the same for all three score types. The higher the frequency, the lower the scores. The decrease in output scores occurs more rapidly with the lower frequency thresholds, and seems to level off with frequencies higher than 20. The decrease does not occur evenly, as can be seen by comparing the output of frequencies 14 and 15. An interesting result is that for a frequency threshold of 2, the boundary location results are better than those of no frequency threshold. This suggests that with a slight restriction on which word forms are added to the lexicon the model performs better than with no restriction. Figure 4 shows a higher precision score than recall, this suggests that the model undersegments. It is performing better on the quality of the forms that it does find than on the number of detected forms.

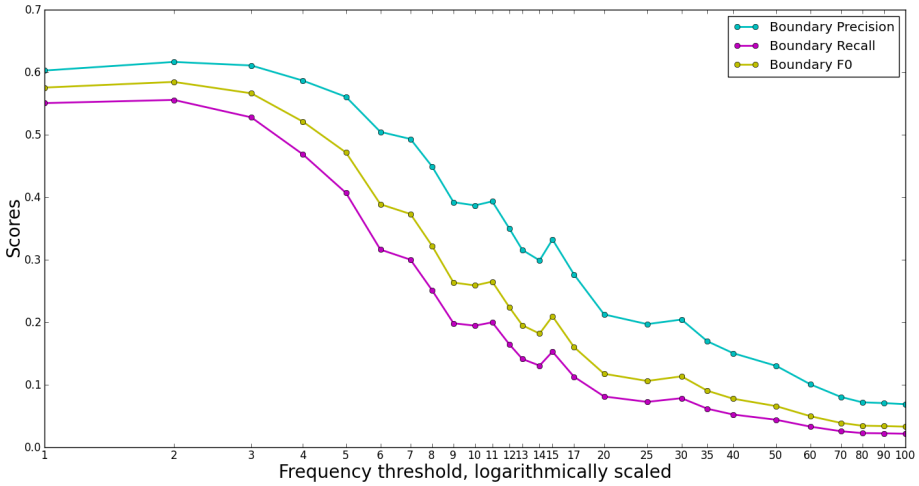


Figure 4: Precision, recall and F0 results for the boundaries as located by the STROLL model with different frequency thresholds.

The hit rate is calculated from the same measures as the recall and thus shows the same pattern of improvement for no versus a small frequency threshold, followed by rapid decline that seems to level off in the high threshold values, see figure 5. As explained in section 3.1, the false alarm rate is the only measure that should be as low as possible for the model to perform well. This rate interestingly stays stable across different frequency thresholds, which is discussed in the next chapter. The two figures on the word boundary result show that the highest performance is achieved with a frequency value of 2. This complies with one of the two highest performances for the lexical results. A frequency threshold of value 2 results in the best segmentation and lexical scores.

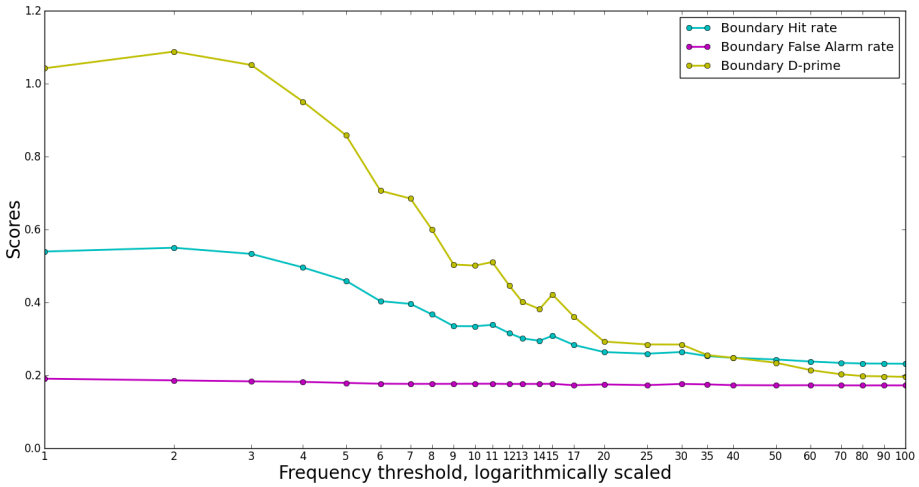


Figure 5: Mean hit rate, false alarm rate and d-prime for the boundaries as located by the STROLL model with different frequency thresholds.

5.1.3 Cue development

Because the word forms in the lexicon are used as a segmentation cue, it is interesting to see the effect of the frequency filter on the effectiveness of this cue. To measure this, the STROLL model is run with the same 27 different frequency thresholds as above. After every 500 analyzed utterances, the size of the lexicon and the F0 scores of the boundaries located thus far are measured. If the influence of the lexical cue is constant and not affected by the frequency cue, the correlation between the growing size of the lexicon and the improving boundary detection ability, measured in F0 scores, should be constant. A constant correlation between the lexical growth and boundary score growth over different frequency thresholds signals that a change in lexicon size leads to an equal change in segmentation performance. The influence of the lexicon on the segmentation performance, as measured in correlation of growth, gets worse with higher frequency thresholds (see figure 6, r = -0.9). The relation between number of forms in the lexicon and the segmentation performance is not constant. It seems that the size of the lexicon has a bigger effect on the segmentation performance than the expected linear one.

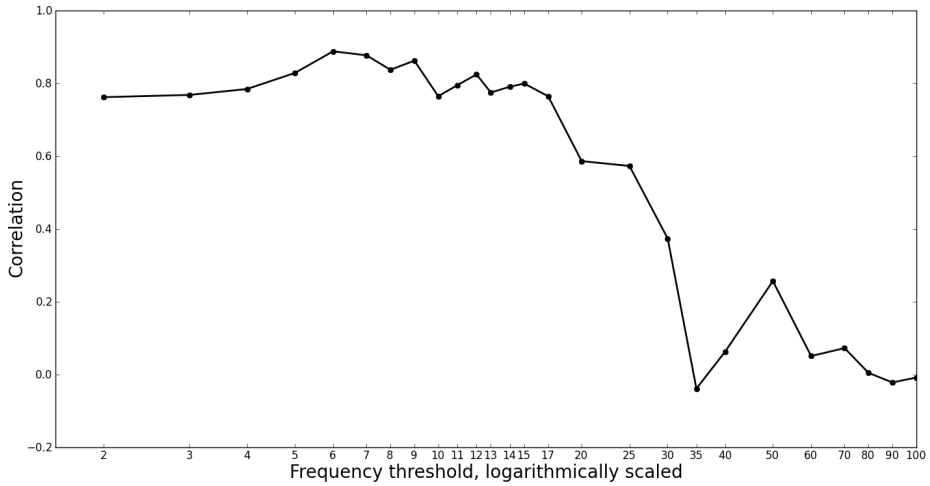


Figure 6: Correlation between growth of lexicon size and improvement of segmentation performance measured at 500 utterance intervals, for 27 different heights of the frequency threshold.

5.1.4 Rearranged corpus

The order in which the utterances are presented to the model could be of influence to the output results. In order to test this possible confounding factor, the STROLL model has been run using a small set of frequency thresholds (2, 5, 10, 24, 50, 100) on the same Bernstein-Ratner (Bernstein-Ratner, 1987) corpus, but with rearranged utterances. The utterances are rearranged in two ways: the first has randomized the order of every single sentence, the second has randomized sets of 250 utterances. The distinction between these types of reordering is mostly of interest for the peak filter, which is discussed in the appropriate

section.

The output of the models run on the corpus used thus far, on the corpus rearranged per sentence and per block of utterances show the same pattern of token, type and boundary F0 scores. There are differences between the scores, but they are small. All three runs show the highest performance for at least one of the six selected frequency thresholds. Figure 7 shows the token F0 results as a representative of the type, token and boundary results. For the frequency threshold, the order in which the input utterances are presented to the model does not cause consequential differences in the results.

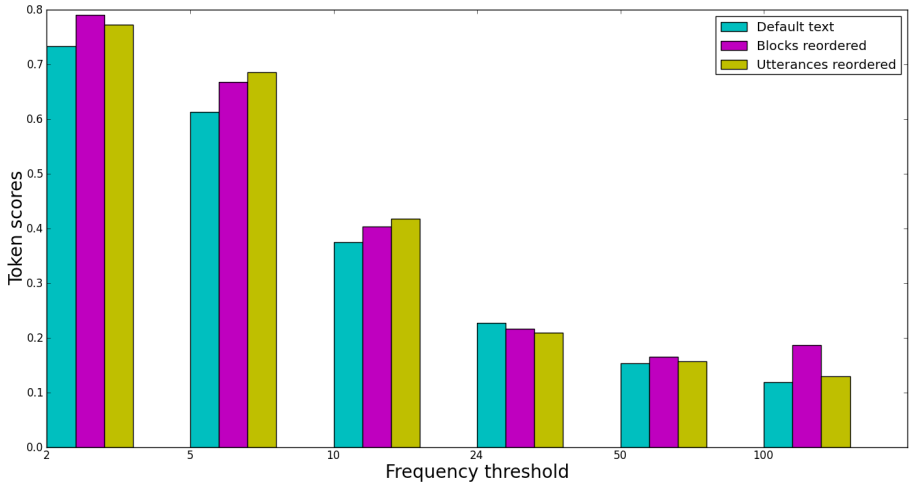


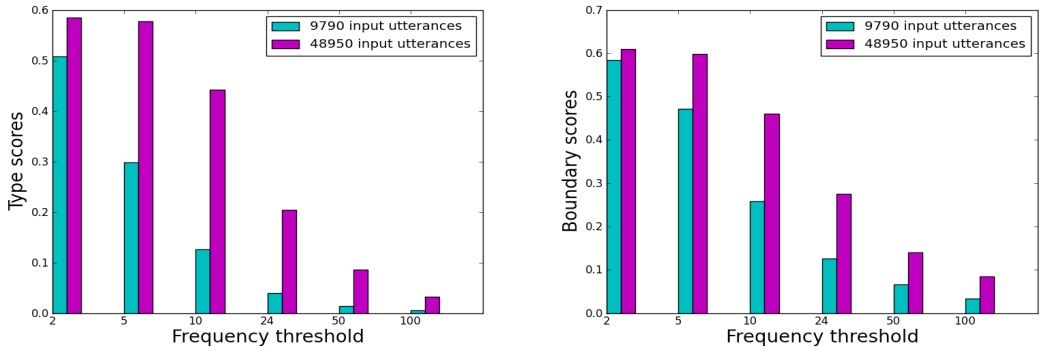
Figure 7: Token F0 scores for the STROLL model with six different frequency thresholds and three different input texts: the one used thus far, one with every sentence reordered and one with blocks of 250 utterances reordered.

5.1.5 Looped corpus

A stricter learning process logically requires more input. If there are restrictions on learning, the process is slowed down. This section tests the performance of the segmentation model with a larger input corpus. To limit the number of varying and possibly confounding factors, the same input corpus as used in the previous simulations is copied several times. This ensures that the same kind of child-directed input is used, and only the factor of size is changed. Each copy of the corpus has been scrambled, changing the order of the sentences per group of 250. The corpus has been copied 5 times, resulting in a total of 48950 utterances. A day of speech input for infants consists of around 25,000 words (Van de Weijer, 1998; Daland and Pierrehumbert, 2010). The single corpus contains 33,399 tokens, little over a day worth of input. The looped corpus, containing five copies of the Bernstein-Ratner corpus, contains 166,995 tokens similar to almost a week of input.

For this experiment a subset of the values for frequency thresholds is used. The STROLL model is run with 6 different frequency thresholds: 2, 5, 10, 24, 50 and 100. The type and the boundary F0 results

are presented in figures 8a and 8b respectively, together with those of the single corpus for comparison. The larger set of input data seems to have a larger effect on the type scores than the boundary results, but overall does the model run with more input outperform the model run with less input.



(a) F0 results of types found

(b) F0 results of boundaries located

Figure 8: F0 scores for types in the lexicon and boundaries located by the STROLL model in 6 runs with different frequency thresholds, for two differently sized input corpora; one containing 9790 utterances, one containing 48950 utterances.

To see whether the threshold is equally restrictive for both amounts of input, the correlations between the height of the threshold and the performance scores are calculated. The correlation signals the influence of the threshold on the performance. A high correlation signals that the height of the threshold is highly predictive of the performance score. The prediction is that more input decreases the restrictive influence of the threshold, causing a lower correlation. The correlation scores are presented in table 11. The prediction is not supported, a larger input corpus is more strongly correlated with the frequency threshold, signaling more influence from it than in the smaller corpus. More input data causes better results for the segmentation and word learning model with a frequency threshold. However, the restrictiveness of the filter does not decline with more input, it actually has a stronger effect.

Score type	9790 input utterances	48950 input utterances
Boundary	-0.78	-0.89
Type (Lexicon)	-0.68	-0.87

Table 11: Correlation scores between frequency threshold and performance scores, for two sizes of input corpus

5.2 Peak filter

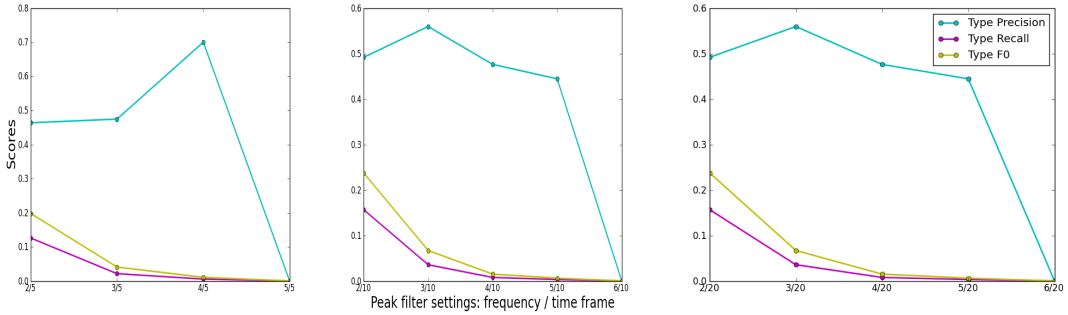
One characteristic of the way words are presented to infants is that they often occur in peaks, as discussed in section 2.3.2. To simulate this environment of word learning, the STROLL model is extended with a peak filter. This filter adds forms to the lexicon when they have been segmented a set number of times within a set number of sentences. The number of sentences represent the time within which a word is

heard multiple times. Similarly to the previous section, an evaluation of the generated lexicon and the segmentation performance are presented first, followed by an analysis of the effect of the peak filter on the lexical cue's influence on the segmentation performance. The last two sections again present the results of the model run with rearranged and extended input corpora respectively.

There are two parameters in the peak filter. Similar to the frequency filter, the filter has a frequency threshold that needs to be reached before a form is added to the lexicon. Besides that parameter the peak filter requires a time frame, denoted in a number of sentences within which the threshold needs to be reached. The threshold is set at three different values: 5, 10 and 20 sentences. For each of these values, the threshold parameter has taken on the values from 2 to the value resulting in an empty lexicon for the entire run. For the time frame of 5 sentences, the threshold of encountering the same form also 5 times resulted in no additions to the lexicon at all. For the 10 sentences time frame, this was 6. With the time frame set at 20 sentences, a threshold of 8 encounters resulted in 0 forms in the output lexicon. All different parameter settings combined leads to 16 separate runs, of which three have no type-evaluation because they resulted in an empty lexicon.

5.2.1 Lexicon generation

Similar to figure 3, presenting the type scores for the frequency thresholds, there is no obvious pattern in the precision score of the model with different peak parameter settings shown in figure 9 with each graph showing the results for a different time frame. Threshold 4 in time frame 5 (4/5 on the x-axis in figure 9a) results in a surprisingly high precision of 70%. The recall score, however, doesn't start out strong and is decreasing with every increase of the threshold. Equal to the pattern shown in figure 9, the F0 score follows the recall score downward. All three graphs show that the higher the threshold, independent of the time frame, the worse the recall and F0. The precision score has its peak at another threshold, 4 encounters for time frame 5 and 3 for time frames 10 and 20. Comparing the height of the scores shows that the model with the peak filter performs significantly worse than the model with the frequency threshold. F0 reaches 25% at most for the peak filters, while it reaches more than 50% for the frequency threshold. The lexical performance of the peak filter suggests that the presence of the time dimension lowers the overall scores, while any raise in the frequency threshold causes a drop in recall and F0. The best results are achieved with the lowest threshold, in any time frame.

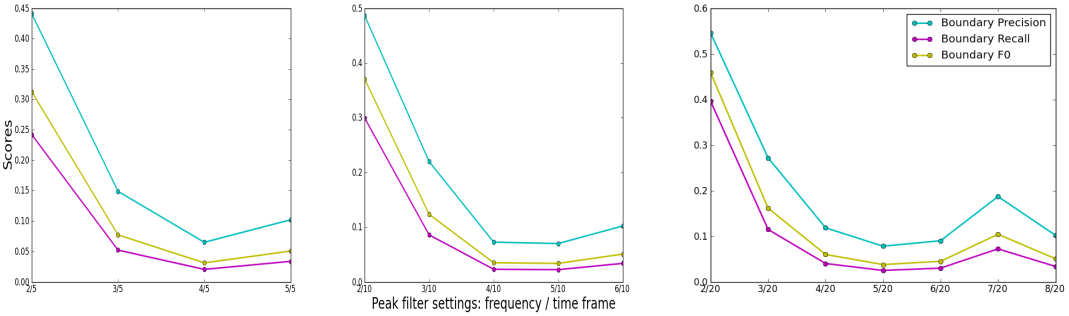


(a) Peak within 5 utt. *(b)* Peak within 10 utt. *(c)* Peak within 20 utt.

Figure 9: Overview of the lexical output evaluation for 16 different parameter settings of the peak filter. The lexical output are the types that are found in the lexicon. For three parameter settings in the figure, the number of types in the lexicon is 0 (5/5, 6/10 and 8/20). The types are scored using Precision (of the found types, how many are correct), recall (of all correct types, how many are found) and F0 (a combinatory score of precision and recall).

5.2.2 Segmentation performance

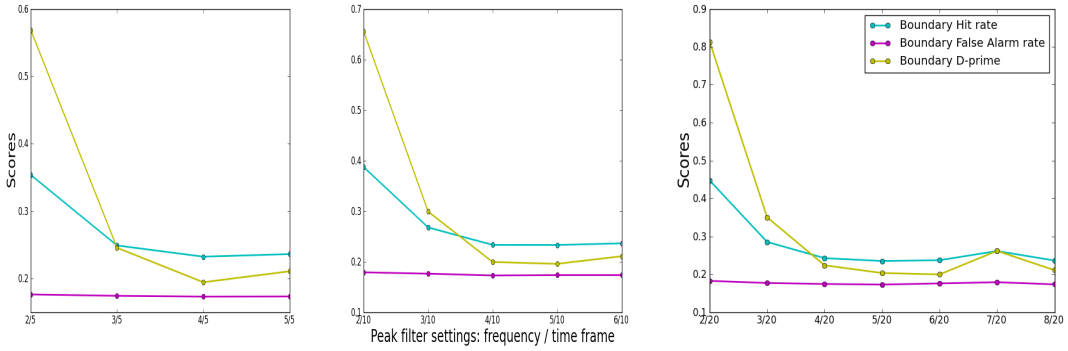
The boundary precision, recall and F0 scores presented in figure 10 show an interesting pattern. For all three different time frame values, a frequency threshold of 2 (2/5 in figure 10a, 2/10 in 10b and 2/20 in 10c) results in the highest scores. Contrary to the linearly decreasing scores with increasing frequency for the STROLL model with a frequency threshold, the peak filter results in a U-shaped pattern for all three time frames. With the exception of the parameter setting of 8 encounters within 20 utterances (8/20 in figure 10c), the increase of the frequency threshold leads to decreasing and subsequently improving segmentation results. An initial explanation could be that with the higher frequency threshold for a certain time frame, the lexical cue improves or declines in a way that affects the segmentation performance. Further interpretation of the U-shape is presented in the general discussion. Again the precision score is higher than the recall, suggesting undersegmentation.



(a) Peak within 5 utt. *(b)* Peak within 10 utt. *(c)* Peak within 20 utt.

Figure 10: Precision, recall and F0 for boundaries located by the STROLL model for 16 different time frame - frequency threshold combinations.

Similar to the results presented in figure 5, the false alarm rate of the STROLL model stays stable at around 0.18 for all parameter settings. The hit rate shows the same U-shape as the recall score, causing a similar U-shape in the d-prime. Similar to the lexical performance, the segmentation performance shows that the time frame dimension does not have a significant effect on the results, though expectedly the longer time frames lead to better results. The frequency dimension, however, causes a large effect on the segmentation results. The lowest frequency leads to considerably better results than higher ones. Interestingly, a U-shape appears, with a rise in the highest frequency threshold for all time frames.

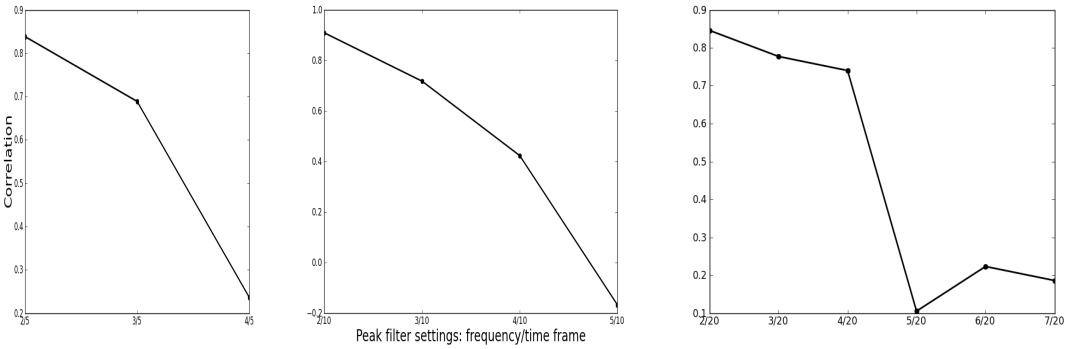


(a) Peak within 5 utt. *(b)* Peak within 10 utt. *(c)* Peak within 20 utt.

Figure 11: Hit rate, false alarm rate and d-prime for boundaries located by the STROLL model for 16 different time frame - frequency threshold combinations.

5.2.3 Cue development

For the peak filter, the influence of the lexicon size on the segmentation performance has been measured as well. Figure 12 shows a downward slope for all three time frame settings. Independent of the number of utterances in which a threshold needs to be reached, the higher the frequency threshold, the less correlation there is between the growth of the lexicon and the growth of the segmentation performance during a run of the STROLL model. The same parameter settings are used as discussed above, but the three settings resulting in an empty output lexicon have been removed. Similar to the model run with a frequency threshold, the effect of the lexicon size on the segmentation performance is not linear, which would have been expected.



(a) Peak within 5 utt. *(b)* Peak within 10 utt. *(c)* Peak within 20 utt.

Figure 12: Correlation between growth of lexicon size and growth of segmentation performance, measured at 500 utterance intervals, for 13 different frequency threshold and time frame combinations.

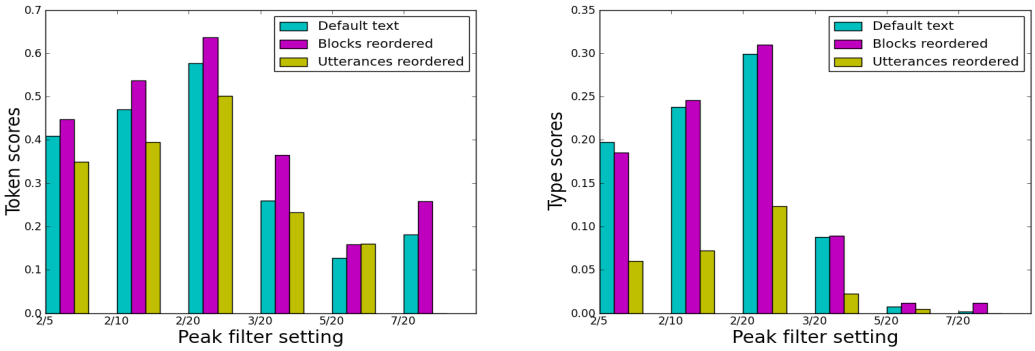
5.2.4 Rearranged corpus

The STROLL model is presented with two rearranged versions of the Bernstein-Ratner (1987) corpus. One version has every sentence replaced, the other has shuffled utterances per block of 250. For the model run with the peak filter, the two versions could result in significantly different results. The peak filter is integrated to simulate a specific environment in which infants are presented with word forms with a high frequency in a short amount of time. The Bernstein-Ratner corpus contains child-directed speech from parents who interacted freely with their child. Assuming that this is a normal circumstance in which peaks occur, the corpus with every utterance replaced should lead to significantly worse results because the peaks are taken apart. Six different settings for the peak filter are used, a subset of the settings run in the previous sections. There are three different time frames, 5, 10 and 20, and for the latter there are four different frequency thresholds to give an overview of the model's performance in both dimensions of the peak filter's parameters. The six settings are: 2/5, 2/10, 2/20, 3/20, 5/20, and 7/20.

The differences between the three versions of the corpus is larger for the STROLL model with a peak filter than with a frequency threshold. The F0 results for tokens and types are presented in figures 13a and 13b respectively. Figure 13a shows a larger divergence between the results for all time frame-threshold combinations, except 5 in 20 utterances, than for the different frequency threshold settings in figure 7. The last setting, 7 encounters in 20 utterances, produced a completely empty lexicon as output. This is visible in figure 13b, the type F0 score is significantly more affected by the reordering of every sentence compared to rearranged blocks. This supports the assumption that the original Bernstein-Ratner corpus contains words occurring in frequency peaks.

The difference between the corpus rearranged in blocks or per individual utterance is smaller for the token scores than for the type scores. The token scores are produced with a combination of the lexical

and the n-gram cue. The latter supplies stable performance for all three types of text, except for the rather strict peak filter of 7 encounters within 20 utterances (7/20 in figure 13a). The number and quality of types that are located by the model show a significant distinction between the two differently rearranged corpora. Reordering every utterance individually removes the frequency peaks that are available for word learning.



(a) Results of tokens found

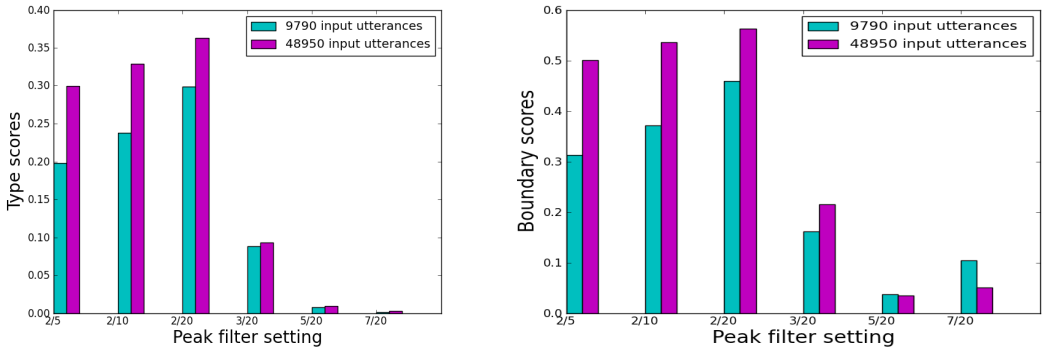
(b) Results of types found

Figure 13: Scores for the STROLL model, run with six different peak filter settings and three versions of the input corpus: the same used thus far, one with shuffled blocks of 250 utterances and one with every sentence rearranged.

5.2.5 Looped corpus

A subset of the peak filter settings has also been tested on a larger input corpus. The corpus has been copied five times, each time with sets of 250 utterances in a different order. This should keep the information needed for the peak filter intact, while providing more learning data. Instead of an input amount of a little more than an infant is presented with in a day, the looped corpus contains a little less than the amount presented in a week (see section 5.1.5). The same six peak filter settings are applied as in the previous section, for an overview in both dimensions of the peak filter.

Both graphs in figure 14 show that for the lower frequency thresholds in all time frames (2/*n*), the model run on the larger corpus outperforms the model run on the smaller one. However, for the higher thresholds, the difference becomes smaller and for the 5 and 7 encounters within 20 utterances, the smaller input even leads to better boundary results. It seems that the factor(s) causing the U-shape which occurs for the boundary results of the STROLL model with increasing frequency threshold for the same time frame affect the boundary scores of the model run with the large corpus less. For lower frequency thresholds, the larger corpus leads to considerably better results. This effect disappears for the higher thresholds.



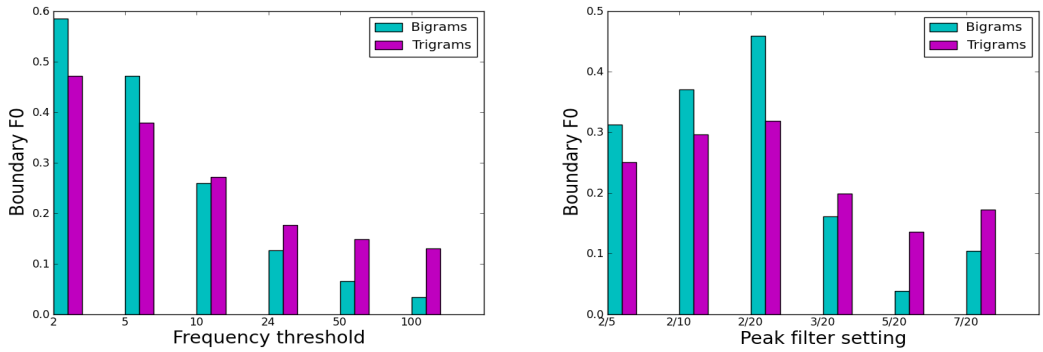
(a) Results of types found

(b) Results of boundaries located

Figure 14: F0 scores for types in the lexicon and boundaries located by the STROLL model in 6 runs with different peak filter settings, for two differently sized input corpora; one containing 9790 utterances, one containing 48950 utterances.

5.3 Trigrams

The PHOCUS model (Blanchard et al., 2010) applies three different versions of the n-gram cue. The model looks at monograms, bigrams and trigrams. The trigram version of PHOCUS including the vowel filter is the best performing version. There is no conclusive evidence whether infants are sensitive to bigrams, trigrams, or both. In this section, the STROLL model is run with the trigram cue and compared to its performance applying the bigram cue. Again, a subset of six frequency thresholds and six time frame and threshold combinations is used. The output data of the F0 values for the boundaries produced by the model are presented. The two graphs in figure 15 – graph 15a containing results for frequency thresholds and 15b for peak filters – contain the F0 values for both the bigram and trigram outputs to enable comparison.



(a) Frequency threshold

(b) Peak filter

Figure 15: Boundary F0 scores for the model with six different frequency thresholds and six different peak filter settings, run with a bigram and a trigram segmentation cue.

The graphs show that for the lower frequency thresholds, the bigrams produce a better output. However, for the higher frequency thresholds the trigram model outperforms the bigram model. An initial analysis of these results suggests that trigrams are a more independent segmentation cue, better able to segment the speech when the lexical cue is absent. When the lexical cue is actively present, because there are many forms added to the lexicon and available for recognition in the speech, the bigram is a better segmentation cue. The trigram functions better on its own and seems to be more robust, the bigram is a better cue in combination with the lexical recognition cue.

5.4 Comparative analysis

This section presents a more detailed look on the STROLL model's performance. The first part closely compares the best performing outputs of the model with a frequency threshold, a peak filter or no learning restriction. The second part compares the lexicon produced by the model with a frequency threshold with the lexicon produced with a peak filter. This gives insight in whether the two restrictions on lexical learning prefer the same types of forms to be learned.

5.4.1 Best results

This sections presents the results of the best performing versions of the two restrictions on lexical learning and the default model without restrictions. This gives an overview of the effect of the restrictions on the segmentation results of the STROLL model. Table 12 presents the precision, recall and F0 scores for word types (lexicon) and boundaries. The type results show that both forms of restriction lead to better precision, with the frequency threshold leading to the best result by far. The recall score shows the opposite results, restrictions on lexical learning lead to a lower recall score. This score also shows the largest divergence for the three versions of the STROLL model. The boundary results show no significant differences in the models' performance, although the model with a small frequency threshold performs the best for all three measures.

	LP	LR	LF	BP	BR	BF
No restrictions	47%	58%	52%	60%	55%	58%
Frequency threshold: 2	65%	42%	51%	62%	56%	58%
Peak filter: 2/20	55%	21%	30%	55%	40%	46%

Table 12: Precision, recall and F0 for types in the lexicon (LP, LR, LF) and for boundaries (BP, BR, BF) produced by the STROLL without restrictions on word learning, with a frequency threshold of 2 segmentations and with a peak filter of 2 segmentations within 20 utterances.

5.4.2 Lexicon comparison

The 50 most frequently generated word forms by the model in different parameter settings are presented in table 13. These forms were generated by the most runs of the model with different frequency thresholds and peak filter settings. The most interesting distinction between the words generated by the model with a frequency threshold and the model with a peak filter, is the number of function words versus undersegmented (but syntactically intact) components. The set of words resulting from the frequency threshold contains approximately 14 function words, for the words resulting from the peak filter there are only 6. The former have 7 undersegmented parts segmented as word forms, e.g. WAtsDIs, while the latter has resulted in 17 of those (e.g. WAtsyurnem, WErIzhi). The number of content words (e.g. Suz, b9It) is equal for the two groups at 14. Interjectory forms (oke, hElo) occur similarly often as well, respectively 15 and 13 times.

The high number of function words occuring in the output lexicon with a frequency filter is not surprising. Functors are the most frequent word forms in English. The results of the peak filter suggest that functors do not occur together within a smaller time frame, or when they do they appear in fixed expressions which are repeated. The functor Iz (is), has frequently passed the frequency threshold, but not the peak filter. It does however occur in multiple combinations: DErhiIz, IzhislipIN, WErIzhi.

Frequency Phonemic word form	threshold Transcription	Peak Phonemic word form	filter Transcription
WAtsD&t	what's that	WAtsD&t	what's that
lUk	look	pik6bu	pikeboo
no	no	ke	(o)key
oke	oke	9ImhANgri	I'm hungry
y&	yeah	Suz	shoes
DEr	there	lUk	look
WAt	what	oke	oke
h9I	hi	pUShImIn	push him in
hElo	hello	lEtmi9Ut	let me out
pik6bu	pikeboo	mOr	more
hir	here	pUS	push
ke	(o)key	WAtduyuwant	what do you want
si	see	WAtsyurnem	what's your name
&nd	and	gUd	good
gUd	good	h9I	hi
r9It	right	pliz	please
yEs	yes	tIkL	tickle
yu	you	wAn	one
&t	at	D&t	that
wAn	one	WAtsDIs	what's this
DIs	this	b9It	bite
D&t	that	hElo	hello
Iz	is	no	no
Olr9It	alright	sItd9Un	sit down
WAtsDIs	what's this	6	eh
huz	who's	DErhiIz	there he is
tu	too	Eg	egg
b9Ib9I	byebye	Its6we	it's away
Suz	shoes	IzhislipIN	is he sleeping
T&Nkyu	thank you	OlgOn	all gone
Tri	three	Su	shoe
gel	Gale	WAt	what
&lIs	Alice	WErIzhi	where is he
mOr	more	b9I	bye
9I	I	b9Ib9I	byebye
D&ts	that's	bAt˜	bottle
IzIt	is it	cEr	chair
bUk	book	gEtIt	get it
blaks	blocks	kIs	kiss
gRl	girl	klozD6dOr	close the door
n9U	now	n9Itn9It	night night
se	say	opD̃6dOr	open the door
arDoz	are those	r9It	right
dOgi	doggy	rIN	ring
duyuwant	do you want	spun	spoon
fOr	for	y&	yeah
fl9URz	flowers	yumed6bIg	you made a big
n9Itn9It	night night	&nd	and
6	eh	6nADR	another
It	it	DIs	this

Table 13: Most frequently generated word forms

5.5 Preliminary summary

This chapter presented the results of a set of simulations done with the STROLL model. This section briefly summarizes some of the most interesting results.

The lexicon produced by both filters performs better on precision, showing a higher quality of the segmented forms than for the model run without a restriction on word learning. The recall and consequently the F0 scores are however lower for the model run with a filter. In order to generate the lexicon containing the highest quality and highest number of types, a filter with the lowest possible value should be applied. The frequency threshold leads to higher lexical scores than the peak filter. There are slightly more function words in the lexicon produced by the frequency threshold, and slightly more undersegmented subphrases in the lexicon produced by the peak filter. There seems to be no linear relation between the size of the lexicon and the segmentation performance, suggesting that there are more factors influencing the performance than just the n-gram cue and the lexical cue measured in size of the lexicon.

The segmentation results of both models suggest undersegmentation, placing too few rather than too many boundaries. The false alarm rate seems not to be sensitive to the presence and settings of the restrictions on word learning. Similar to the lexical performance, the frequency threshold leads to better segmentation results than the peak filter. Several of the different peak filter results seem to occur in a U-shape.

A larger input corpus leads to better results, but not a lower restrictiveness of the frequency threshold. The natural order in which utterance are spoken to infants contain frequency peak. These can be useful as a word learning cue. When the utterances in the corpus are randomly rearranged, the frequency peaks are removed. The trigram cue leads to more robust results across different settings of both filters than the bigram cue.

6 General discussion

The first two sections of this chapter discuss the results of the previous chapter. The third section presents a discussion of the STROLL model. The final section of this chapter discusses the topic of this thesis, a computational segmentation model that includes factors of the infant word learning process.

6.1 Frequency threshold

The frequency threshold has an improving effect on the quality of the forms that are added to the lexicon. As can be seen in figure 3, the precision of the located words improves from 47% to 65% for a frequency threshold of 2 with respect to no frequency filter. The filter causes a higher precision for the located words, a higher quality. This effect is the result of the fact that the model with a frequency threshold needs to be more certain of a word form before it is added to the lexicon than a model without such a restriction. Interestingly, the requirement to encounter a form multiple times before it is considered a word form does not cause a parallel improvement in the lexicon. Only for the first few increases of the threshold the precision rises. The precision score fluctuates between 70% and 80% for most of the frequency thresholds until the threshold is higher than 50. These results suggest that the lexical performance of the STROLL model increases with a frequency threshold. This indication is specified by the recall results. With every increase of the frequency threshold, the recall results decline. This too is an expected result. The higher the threshold for a form to be added to the lexicon, the less forms that pass it. Considering the combined precision and recall results, the STROLL model performs better with a frequency threshold than without, but the threshold should have the lowest value possible.

The effect of the frequency threshold is predicted to decline with more input data. Once a form passes the threshold, it is added to the lexicon and together with the increasing frequency count in the lexicon the influence of the form as a segmentation cue increases. A less restrictive effect of the frequency filter is helpful for infants in raising the recall score, and enabling them to learn more word forms. Increasing the amount of input enables more forms to pass the treshold. To test whether the larger corpus diminishes the restriction from the threshold, the height of the threshold should be of less influence to the segmentation results for the larger corpus. Comparing the performance for different threshold values is a measure for the influence of the threshold. The correlation between the height of the threshold and the segmentation performance should be higher for the smaller corpus, suggesting a larger influence of the frequency threshold. This seems not to be the case when looking at the correlation scores presented in section 5.1.5 and table 11, which show a higher correlation between the height of the frequency threshold and the segmentation performance with the larger input corpus. This suggests that more input does not directly counter the frequency threshold.

Daland and Pierrehumbert (2010) have inserted a frequency threshold in the Lexical-DiBS model. The threshold was only tested on three values in the DiBS model – 10, 100 and 1000 – inducing a more coarse-grained analysis of the effects of the threshold than for the STROLL model. The large input corpus enabled the DiBS model to be successful despite a frequency threshold of 1000. The frequency threshold of 100 already limits the STROLL model to generate a lexicon containing only 6 forms. The input corpus for the DiBS model contains more than 20 times as many tokens as the input to the STROLL model, and even compared to the simulation with the enlarged corpus the DiBS model received considerably more input. However, the larger corpus did not seem to lower the restriction of the frequency threshold in STROLL. Further research is required to understand why the frequency threshold in the STROLL model is more restrictive and fine-grained adaptations cause a larger effect than for the Lexical-DiBS model.

The qualitative analysis of the lexicon and the types it contains presented in table 13 shows 28% function words. Section 2.3.1 discussed experimental research on the function of word types – especially function versus content words – in the word learning process. This research suggests that because of the fact that functors are usually reduced in speech or cliticized, and despite their high frequency, function words take longer to be recognized and accurately perceived. This starts to develop around 11 months (Shi et al., 2006), while content words are accurately recognized by infants of 7.5 months (Jusczyk and Aslin, 1995). Given the focus of the STROLL model with a frequency filter on the number of encounters for a form, a high percentage of acquired functors is expected. Of the 50 most frequent forms in the gold text 68% is a function word. This is a considerably higher number than the 28% located by the STROLL model. The frequency filter in the STROLL model simulates the infant behavior of having trouble with functors despite their high frequency.

The STROLL model run with trigrams shows interesting results compared to the model run with bigrams. For lower frequency values – both for the frequency threshold and for the peak filter with any time frame setting – the bigrams lead to better performance. For frequency thresholds higher than 5, the trigrams perform better. In the latter case the lexicon is smaller, as is the role of the lexical cue in the segmentation process, leaving a bigger role for the trigram cue. The STROLL model seems to be more robust with the trigram than with the bigram cue, it is less affected by the frequency threshold or the peak filter in the former case.

The segmentation results presented in figure 4 show a similar undersegmentation pattern as the lexical results of figure 3. The fact that precision scores are higher than recall suggests that the model is under-segmenting. It inserts less boundaries than there are in the gold text. Recall denotes the percentage of all correct word boundaries that are found by the model. A low score here is a sign of undersegmentation. Especially in combination with a higher precision score, which states that the boundaries that are located are correct. Combined, the measures denote that correct boundaries are located, but not enough of them.

Predictably, a restriction on word learning, which decreases the influence of one of the two segmentation cues, leads to correct boundaries, but also lowers the number of boundaries that are detected.

Figure 5 in the previous chapter shows almost no change in the false alarm rate for different frequency thresholds. In the calculation of the false alarm rate, the measures for the number of false positives (located boundaries which are incorrect) and true negatives (location which correctly is defined as no word boundary) are used. In other words, this is the percentage of located boundaries which should not have been placed. The stability of the false alarm rate with the different thresholds suggests that it is not influenced by the size or quality of the lexicon. Instead, it could be a reflection of the model's default segmentation performance, only based on the segmentation using the n-gram cue. Or the false alarm rate could mainly be influenced by the model's willingness to place a boundary, which is the level of certainty needed by the model for placing a boundary and which is constant across different frequency thresholds. The false alarm rate for the different settings of the peak filter are constant as well, at the same result of around 0.17 as for the frequency threshold. The six threshold and six peak filter settings that are run with the trigram cue result in the same 0.17 average false alarm rate.

6.2 Peak filter

A quantitative analysis of the lexicon produced by the model run with a peak filter shows a similar pattern as for the frequency threshold. The precision is high for any peak filter setting, as long as it is not so restrictive that there are no forms added to the lexicon. For all peak filter values, the recall scores are small. A comparison of the qualitative analyses of the produced lexicon for the two types of restrictions shows an interesting difference. The STROLL model with a peak filter results in 34% of the most frequently produced forms in the lexicon being undersegmented components. It appears that these parts of speech are common in the infant directed speech, and they especially occur in frequency peaks. Undersegmentation is preferred over oversegmentation in the segmentation process. An undersegmented form can be further segmented to the correct form in a later stage, while an oversegmented form requires a whole new procedure of joining forms to come to the correct ones.

The difference between the forms that are most frequently added to the lexicon by the two types of restrictions on word learning show a difference in the word learning strategies. Purely focusing on the frequency of a form leads to acquisition of more function words, even though in the actual language acquisition process this is impaired by the lack of phonetic salience of these forms. Focusing on peaks in frequency leads to standard phrases, that are considered as one word form. Both these sources can be useful in the word form acquisition process. It is interesting to see that different factors of the word learning process lead to different forms in the output lexicon. This suggest that similar to the segmentation process, which combines distinct segmentation cues, the word form learning process could be a

combination of different learning strategies using different learning cues and resulting in different types of forms to be learned.

As discussed in the introduction, the experimental side of the word segmentation research is focused on defining which cues are available to infants in the segmentation process. The computational side focuses on whether these cues extract enough information from the input speech to correctly locate word boundaries. This computational thesis has shown the usability of the peak filter for the word learning process. Section 5.2.4 presented the STROLL model with six peak filter settings run on three distinct input corpora: the original corpus, the corpus with utterances rearranged in blocks and the corpus with each utterance individually reordered. The original corpus is a combination of the natural speech of nine mothers to their infants. If there are useful frequency peaks in the natural speech infants receive, the produced lexicon should contain considerably less forms when the utterances in the input corpus are put in a random order compared to when the utterances are kept in the order in which they were uttered. Figure 13 shows that this is the case. Frequency peaks are present in the input infants receive and consequently, they can be used as a word learning cue.

The usability of the peak filter in the actual infant development depends on several factors. The results presented in this thesis show the presence of frequency peaks in the input and that using these as a word learning cue provides useful information. Brent and Siskind (2001) also showed the potential usefulness of forms that occur with a high frequency in a short amount of time. They determined that 27.2% of forms that were produced in isolation were repeated at least once within 30 seconds. There is no experimental evidence for which time span is optimal for infants to notice a peak in frequency and use this peak. The different durations used in this thesis were 5, 10 and 20 utterances, representing approximately 10 to 40 seconds of speech input. Jusczyk and Hohne (1997) exposed infants to 30 minutes containing three short stories. The length of these stories was around 10 minutes each, increasing the duration of the peak filter drastically from 30 seconds. This thesis has shown the presence in the speech input of the peak filter and its usefulness as a cue to word learning. Further experimental research is necessary to determine the sensitivity of infants to this cue and for which frequency density in which time span this sensitivity is optimal. A similar lack of knowledge pertains the height of the frequency threshold, the number of times an infant needs to encounter a word form before it can be acquired.

For all runs with the peak filter, except the ones with the individually rearranged utterances, the token and boundary scores show a U-shape (see figures 10,13 and 15). Low frequency thresholds result in high scores – independent of the time frame – with increasing frequency thresholds the scores decline and then rise up again. This result pattern also occurs for the model run with trigrams and for the rearranged corpus, but it seems absent in the model run with the larger input corpus. An explanation for the re-occurring U-shape can be found in the size of the output lexicon and the number of forms the model has

categorized as familiar. The moment the token and boundary results start improving occurs when the number of forms in the output lexicon is less than 10. For these runs the lexical cue plays a very limited role. Conceivably, the lexical cue inhibits the segmentation process when the lexicon contains a small number of forms. It functions properly when there is a rich source of familiar forms. When there are very few – less than 10 – known forms the model relies more on the other segmentation cue, the n-grams, and the impeding effect of the lexical cue is limited. The fact that there seems to be no or a smaller U-shape pattern for the output of the model run with the larger input corpus is understandable given that for the settings of 5 and 7 encounters within 20 utterances there were respectively 13 and 5 forms in the lexicon, both with the small and large corpus input. Apparently the model is not able to locate more forms that pass these filter settings, independent of the number of times the text is presented to the model.

6.3 The STROLL model

This section discusses aspects of the STROLL model in detail, the assumptions it makes and how it performs compared to other computational models. Marr (1982) distinguished two levels of analysis in computational models: the computational and the algorithmic level. The former is concerned with *what* the model does, the latter with *how* it functions. A computational model that aims to be a representative simulation of the infant development process should not make assumptions that lack experimental support on either level. The computational level is relevant to the strategy and the target of the model; the input the model receives, the output that is aimed for and the cues the model uses to get from the input to the output. On this level, the STROLL model is experimentally supported.

Computational level assumptions

The input the model receives is based on speech from a natural situation in which mothers talked to their infants. The content and difficulty level of the speech input for the model is the same as what infants receives. However, there is a difference in the transcription of the speech. Phonetic information is left out that can be helpful in the segmentation process, such as stress, while other information that makes the segmentation task more difficult – phonetic variation in pronunciation – is precluded as well. Adriaans and Kager (2010) have used a phonetically transcribed corpus, requiring their model to handle the pronunciation variation it contains. However, this corpus contained adult conversation of a significantly higher difficulty level than what infants are confronted with. Adding more phonetic information could simplify the task as well. Gambell and Yang (2005) present their model with input that contains prosodic information and syllable boundaries. Especially the latter source of information makes the word boundary locating task more simple than what infants have to face. Ideally, the input corpus is expanded with phonetic variation and prosodic information to be an even more realistic representation of the input infants receive. However, as both symplifying and complicating information is missing, the corpus is

comparable in difficulty level to the speech input.

The segmentation cues available to the STROLL model have been found to be available to infants (Mattys et al., 1999; Bortfeld et al., 2005). The n-gram cue is supported by infant word learning research as well. Storkel (2004) has found both word length and neighborhood density of a form to be predictive of the age at which a form is acquired. The implementation of the n-gram cue for novel forms causes the same effect as these two factors, a longer word receives a lower probability and a form with more phonological neighbors receives a higher probability. This is caused by the fact that a novel form receives a score based on the product of the n-grams it contains. The more n-grams, the more arguments are multiplied. Since each argument has a less than one probability, every new multiplication leads to a lower probability of the form, causing a preference for shorter forms over longer forms. The neighborhood density factor is implemented in a more indirect way. The probability of a novel form is the product of the probabilities of its n-grams, which are based on the n-grams that have been encountered so far. The probability of a form with a high neighborhood density will be high because there are many other forms that share the n-grams, raising the likelihood of the form.

It would be interesting to extend the STROLL model with more segmentation cues. Experimental research has shown that infants rely on multiple cues that are used complementarily in different proportions during the infant development (Mattys et al., 1999; Johnson and Jusczyk, 2001; Mattys et al., 2005). Ideally, all cues that are used by infants are applied in the simulation of the segmentation process, enabling the combination of segmentation cues (Christiansen et al., 1998). The structure of the STROLL model is capable of integrating more cues, simply by adjusting the probability of the located forms in a segmentation according to compliance with a cue. For example, every form that conforms with the trochaic stress pattern, dominant in the English language, could receive a boost of probability to incorporate the stress cue that infants apply (Jusczyk et al., 1999b).

Algorithmic level assumptions

Marr (1982) has defined the algorithmic level of analysis as focused on the implementation of the computational level, the mechanisms that underly the performance of the model. These mechanisms are a representation of the computational power and the memory capacity that infants have available for the simulated processes. A computational model that assumes a perfect memory makes unrealistic assumptions about the infant's situation (Pearl et al., 2010). The algorithmic level assumptions the STROLL model makes are less plausible than its computational level assumptions. The STROLL model generates every possible segmentation of an input utterance and calculates its probability before selecting an optimal segmentation. This requires a lot of memory capacity, even if the segmentations with non-maximal probability scores are instantly removed from memory. Pearl et al. (2010) have looked at Bayesian learners with limitations, among others limitations on memory. Their adjusted models were able to perform

the segmentation task despite the memory limitations, showing an interesting opportunity for a more realistic implementation of a Bayesian model on the algorithmic level.

The computational requirements are high as well, since the entire sentence needs to be perceived before the segmentation process can take place. Between perceiving an utterance entirely and analyzing it before the next utterance starts there is little time available, requiring a very fast computation of the optimal segmentation. The theoretical model of Brent (1997) contained a procedure to deal with partial utterances and to start the segmentation process before the entire utterance is perceived. Integrating an implementation of this procedure in the STROLL model increases the feasibility of its algorithmic level representation.

Segmentation cues

The STROLL model presented in this thesis applies two types of segmentation cue, familiar forms and phonotactics through transitional probabilities. Section 2.1 has presented many more cues that can be applied to signal the locations of word boundaries. A segmentation model with more cues integrated is more representative of the infant word segmentation process. The same applies to the factors that are relevant in the infant word learning process. Mattys et al. (2005) discussed the integration of multiple segmentation cues in the adult speech perception process. They discussed the developmental trajectory of the infant segmentation process as well (see section 2.1). This trajectory defines three stages. Two sublexical stages – the metrical stage and segmental stage – and the lexical word recognition stage. The segmental stage requires more detailed information, segments must be perceived individually while a more coarse-grained perception is sufficient to perceive stress patterns. However, in order to learn word forms the detailed segmental information is necessary. This explains the infant trajectory of first relying on metrical information (Jusczyk et al., 1999b), followed by segmental information (Morgan and Saffran, 1995), with an increasing role for familiar word forms (Mattys et al., 2005).

The addition of the prosodic cue to the STROLL model was already discussed shortly in this chapter. This would be simple to integrate. Jusczyk et al. (1999b) found that infants are able to apply the dominant English trochaic stress pattern in a segmentation task. This could be integrated in the STROLL model by increasing the probability of a novel form that follows this stress pattern. Consequently, the form will be segmented and acquired faster. According to experimental research, the prosodic cue should be available before the phonotactic n-gram cue. One relevant obstacle in the computational segmentation research is the availability of infant-directed speech corpora. As discussed, an ideal corpus would contain representative infant-directed speech that is phonetically transcribed keeping the pronunciation variation and prosodic information available. As long as there is a lack of representative input corpora containing all the information, the creation of computational models sensitive to this information is limited. The inclusion of a co-articulation cue would be more interesting if an input corpus containing detailed phonetic

information is available. Infants are sensitive to this segmental cue (Johnson and Jusczyk, 2001), which distinguishes the pronunciation of segments that occur within words from those occurring across a word boundary. This cue is more saliently present in a corpus containing detailed phonetic information.

Independent of the lack of available information in infant-directed speech corpora, the choice for the phonotactic cue is well-supported. The information of transitional probabilities between phonemes is more robust than co-articulation information, which is more dependent on between-speaker differences. Probabilities of phoneme transitions are a result of the phonotactic grammar, which is a property of the language. Co-articulation is a property of the phonetic pronunciation, which is subject to differences between people. The prosodic cue is similar in robustness to the phonotactic cue, since it is also part of the phonological grammar of the language. Because of the requirement of more detailed perception, the phonotactic cue is available later in the infant's life. However, Davis (2010) has shown that for computational segmentation, not every phonetic feature needs to be accurately specified for the phonotactic cue to be functional. Jusczyk and Aslin (1995) have found that infants at 7.5 months of age are able to distinguish forms differing in only one or two phonetic features of the initial phoneme. This shows that the detailed perception is available to infants in the word segmentation process.

Comparison to other segmentation models

The STROLL model is based on the PHOCUS model (Blanchard et al., 2010), which is inspired by the Bayesian models presented in section 3.5. An important distinction between these models and PHOCUS and STROLL lies in the level of cognitive plausibility in the computational assumptions. Some of the Bayesian models function in batch mode (Brent and Cartwright, 1996; Goldwater et al., 2009), which has already been determined to be an unrealistic representation of the infant segmentation process since it requires the model to access the entire input corpus before it starts to segment. Even the incremental Bayesian models (Brent, 1999; Blanchard and Heinz, 2008) require the entire corpus to be available throughout the segmentation process, since the optimal segmentation of a single utterance depends on the optimal segmentation of the entire corpus. PHOCUS and STROLL only require the availability of a lexicon and the n-gram probability distribution, making more plausible assumptions on the cognitive requirements.

The lexical and segmentation results produced by the STROLL model are lower than the scores of the PHOCUS model. The baseline version of STROLL without the frequency threshold or the peak filter is outperformed by the PHOCUS model. Table 14 gives an overview of the performance of the two models in baseline using bigrams, which encompasses no restrictions for STROLL and the optional universal phonotactic constraint activated for PHOCUS. The token and type results are comparable, STROLL even performs slightly better when looking at the F0 scores. However, the boundary scores produced by the PHOCUS model are far better than the ones resulting from the STROLL model. One of the largest

distinctions between the two models is at the implementation of the n-gram cue, which mostly affects the boundary results. Further research is necessary to determine whether the source of calculating the n-gram distribution is the cause of the discrepancy in the boundary results between the two models.

Model	WP	WR	WF	BP	BR	BF	LP	LR	LF
PHOCUS-2s	75.2%	64.2%	69.3%	93.7%	74.5%	83.0%	43.0%	63.7%	51.4%
STROLL	72.7%	66.7%	69.6%	60.3%	55.1%	57.5%	47.3%	57.8%	52.1%

Table 14: Results PHOCUS and STROLL. P is precision, R is recall and F is F0. W stands for word (token), B for boundary and L for lexicon (type).

6.4 A combined model of word segmentation and word learning

The research fields of infant word segmentation and infant word learning are narrowly connected. Word forms need to be segmented before they can be learned, and acquired forms are applied as a cue to locate word boundaries in the segmentation process. Experimental studies in this field aim to map the abilities infants have at their disposal to segment word forms and to find the relevant factors that determine which forms are learned by an infant. The computational research is aimed at modeling the usability of the segmentation cues that infants are able to use and simulate the learning process. Despite the connectedness of the two processes, most segmentation models ignore the aspects that are relevant in the infant word learning process. Word learning is reduced to the storage of forms in these models. The model presented in this thesis has included two factors of the infant word learning process in the word segmentation task.

It has shown that a frequency threshold which requires forms to be segmented a set number of times before they are added to the lexicon helps in improving the quality of the forms in the lexicon. Other lexical characteristics that enable faster word learning were already present in the model. Word length and neighborhood density were a part of the segmentation strategy, preferring novel forms that conformed to these factors and enabling them to be learned faster. The second restriction on lexical learning was based on the environment in which infants receive speech input. These environments often contain peaks in which certain content words are uttered with an explicitly higher frequency count than average. The model has shown the presence and usefulness of this word learning cue in natural language, and its absence when the order of the utterances in the naturally collected speech are randomized.

The word segmentation and word learning model presented in this thesis has combined two processes that are connected in infant development. Of both processes, only a subset of the relevant cues and factors are modeled. Future research needs to focus on including all the aspects that take part in this stage of infant language development. Combining the two processes is a first step towards a computational model that fully simulates infant word segmentation and word learning.

7 References

Adriaans, Frans and Kager, René. 2010. Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, 62(3):311–331.

Apoussidou, Diana and Kager, René. 2013. UP on StaGe: Growing a lexicon on phonotactic segmentation. *Unpublished manuscript*.

Bernstein-Ratner, Nan. 1987. The phonology of parent–child speech. *Children's language*, 6:159–174.

Bertoncini, Josiane and Mehler, Jacques. 1981. Syllables as units in infant speech perception. *Infant behavior and development*, 4:247–260.

Blanchard, Daniel and Heinz, Jeffrey. 2008. Improving word segmentation by simultaneously learning phonotactics. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 65–72. Association for Computational Linguistics.

Blanchard, Daniel; Heinz, Jeffrey, and Golinkoff, Roberta. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of child language*, 37(3):487.

Booth, Amy E and Waxman, Sandra R. 2009. A horse of a different color: Specifying with precision infants' mappings of novel nouns and adjectives. *Child development*, 80(1):15–22.

Bortfeld, Heather; Morgan, James L; Golinkoff, Roberta Michnick, and Rathbun, Karen. 2005. Mommy and Me: Familiar Names Help Launch Babies Into Speech-Stream Segmentation. *Psychological Science*, 16:298–304.

Brent, Michael R. 1997. Toward a unified model of lexical acquisition and lexical access. *Journal of Psycholinguistic Research*, 26(3):363–375.

Brent, Michael R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1-3):71–105.

Brent, Michael R and Cartwright, Timothy A. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61(1):93–125.

Brent, Michael R and Siskind, Jeffrey Mark. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):B33–B44.

Brown, Roger. 1973. *A first language: The early years*. Cambridge, Massachusetts.

Chomsky, Noam and Halle, Morris. 1965. Some controversial questions in phonological theory. *Journal of linguistics*, 1(2):97–138.

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

Daland, Robert. 2009. *Word segmentation, word recognition, and word learning: a computational model of first language acquisition*. PhD thesis, Northwestern University.

Daland, Robert and Pierrehumbert, Janet B. 2010. Learning Diphone-Based Segmentation. *Cognitive Science*, 35(1):119–155.

Daland, Robert; Hayes, Bruce; White, James; Garellek, Marc; Davis, Andrea, and Norrmann, Ingrid. 2011. Explaining sonority projection effects. *Phonology*, 28(2):197–234.

Davis, Andréa. 2010. Word Segmentation: The Role of Contrast. *Unpublished MA thesis*.

Elman, Jeffrey L. 2004. An alternative view of the mental lexicon. *Trends in cognitive sciences*, 8(7): 301–306.

Ernestus, Mirjam and Baayen, R Harald. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, pages 5–38.

Gambell, Timothy and Yang, Charles. 2005. Mechanisms and constraints in word segmentation. *Manuscript, Yale University*, 31.

Goddijn, Simo and Binnenpoorte, Diana. 2003. Assessing manually corrected broad phonetic transcriptions in the Spoken Dutch Corpus. In *Proceedings of ICPhS*, pages 1361–1364.

Goldinger, Stephen D; Luce, Paul A, and Pisoni, David B. 1989. Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, 28(5):501–518.

Goldwater, Sharon; Griffiths, Thomas L, and Johnson, Mark. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Good, Irving J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264.

Goodman, Judith C; Dale, Philip S, and Li, Ping. 2008. Does frequency count? Parental input and the acquisition of vocabulary. *Journal of child language*, 35(3):515.

Graf Estes, Katharine; Evans, Julia L; Alibali, Martha W, and Saffran, Jenny R. 2007. Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, 18(3):254–260.

Huttenlocher, Janellen; Haight, Wendy; Bryk, Anthony; Seltzer, Michael, and Lyons, Thomas. 1991. Early vocabulary growth: Relation to language input and gender. *Developmental psychology*, 27(2): 236.

Johnson, Elizabeth K and Jusczyk, Peter W. 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44(4):548–567.

Jusczyk, Peter W and Aslin, Richard N. 1995. Infants' Detection of the Sound Patterns of Words in Fluent Speech. *Cognitive psychology*, 29(1):1–23.

Jusczyk, Peter W and Hohne, Elizabeth A. 1997. Infants' memory for spoken words. *Science*, 277 (5334):1984–1986.

Jusczyk, Peter W and Luce, Paul A. 1994. Infants' Sensitivity to Phonotactic Patterns in the Native Language. *Journal of Memory and Language*, 33(5):630–645.

Jusczyk, Peter W; Hohne, Elizabeth A, and Bauman, Angela. 1999a. Infants' sensitivity to allophonic cues for word segmentation. *Perception & psychophysics*, 61(8):1465–1476.

Jusczyk, Peter W; Houston, Derek M, and Newsome, Mary. 1999b. The beginnings of word segmentation in English-learning infants. *Cognitive psychology*, 39(3):159–207.

Keating, Pat. 2005. D-prime (signal detection) analysis. Website. URL `http://www.linguistics.ucla.edu/faciliti/facilities/statistics/dprime.htm`. last checked: 04.12.2013.

Kimper, Wendell. 2010. Positivity, Serialism, and Finite Goodness. *Presentation at the 18th Manchester Phonology Meeting*.

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

Lalonde, Chris E and Werker, Janet F. 1995. Cognitive influences on cross-language speech perception in infancy. *Infant Behavior and Development*, 18(4):459–475.

Legendre, Géraldine; Miyata, Yoshiro, and Smolensky, Paul. 1990. *Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations*. University of Colorado, Boulder, Department of Computer Science.

Lignos, Constantine. 2012. Infant Word Segmentation: An Incremental, Integrated Model. *Cascadilla Proceedings Project*.

Luce, Paul A and Pisoni, David B. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, 19(1):1.

MacWhinney, Brian and Snow, C. 1985. The child language data exchange system. *Journal of Child Language*, 12:271–296.

Marr, David. 1982. *Vision: A computational approach*. Freeman & Co., San Francisco.

Mattys, Sven L and Jusczyk, Peter W. 2000. Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78(2):91–121.

Mattys, Sven L; Jusczyk, Peter W; Luce, Paul A, and Morgan, James L. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive psychology*, 38(4):465–494.

Mattys, Sven L; White, Laurence, and Melhorn, James F. 2005. Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology: General*, 134(4): 477.

McClelland, James L and Elman, Jeffrey L. 1986. The TRACE model of speech perception. *Cognitive psychology*, 18(1):1–86.

Morgan, James L and Saffran, Jenny R. 1995. Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child development*, 66(4):911–936.

Nazzi, Thierry; Bertoncini, Josiane, and Mehler, Jacques. 1998. Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, 24(3):756.

Norris, Dennis. 1994. Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52 (3):189–234.

Norris, Dennis and McQueen, James M. 2008. Shortlist B: a Bayesian model of continuous speech recognition. *Psychological review*, 115(2):357.

Norris, Dennis; McQueen, James M, and Cutler, Anne. 1995. Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(5): 1209–1228.

Norris, Dennis; McQueen, James M; Cutler, Anne, and Butterfield, Sally. 1997. The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, 34(3):191–243.

Oostdijk, Nelleke H. 1999. Building a corpus of spoken dutch. In Monachesi, Paola, editor, *Computational Linguistics in the Netherlands 1999. Selected papers from the tenth CLIN meeting*, pages 147–158.

Pearl, Lisa; Goldwater, Sharon, and Steyvers, Mark. 2010. Online learning mechanisms for bayesian models of word segmentation. *Research on Language and Computation*, 8(2-3):107–132.

Peterson, William W T G; Birdsall, T, and Fox, W. 1954. The theory of signal detectability. *Information Theory, IRE Professional Group on*, 4(4):171–212.

Pollack, L and Norman, D A. 1964. Non-parametric analysis of recognition experiments. *Psychonomic Science*, 1:125–126.

Prince, Allan and Smolenksy, Paul. 1993. *Optimality Theory: Constraint interaction in generative grammar*. New Brunswick, NJ: Rutgers University Center for Cognitive Science, Rutgers University.

Rissanen, Jorma. 1989. *Stochastic complexity in statistical inquiry theory*. World Scientific Publishing Co., Inc.

Roy, Brandon C; Frank, Michael C, and Roy, Deb. 2009. Exploring word learning in a high-density longitudinal corpus.

Roy, Deb. 2009. New horizons in the study of child language acquisition.

Saffran, Jenny R; Aslin, Richard N, and Newport, Elissa L. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

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

Shi, Rushen; Werker, Janet F, and Morgan, James L. 1999. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72(2):B11–B21.

Shi, Rushen; Werker, Janet F, and Cutler, Anne. 2006. Recognition and representation of function words in English-learning infants. *Infancy*, 10(2):187–198.

Storkel, Holly L. 2004. Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, 25(02):201–221.

Swingley, Daniel. 2009. Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536):3617–3632.

Thiessen, Erik D and Saffran, Jenny R. 2003. When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental psychology*, 39(4):706.

Van de Weijer, Joost C. 1998. *Language input for word discovery*. PhD thesis, Max Planck Institute for Psycholinguistics, Nijmegen.

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

Waxman, Sandra R and Booth, Amy E. 2001. Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive psychology*, 43(3):217–242.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort.* addison-wesley press.