

UNIVERSITEIT UTRECHT

Word Segmentation: The Role of Contrast

by

Andréa K Davis

3303918

A thesis submitted in partial fulfillment for the
degree of Master of Arts

in the

Linguistics Department

UiL-OTS

Advisors Rene Kager and Kie Zuraw

June 2010

Declaration of Authorship

I, Andréa K Davis, declare that this thesis titled, ‘Word Segmentation: The Role of Contrast’ and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UNIVERSITEIT UTRECHT

Abstract

Linguistics Department

UiL-OTS

Advisors Rene Kager and Kie Zuraw

Master of Arts

by [Andréa K Davis](#)

3303918

Phonotactics are known to be a potential cue for word boundaries ([Mattys & Jusczyk, 2001](#); [Mattys et al., 1999](#)); however, to learn and make use of phonotactics in this way, an infant must both perceive native language phones correctly most of the time, and also categorize them in an adult-like way. The role of perception in phonology has been a focus of study in both diachronic and synchronic phonology ([Ohala, 1981](#); [Steriade, 2001](#)). That ease of perception should be a driving force for the organization of phonological grammar is nothing new. However, its role in word segmentation has not been examined in detail. Given that phonotactic cues may be important for word segmentation, it is possible that failure to perceive certain contrasts could be devastating to the learner, in which case phonotactic cues would hardly be practical as reliable cues for word boundaries. Alternatively, it could be that those contrasts which are least perceptible cause less damage to the learner's ability to reliably locate word boundaries. I test this hypothesis on a supervised computational learning model, a modified version of DiBS [Daland \(2009\)](#), altering the input to take away specific contrasts. As a case study, I then ran input neutralized for place of articulation and input neutralized for voicing through the unsupervised learner StaGe [Adriaans & Kager \(2010\)](#). From these modeling tests, I conclude that a) a rich inventory of phones (ie, types) is not as crucial to segmentation as one would think; b) of contrasts tested, place of articulation is the least useful for phonotactic constraints as cues for word boundaries in Dutch, and c) there is a correlation between perceptual salience of a contrast and the usefulness of that contrast in cueing word boundaries.

Acknowledgements

I could not have written this thesis without the support, advice, and inspiration given by my teachers, advisors, and friends. A thousand thanks to my advisors, Rene Kager and Kie Zuraw. Rene, thank you for all your advice and encouragement as I first set up this project, and for your comments later on; your suggestions were invaluable. Thank you, Kie, for your advice and patient reading as this thesis developed; your guidance and recommendations were likewise beyond measure.

Many thanks also to Frans Adriaans, for sharing StaGe with me, and answering all my many questions. Also, a big thank you to the linguistics departments at Universiteit Utrecht and at UCLA, with especial thanks to Megha Sundara at UCLA, for your comments, and to Robert Daland, for both your comments and for sharing DiBs with me - couldn't have done it without you!

Thank you to my friends in Utrecht, who made my time there joyful as well as productive: Liquan Liu, Tim Schoof, Coppe van Urk, Joe Wolfson.

Thank you to Denia Djokic, Kim Pham, Nicholas Hosein, for everything. I would not be the person I am without you.

Finally, thank you to my parents, Rob and Karen Davis, for all your love and support, and for thinking that, when I told you I wanted to be a linguist, it was a good idea. I dedicate this thesis to you.

Contents

Declaration of Authorship	i
Abstract	iii
Acknowledgements	iv
List of Tables	vii
1 Introduction	1
1.1 Computational Models of Word Segmentation	2
1.2 Assumptions made about the infant learner	4
1.3 Statement of research questions and hypotheses	6
1.3.1 Set-up	7
2 Experiment 1: Are some contrasts more important than others?	8
2.1 Experiment 1a: Place, Voice, Nasal, and Continuant	9
2.2 Experiment 1b: Place of Articulation	14
2.3 Summary	16
3 Experiment 2: A Case Study: StaGe	17
3.1 The StaGe Model	17
3.1.1 Statistical Component	18
3.1.2 Generalization Component	20
3.1.3 Forming and generalizing constraints	20
3.1.4 Ranking constraints and evaluating word boundaries	21
3.1.5 Summary	22
3.2 Experiment 2a: Place	23
3.2.1 Why some constraints and biphones do not matter as much as others	25
3.3 Experiment 2b: Voice	29
3.3.1 Exploring the segmentation decisions	30
3.4 Comparing the StaGe and O/E models	36
4 General Discussion and Conclusions	38
4.1 Frequency Distributions in Dutch	39
4.2 Perception as a driving force of phonology	46

4.3 Conclusion and Future Directions	47
A Terminology	49
B Differently Segmented Biphones	51
Bibliography	60

List of Tables

2.1	Conditions Tested	10
2.2	Results for Experiment 1a	11
2.3	Undersegmentation Errors for the Nasal Condition	12
2.4	Oversegmentation Errors for the Nasal Condition	13
2.5	Results for total place of articulation neutralization	13
2.6	Conditions Tested	14
2.7	Results for Experiment 1b	14
2.8	Oversegmentation Errors for the Stops Place Condition	15
3.1	Examples of biphones and O/E values	18
3.2	Results with full Contrast	22
3.3	Neutralizations in the input	23
3.4	Results with place neutralization	24
3.5	Results with full Contrast	25
3.6	Homo-organic nasal+stop cluster frequency across a word boundary	26
3.7	Non-homo-organic nasal+stop cluster frequency across a word boundary	27
3.8	Plosive and nasal geminate frequencies	28
3.9	Neutralizations in the input	29
3.10	Results with voice neutralization	30
3.11	Results with full contrast	30
3.12	Obstruent biphones agreeing in voice	31
3.13	Frequent obstruent biphones and their segmentation rate	33
3.14	Differently segmented vowel+obstruent biphones	35
3.15	Differently segmented vowel+obstruent biphones	36
4.1	Frequent Dutch Words	44
4.2	Phone Type Frequencies	46

For my parents

Chapter 1

Introduction

Phonology may be described as the study of phonetic contrasts. A phonological grammar describes both what is an allowable contrast in a given language, as well as the environment in which that contrast is permitted. However, attention is not often given to what a phonological grammar is useful for, from the perspective of the language user.

One role for phonology is breaking up speech into smaller, more manageable chunks: speech segmentation. Natural speech does not normally contain audible pauses between words (Cole & Jakimik, 1980), nor any other non-language-specific cue (Cutler & Carter, 1987). As a result, one of the first tasks of language learning is to break up continuous speech into words.

Experimental work has found that infants may use a variety of cues for finding word boundaries, including: probabilistic phonotactic cues (Mattys & Jusczyk, 2001; Mattys et al., 1999), prosodic stress (Cutler, 1994, 1990; Cutler & Norris, 1988), distributional cues (Saffran et al., 1996), and phonetic detail (allophonic cues) (Cho et al., 2007; Fougeron & Keating, 1997; Johnson & Jusczyk, 2001). Phonotactics has also been found to influence word processing in adults: phoneme perception (Coetzee, 2005), wordlikeness judgments (Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997), as cues for word boundaries (McQueen, 1998), and, to some extent, as cues for morpheme boundaries (Hay & Baayen, 2004). Once learned, then, phonotactics has potential as a way of organizing and breaking up chunks of speech.

Learning any of these cues, however, is not a simple problem. In the case of learning distributional cues, infants must keep track of a huge number of co-occurring sounds. In the case of phonotactics, infants must learn what phone clusters are legal within a word, and which are not, without knowing word boundaries. That is, infants must learn phonotactics from continuous speech. While experimental work has shown that infants

are capable of using phonotactics to predict word boundaries, it does not make clear how infants do this. In order to solve this problem, a number of computational models have employed a variety of strategies to learn phonotactics from an unsegmented input, and then applied this knowledge to predict word boundaries in unsegmented speech.

While computational models of word segmentation have considered various strategies for predicting word boundaries, no studies have yet focused on the nature of the input. Given that infants are learning contrasts at the same age as they are beginning to segment speech (Maye et al., 2002), and that knowledge of contrasts may not be as perfect in real life as in laboratory settings, it is conceivable that input corpora for these models over-estimate an infant's ability to keep track of contrasts.

1.1 Computational Models of Word Segmentation

Computational models allow the testing of explicit models of speech segmentation. Complementing experimental research in infant acquisition, the researcher using a computational model is able to ask very complex questions in a direct way, since it is easier both to control the settings, and to see how the results were obtained.

In the past two decades, a number of computational models have been proposed for segmenting words from continuous speech. In this paper, I focus on phonotactic, n-phone-based (ie, n-gram) models, which focus on n-phone patterns as a segmentation cue. Biphones (eg, da) were first shown to be a good segmentation cue by Cairns et al. (1997); however it was not clear how infants could learn which biphones should be segmented.

N-phone (often biphone) models generally use a probability statistic to predict where word boundaries may occur, akin to how infants presumably use distributional information (Thiessen & Saffran, 2003). There are a number of probability statistics which may be employed. The most popular in the psycho-linguistic literature is transitional probability (Aslin et al., 1998; Saffran et al., 1996; Thiessen & Saffran, 2003). Transitional probability is the probability of a phone, y, given a preceding phone, x. It is calculated:

$$TP(x, y) = \frac{P(xy)}{P(x)} \quad (1.1)$$

Mutual information is related to observed/expected; it is simply the log of O/E:

$$MI = \log(O/E) \quad (1.2)$$

All of these probability statistics are obtained by calculating frequencies of particular n-phones from a corpus - the training set. Information thus obtained may be employed directly to predict word boundaries, or it may be used to discover phonotactic constraints. Any model that uses phonotactics as a cue follows a similar pattern: the model must first learn the sequences which are phonotactically legal, and then using this information to predict word boundaries.

In all cases, the focus of these kinds of models has generally been to maximize accuracy of prediction, without assuming that the learner has access to any more information than an infant might be expected to have - hence, to have the best performance given a psycho-linguistically plausible model.

Despite overall similar strategies of segmentation, these models differ in the particulars. Here, I review some of the models developed in the past two decades; for a more complete review of word segmentation models, see [Brent \(1999\)](#) or [Daland \(2009\)](#).

[Brent & Cartwright \(1996\)](#) approach the problem of learning phonotactics by assuming that all legal phonotactic sequences eventually occur at utterance boundaries, which, unlike word boundaries, are marked by pauses. Hence, the learner may assume that sequences which occur at the beginning of utterances also are good as word onsets, while sequences which occur at the end of utterances are also good word-finally. From this, infants assume that sequences which do not occur at utterance boundaries are phonotactically illegal. Using this information, the learner may then posit word boundaries.

The StaGe model ([Adriaans & Kager, 2010](#)), which will be discussed further in section 3.1, uses biphone probabilities to discover phonotactic constraints, which are then employed to predict word boundaries. It uses a different statistic than transitional probability: observed/expected. Observed/expected is simply the ratio of an observed frequency of a biphone to the expected frequency of a biphone:

$$O/E = \frac{\text{countof}xy}{\text{expectedcountof}xy} \quad (1.3)$$

where x and y are phones and xy is a sequence of x followed immediately by y.

The expected count of a biphone is based on the frequency of each phone being preceded or followed by any element:

$$E = f([xY]) * f([Xy]) * n \quad (1.4)$$

where Y is any element following x , X is any element preceding y , and n is the total number of biphones in the corpus.

[Adriaans & Kager](#) find that StaGe, which employs this statistical information to form generalized, phonotactic constraints, similar to Optimality Theory constraints, does better than a model which applies statistical information to the word boundary problem directly.

Another model, DiBS ([Daland, 2009](#)), uses phonotactic probability to predict word boundaries, but uses a different statistic: Maximum Likelihood Decision Threshold (MLDT). For each possible word boundary, it makes a decision based on the probability of a word boundary, $p(\# | xy)$, rather than the probability of a biphone. There are two versions of DiBS, a baseline, supervised learner, and an unsupervised learner. In the former case, computing the probability of a word boundary for biphone xy is easy, since word boundaries are known; hence, the probability of a word boundary is calculated from the frequency of a word boundary for a given biphone. In the latter case, however, [Daland](#) has the model infer the probability of a word boundary based on the probability of a) biphones at utterance boundaries and b) context free probability of a word boundary. He finds that this unsupervised learner performs nearly as well as the baseline (supervised) learner, meaning that it achieves nearly optimal performance.

1.2 Assumptions made about the infant learner

The nature of the input is especially important for computational models, since their validity depends on the input given to the "learner" being very close to what an infant perceives. Input, however, has been somewhat of a background question in modeling. While occasionally, variation - whether in the form of random noise, variation in pronunciations, speech errors, or adult vs. child directed speech - has been examined, it has been examined in view of testing how well a particular model withstands variation in the input.

One such difficulty - the limits of what an infant perceives, and what an infant knows about the contrasts of her native language - has not been examined at all. Infants are learning phones/phonetic contrasts at the same time that they are learning word boundaries. It has been assumed that infants have access to phonetic contrasts by the time they begin to segment speech (7.5 months), or at least by the time they use more complex linguistic information (ie, phonotactics, metrical stress) as cues for word boundaries. This assumption is not without empirical evidence; infants between 6 and 8 months are able to discriminate between phonetic categories, at least in a laboratory

setting (Maye et al., 2002; Jusczyk & Aslin, 1995). Further evidence comes from a learning model which is able to learn many categories directly from the acoustic signal (Lin, 2005). However, in the latter study, neither voice nor place of articulation were learnable from acoustic information alone. Further, studies which show that infants can distinguish between contrasts such as p/t/k or p/b are always in a laboratory setting, normally with very simple speech with few, if any, consonant clusters or reductions.

A timeline of infant acquisition is given below:

2 months	discrimination of speech sounds both native and non-native (Werker & Tees (1983))
4 months	native discrimination of vowel contrasts (Polka & Werker, 1994) discrimination of consonant contrasts both native and non-native
6-9 months	beginning to segment speech sensitivity to distributional cues (Thiessen & Saffran, 2003)
7.5 months	sensitivity to metrical cues (Jusczyk et al., 1999)
9 months	sensitivity to phonotactic cues (Mattys & Jusczyk, 2001; Mattys et al., 1999)

Since learning contrasts depends on accurate perception, it would seem that the perceptual salience of a contrast could influence both a) its learnability and b) accuracy of transmission in phonotactic rules. Infants, at least at first, must depend primarily on acoustic information, as initially they have no lexicon, nor have they learned the sound patterns of their native language; hence, it makes sense if more perceptually salient contrasts are learned earlier, and are retained as being phonotactically important (ie, useful for word segmentation).

Experimental perceptual studies converge on place of articulation being relatively more difficult to perceive. Cutler et al. (2004), in an experiment comparing Dutch and English perception under noise, found that both groups had the most trouble with correctly identifying place of articulation contrast, while for the Dutch listeners, voice, while less difficult than place, was more difficult than identifying fricative or nasal.

The classic study by Miller & Nicely (1955) found similar results, regarding which contrasts were most difficult (for English speakers). Voice and nasality were the least likely to be confused by English speakers when noise was introduced to a speech signal, while place of articulation was the most confusable. Smits, Warner, McQueen, & Cutler (2003) also found that place of articulation was a major source of error, but unlike the other studies found that voice along with place of articulation were the two main sources of error for Dutch listeners listening to (Dutch) syllables embedded in nonsense speech.

Position mattered for confusability, syllable final position being more confusable, particularly for voicing contrast; this latter confusion could be due to the illegality of voice in coda position in Dutch.

Yet another study of English speakers' perception in noise found that within place of articulation, there seems also to be a hierarchy: nasal place of articulation is most confusable, followed by stops, followed by fricatives (Hura, Lindblom, & Diehl, 1992).

While undoubtedly phonology influences perception (Coetzee, 2005; Dupoux et al., 2001), it has also been argued that perception influences phonology (Hura et al., 1992; Kohler, 1990; Mielke, 2003a,b; Ohala, 1981; Steriade, 2001), due to certain contrasts being more perceptually salient than others, with others being more vulnerable to misperception, regardless of language-specific influences. For example, Kohler (1990) observed that nasals and stops were more likely to undergo assimilation than fricatives, and posited that the reason was perceptual; this is in line with the results from Hura et al.. If perception does indeed influence phonology, then it is imaginable that more perceptually salient contrasts are more likely to be retained, and hence make better word boundary cues.

1.3 Statement of research questions and hypotheses

My broad goal in the following experiments was to explore the potential role of contrast in word segmentation. That is, what kinds of contrasts play an important role in phonotactic cues for word segmentation?

Within this broader question, there are two secondary questions.

1. How damaged is the ability of the learner to make accurate predictions if the learner does not correctly perceive or does not correctly categorize native phones?
2. Are all contrasts equally important, or are some more important than others? Is there any correlation between perceptual salience and the importance of a contrast as a cue?

Question 1 relates to the problem of using phonotactics to segment words without having necessarily fully segmented or correctly categorized native phones. If removing contrasts devastates the ability of the learner to make reliable predictions (ie, accuracy, measured with d' , and precision both approach 0; the model segments nearly at chance), then we may assume that infants, if they use phonotactics in segmentation, must perceive with very good accuracy. Further, we must assume that infants are recognizing roughly

adult-like phone categories. Alternatively, if the neutralization of various contrasts in the input does not devastate the model's performance, then one may conclude that a) the model is fairly robust to error in the input and b) it need not be assumed that infants have perfect perception and a fully adult phonological system, in order for them to make use of phonotactics as word boundary cues.

Question 2 extends the broad literature of the influence of perception on phonological grammar to the word segmentation problem. If perceptual salience influences the potential usefulness of certain contrasts for word segmentation (ie, if language is organized such that perceptually salient contrasts are more useful for predicting word boundaries), then it may be expected that removing place contrast will have little effect on the ability of the learner to use phonotactics as word boundary cues, and the more salient manner contrasts, such as nasal and continuant, to be of greater potential benefit.

1.3.1 Set-up

In Experiment 1, I used a supervised learner to test the potential importance of particular contrasts. This will show, given the best possible performance, if particular contrasts are of greater potential usefulness than others.

Experiment 2 is a case study, examining the effects of neutralizing particular contrasts for the model StaGe, an n-gram(biphone)-based threshold learner. This will show the effects of neutralization on an unsupervised learner, and the importance (or unimportance) of contrast for a more realistic learner.

Chapter 2

Experiment 1: Are some contrasts more important than others?

Given that some contrasts are easier to perceive than others, it could be the case that some contrasts are a) easier to learn than others and/or b) easier to use as cues for word boundaries. Further, if perception is a driving force of phonological alternations, this predicts that the most perceptible cues are the least likely to reduce in natural speech. It would then make sense if the cues which were most informative were also the easiest to perceive.

Could it be the case that the contrasts which are important for cueing word boundaries are also the ones which are the easiest to perceive? In order to test this hypothesis, I used a supervised word segmentation model.

Using a supervised learner gives the optimal performance for the set of possible biphones, which depends on the contrasts available to the model. For example, if the input set doesn't include place of articulation, biphones *pa* vs. *ta* vs. *ka* would not exist, but would be collapsed as a single biphone: *Ta*. Hence, the fewer contrasts available, the fewer biphones in the set of possible biphones.

Based on perceptual experiments, the contrasts that are most difficult to perceive is place of articulation in stops and nasals and voice (at least in Dutch), which would predict that these contrasts would also be less important. These contrasts are also not classifiable by acoustic information alone (Lin, 2005), suggesting that they are inherently more difficult to learn and perceive than manner contrasts: fricative (ie, continuant) and

nasal contrasts are found to be easiest to perceive, and therefore perhaps most exploitable for word boundary cues.

2.1 Experiment 1a: Place, Voice, Nasal, and Continuant

Method

The supervised model I used is a modified version of that used by [Daland \(2009\)](#). It is a biphone-based phonotactic learner, which works as follows:

1. It takes a segmented corpus of utterances as input.
2. It calculates total frequency of a biphone, frequency of that biphone within a word (word internal frequency), and frequency of that biphone at a word boundary (word boundary frequency).
3. By definition, the best performance will be to segment biphones which occur at a word boundary $> 50\%$ of the time.
4. Segmentation decisions are made for each biphone, such that a biphone occurring at a word boundary $> 50\%$ of the time is segmented, and a biphone occurring at a word boundary $\leq 50\%$ of the time will not be segmented.
5. The model reports the true positives, false positives, true negatives, and false negatives, the recall (hit rate), the precision, and the segmentation decisions made for each biphone.

I tested a number of possible neutralizations, including : place of articulation neutralized for stops and nasals, voicing neutralized for obstruents, nasal neutralized for nasal-s/stops, and continuant neutralized for obstruents. A summary of the various conditions is given in the table below:

In each case, the symbol chosen for the neutralized version of a class (eg, p,t,k) does not matter, so long as it is consistent. The one case where it did matter was for continuant neutralization, where the fricative version was chosen, so that the place contrasts s vs. \int and z vs. ζ , which do not have equivalents in the stop class, were not neutralized.

Below is an example of the unaltered and altered corpus:

unaltered:

Condition	Neutralizations
Control	no neutralizations
Place	m,n,ŋ > n p,t,k > t b,d,g > d
Voice	b > p, d > t, g > k v > f, z > s, ʒ > ʃ, ʁ > x
Nasal	m > b, n > d, ŋ > g
Continuant	p > f, t > s, k > x b > v, d > z, g > ʁ

TABLE 2.1: Conditions Tested

@tx@ramt@statOp@mvirpot@ndij@swArt@dIN@did@rOnd@rLtstek@

neutralized for place in stops and nasals:

@tx@rant@statOt@nvirtot@ndij@swArt@dIn@did@rOnd@rLtstet@

Corpus

In all cases, I used the Dutch Spoken Corpus (Corpus Gesproken Nederlands, from now on CGN), a corpus of natural speech transcribed with natural variation in pronunciation (Oostdijk, 1999). Using a corpus of natural speech is non-trivial to the question of what kinds of contrasts may be most important: Daland (2009) found that testing with a corpus containing natural variation in pronunciation made a difference in performance for various word segmentation models, including DiBS (Daland, 2009), Goldwater 2006, Fleck 2008. Overall, the phonotactically-based models are more robust to variation in pronunciation. He posits that the reason for this is that, in natural speech, the cues to word boundaries which are most important are the least likely to be reduced.

Results and Discussion

Hit and false alarm rates are reported for each condition in Table 2.2. These are calculated :

$$H = \frac{TruePositives}{(TruePositives + FalseNegatives)} \quad (2.1)$$

$$F = \frac{FalsePositives}{(FalsePositives + TrueNegatives)} \quad (2.2)$$

d' is calculated from the mean scores of the hit and false alarm rates. Because the z score of hit and false alarm rates is used to calculate d' , it is a bias-free measure of performance. d' is calculated:

$$d' = z(H) - z(F) \quad (2.3)$$

Finally, precision is also reported. Precision is defined as the number of true positives divided by the sum of true positives and false positives:

$$p = \frac{TruePositives}{(TruePositives + FalsePositives)} \quad (2.4)$$

In Table 2.2, the Control is when the corpus is unaltered (no neutralization), Place is where place contrasts are neutralized, Voice is where voice contrasts are neutralized, Nasal is where nasal contrasts are neutralized, and Continuant is where continuant contrasts are neutralized.

Condition	Control	Place	Voice	Nasal	Continuant
False Positives	123501	102553	101920	114475	123025
True Positives	355613	312260	306327	327946	331804
False Negatives	226741	270094	276027	254408	250550
True Negatives	1468824	1489772	1490405	1477850	1469300
Hit Rate	0.6106	0.5362	0.5260	0.5631	0.5698
False Alarm Rate	0.0776	0.0644	0.0640	0.0719	0.0773
Precision	0.7422	0.7528	0.7503	0.7413	0.7295
d'	1.6840	1.6550	1.6300	1.6270	1.5810

TABLE 2.2: Results for Experiment 1a

When measured in d' , there is less of a change between full contrast and the place condition than is the case for any other condition. In the place condition, d' is higher than in the other conditions, though lower than in the control. There is less difference between the other conditions, although continuant neutralization results in greater change to both d' and to precision. In any measure, then, it seems that continuant neutralization has the biggest negative effect on performance.

Precision is actually greater for place neutralization than for the control; essentially, this means that fewer mistakes are made (because there are fewer false positives) when place is neutralized. When performance is measured in precision, neutralizing place of

articulation could even be helpful to the infant. The usefulness of place of articulation could in fact be in increasing the hit rate; this in fact follows what would be the expected pattern for an infant - under-segmenting early on, but with few false positives. While there has been little study on whether infants over-segment or under-segment, what study there has been suggests that young children under-segment (Peters, 1983).

Voice neutralization also results in an increase in precision as compared with the control, although not quite as much as in the case of place neutralization. d' is lower than either the control or the place neutralization condition, due to a decrease in hit rate.

Precision falls for both nasal and continuant neutralization; d' is also lower. Hit rate is higher, but without nasal or without continuant information, more errors are made; the model is too general.

The differences in performance between conditions are due for the most part to a few very frequent biphones. For example, in the case of nasal neutralization, hit rate falls when some high frequency biphones of the form V+voiced stop are not segmented. Table 2.3 shows the most frequent (> 200) under-segmentation errors made when the contrasts b-m, d-n, and g-ŋ are neutralized; because em, am, an, and en are more frequently word internal than at a word boundary, eB, aB, aD, and eD (where B and D are the neutralized b-m and d-n respectively) are better not segmented in the nasal neutralization condition. However, while this is the best decision given the contrasts available, it means that eb, ab, ad, and ed will be under-segmented.

Biphone	Internal Frequency	Boundary Frequency	Total Frequency	% Word Boundary
eb	208	242	450	0.5377778
ab	342	507	849	0.5971731
ad	1818	3676	5494	0.6690936
əd	6343	14420	20763	0.6945046

TABLE 2.3: Undersegmentation Errors for the Nasal Condition

When hit rate falls, false alarm rate inevitably falls as well. However, because many biphones in the nasal neutralization condition are over-segmented, it does not fall enough to either maintain as high of precision or d' as in the control; both are lower. Over-segmented biphones are generally of the form C+D, where D is a phone neutralized for nasality. Hence, oversegmentation errors occur in the nasal condition when the nasal counterpart of a C+D biphone would be segmented, but the stop counterpart would not. Oversegmented biphones of frequency greater than 200 are given in Table 2.4.

One criticism that could be made of this conclusion is that for place of articulation, only nasals and stops were neutralized. Place of articulation was preserved in non-nasal sonorants and in fricatives. The reasoning was that perceptual experiments found that nasal

Biphone	Internal Frequency	Boundary Frequency	Total Frequency	% Word Boundary
rd	4923	4435	9358	0.4739261
ld	2289	1675	3964	0.4225530
yd	291	194	485	0.4000000
dz	135	117	252	0.4642857
dI	3323	2155	5478	0.3933917
ab	342	507	849	0.5971731
ad	1818	3676	5494	0.6690936
æd	6343	14420	20763	0.6945046

TABLE 2.4: Oversegmentation Errors for the Nasal Condition

and stop place of articulation was more often misperceived than other contrasts. A second reason for doing this was that, while place may not be very salient perceptually, it is contrastive for nearly all natural classes. This is not true of other contrasts. For example, while voicing may be said to be contrastive between obstruents, it is not contrastive for sonorants. Likewise, continuant is only contrastive for obstruents, since non-continuant sonorants have other features which distinguish them from other sonorants (ie, nasal or lateral). The reasons for this are probably articulatory, not perceptual.

Thus, neutralizing place of articulation for all classes would neutralize far more contrast types than does neutralizing voice or manner contrasts like continuant or nasal, and indeed, when fricative place of articulation is also neutralized, performance is much worse:

False Positives	141139
True Positives	334837
False Negatives	247517
True Negatives	1451186
Hit Rate	0.5750
False Alarm Rate	0.0886
Precision	0.7035
d'	1.5430

TABLE 2.5: Results for total place of articulation neutralization

For a more fair comparison, in which roughly the same number of types are neutralized, various classes may be neutralized for place of articulation and compared. [Hura et al. \(1992\)](#) found that not all classes are equally likely to have place of articulation misperception; hence, it seems some places of articulation are more perceptually salient than others. Nasal place is more confusable than stop place, which is in turn more confusable than fricative place. If more perceptually salient contrasts are more useful for word segmentation, and the perceptual hierarchy of nasal place > stop place > fricative place is indeed true, it is expected that neutralizing nasal place of articulation will affect

performance less than neutralizing stop place, which will in turn affect performance less than neutralizing fricative place. This is tested in the next section.

2.2 Experiment 1b: Place of Articulation

Method

Experiment 1b followed the same method as described in Experiment 1a, except that the following neutralizations were made to the corpus:

Condition	Neutralizations
Nasal Place	m,n,ŋ > n
Stop Place	p,t,k > t b,d,g > d
Fricative Place	f,s,ʃ,x > s v,z,ʒ,ʝ > z

TABLE 2.6: Conditions Tested

Results and Discussion

Condition	Nasal Place	Stop Place	Fricative Place
False Positive	108347	120737	154868
True Positives	329831	340203	370010
False Negatives	252523	242151	212344
True Negatives	1483978	1471588	1437457
Hit Rate	0.5664	0.5842	0.6354
False Alarm Rate	0.0680	0.0758	0.0973
Precision	0.7527	0.7381	0.7049
d'	1.6520	1.6070	1.6400

TABLE 2.7: Results for Experiment 1b

Again, the same trend is observed for precision: the more perceptually salient the contrast, the lower the precision when that contrast is neutralized. This does not entirely hold true for d': while place neutralization in stops results in a lower d' than does nasal place neutralization, d' is actually higher when place is neutralized in fricatives. This comes at the cost of a very high false alarm rate, however; in fact, false alarm rate is highest of any condition in either Experiment 1a or 1b, when place is neutralized in fricatives. Overall, within this comparison, the trend seems to hold: less perceptually salient contrasts damage word segmentation performance less when neutralized.

However, note that precision and d' is lower for place neutralization in stops alone than it is for place neutralization in stops and nasals. This is because when nasals are neutralized along with stops, the resulting generalization (for example, əm , ən , $\text{əŋ} = \text{əN}$ when place is neutralized) prevents some oversegmentation errors. For example, the biphone yn is not segmented when nasals are neutralized, but when they are not, yn happens to cross the 50% mark; that is, it is at a word boundary 50.1845% of the time. Hence, it is segmented, resulting in more hits, but almost as many misses. When hits and misses are nearly equal, precision falls, and false alarm rate rises. This will occur whenever probability of a word boundary approaches 0.5. When probability of a word boundary is slightly above 0.5, it results in many oversegmentation errors.

Below are given the oversegmentation errors for the Stops Place condition, for biphones of frequency > 200 :

Biphone	Internal Frequency	Boundary Frequency	Total Frequency	% Word Boundary
im	1013	1502	2515	0.5972167
əm	8201	8649	16850	0.5132938
yn	270	272	542	0.501845
nɛ	1659	2392	4051	0.5904715
nə	7766	11044	18810	0.5871345

TABLE 2.8: Oversegmentation Errors for the Stops Place Condition

In contrast, there is only one biphone which is segmented when place is neutralized in both nasals and stops, but not segmented when only stops are neutralized: vn . Further, this does not result in an oversegmentation error, since the probability of a word boundary is 0.8645; additionally, vn is not especially frequent, as it occurs only 310 times in total.

Hence, while at first it is surprising that the model does worse when only one contrast is neutralized than when two are, when the former contrast is a subset of the latter, it becomes clear on inspection; certain very frequent biphones cross the 0.5 boundary and are segmented, when nasal place is no longer neutralized. It would seem that neutralizing nasal actually helps the learner, since it allows generalization across nasals, but without being overly-damaged by over-generalization.

When fricative place is neutralized, d' is higher than for when stops are neutralized, though false alarm rate is very high (the highest of any condition in either Experiment 1a or 1b). Accordingly, precision is very low. This makes it somewhat difficult to compare with the other conditions. However, on the assumption that undersegmentation is preferable to a high false alarm rate, neutralizing place of articulation in fricatives is very damaging to the learner, arguably more so than voice, nasal, or continuant neutralization, despite having a higher d' . Place neutralization in fricatives, then, may

be an exception to the general trend of less perceptually salient contrasts being less crucial to perceive correctly in learning phonotactic cues for word boundaries.

2.3 Summary

Here, I have approached the question of whether there is a correlation between the perceptual salience of a contrast and the potential usefulness of that contrast in phonotactic cues for word boundaries by testing on a supervised model; hence, how useful a particular contrast could be under optimal conditions. It does indeed seem to be the case that perceptual salience broadly correlates with usefulness in word segmentation: place of articulation, which is less perceptually salient both cross-linguistically and independently of context, seems to be less helpful for segmentation than other contrasts in Dutch. This extends to voice contrasts, which at least in Dutch seem to be less perceptually salient; neutralizing voice, while having a greater effect than neutralizing place of articulation, has less effect than neutralizing manner contrasts continuant and nasal. Further, it looks as though place of articulation in nasals, which is less perceptually salient than place of articulation in stops or in fricatives, is likewise less helpful for word segmentation.

Additionally, the contrasts that are more perceptually salient - continuant and nasal - are more useful than either place or voice for biphone (phonotactic) word boundary cues. Precision and *d* were both lower when these contrasts were removed from the input. While the relation between perceptual salience and usefulness in word boundary cues is less clear in Experiment 1b, overall the relation holds, with fricative place of articulation being an exception to the general trend.

Thus, all in all these tests are in support of the notion that there is a correlation between perceptual salience of a contrast and its usefulness in word segmentation. However, a supervised model is not a psycholinguistically plausible model, as an infant cannot possibly already know word boundaries *while* learning them. Thus, testing on a supervised model answers the question: what is the best the learner can hope to do, given limitations in the input, but for a more realistic learner, I now turn to a case study: the unsupervised learner, StaGe.

Chapter 3

Experiment 2: A Case Study: StaGe

In Experiment 1, it was found that there is a correlation between the neutralization of a perceptually salient contrast and the optimal performance of a learner. In other words, it was shown that the perceptual salience of a contrast affects the performance of a *supervised* learner, when that contrast is neutralized.

Testing a supervised learner first makes sense, since it makes the results non-specific to a particular model, and answers the question as to the best a learner can do, given limitations of the input. However, a more realistic situation is that of the unsupervised learner, and is interesting as a case study. A number of unsupervised learning computational segmentation models have been developed in the last decade ([Adriaans & Kager, 2010](#); [Daland & Pierrehumbert, submitted](#); [Goldwater et al., 2009](#); [Brent, 1999](#); [Blanchard & Heinz, 2008](#); [Goldwater, 2007](#); [Venkataraman, 2001](#)). Here, I run neutralized input through the StaGe Model as a case study, to observe the consequences of neutralizing contrasts in an unsupervised model.

3.1 The StaGe Model

The StaGe Model was developed by [Adriaans & Kager \(2010\)](#) to show that generalization of features to form general constraints is superior to using only specific constraints. Like DiBS, it is a biphone-based learner, threshold learner. Unlike DiBS, there are two components to this model: a statistical component and a generalization component. I describe the model in greater detail in this section.

3.1.1 Statistical Component

The statistical component of StaGe is relatively simple. It is a biphone model, which computes the frequency distributions of every biphone discovered in the training corpus, where a biphone is a pair of phones, xy (x and y being phones). Most experimental work has considered the transitional probabilities - henceforth TP - of biphones (Saffran et al., 1996; Aslin & Newport, 2004). StaGe uses observed/expected, which has also been used in studies of phonotactics (Frisch et al., 2004).

To give an idea of what this looks like when computed over a corpus, a few biphones and their observed, expected, and O/E values are given below:

biphone	observed	expected	O/E
he	7243	958.5529	7.5562
di	10355	2735.1631	3.7859
fa	209	115.8549	1.8040
xo	1234	1442.2410	0.8556
ŋr	327	877.6323	0.3726

TABLE 3.1: Examples of biphones and O/E values

Both the O/E model and the StaGe model are threshold-based segmentation models. Given a high or low enough O/E value for a biphone, the biphone will either always be segmented, or always be left contiguous. This differs from a trough-based model, where a biphone is only segmented depending on the probability of surrounding biphones. Given a sequence $wxyz$, a trough-based model would place a word boundary between the two phones with the lowest transitional probability, such that if wx has a high TP, xy has a low TP, and yz has a high TP, a word boundary will be placed between x and y .

While threshold models have their disadvantages, since there are very few biphones which are segmented 100 percent of the time, they have the advantage over trough-based models in being able to segment unigram words. In a trough based model, a word is minimally a biphone.

While well documented in acquisition literature, one may reasonably ask why it is expected that a probability statistic like O/E or TP should cue a word boundary, or a non-boundary, as the case may be. Below is a graph plotting the percent word boundary against the log O/E ratio, for biphones of frequency 5,000 and above:

occurs as often as expected, and thus has neutral representation, anything with an O/E value less than 1.0 is under-represented and anything with an O/E value greater than 1.0 is over-represented. As such, a word boundary is placed for biphones with O/E values less than 1.0. Anything with an O/E value of exactly 1.0 will be segmented or not segmented at random. Everything else will remain intact. Going back to Table 3.1, the O/E learner would segment *xo* and *nr*, since these have O/E values less than 1.0. In the corpus segmented by this learner, we see that this biphone is indeed segmented:

17075¹ dAtIzb@lAN rKk@mIsx inwEl dAn @ t dAn @ tfA kwAtj@xeft

3.1.2 Generalization Component

StaGe does not apply the statistical information acquired from computing O/E values directly to word segmentation. Rather, constraints are formed in a modified version of Optimality Theory. Strict domination holds, as in traditional OT, but ranking is unsupervised, and rather than the classic markedness and faithfulness constraints, StaGe makes use of markedness and contiguity constraints. Faithfulness constraints as such are not available to the learner, since the learner can have no knowledge of underlying forms. Markedness constraints induce segmentation; contiguity constraints prohibit segmentation. Thus, these two kinds of constraints are in competition with one another, as markedness and faithfulness constraints are in classic OT.

3.1.3 Forming and generalizing constraints

Constraints are formed in a threshold-based way. Biphones that have an O/E value less than 0.5 are turned into markedness constraints (segment the biphone). Biphones that have an O/E value greater than 2.0 are turned into contiguity constraints (keep the biphone intact). The threshold values are arbitrary, but subsequent testing of the model has shown that the exact values, provided they do not approach 1.0, do not effect the greater accuracy of StaGe as compared with the O/E learner (see [Adriaans & Kager \(2010\)](#) for details).

Returning once more to Table 3.1 of biphones and their O/E values, contiguity constraints would be formed for *he* and *di*, but not for *ja*, since, while over-represented (O/E value > 1.0), it is not above 2.0, the threshold value for StaGe. Likewise, a segmentation constraint will be formed for *nr* but not for *xo*. Whether or not these particular biphones are kept intact or segmented depends, however, not only on whether or not constraints exist, but also on whether or not they outrank conflicting constraints.

¹This number refers to the utterance; it is useful for finding a particular utterance in the corpus

This, of course, leaves a lot of biphones unaccounted for; many biphones have O/E values that fall in between 0.5 and 2.0 - namely, *fa* and *xo*. These biphones with neutral O/E values tend to be accounted for, however, once generalization takes place.

Generalization works through single-feature abstraction, using the notion of "constraint neighbors" first discussed in Hayes (1999). Constraints are neighbors when they differ in one feature only. For Dutch, StaGe uses the following features: syllabic, consonantal, approximant, sonorant, continuant, nasal, voice, place, anterior, and lateral for consonants, and high, low, back, round, long, tense, and nasalized for vowels. Constraints $*[v][d]$ and $*[v][t]$, which have O/E values of 0.4463 and 0.0200 respectively, are neighbors, since they differ in a single feature: voice. They will form a more general constraint, $*[v][td]$. In turn, this can become more general; constraint $*[z][t]$, with O/E value 0.0408, differs from $[v][d]$ by one feature, place, thus forming $*[vz][td]$. This has the effect of also banning *zd*, which has an intermediate O/E value of 0.9475. Further generalization may occur when $*[f][d]$ (O/E value 0.2169), $*[v][b]$ (O/E value 0.4091), $*[s][b]$ (O/E value 0.3879) (and potentially others) are added: $*[fsvz][btd]$.

Generalization has the advantage of capturing biphones that would otherwise not be included in any constraints. For example, in the above constraint, $[sd]$ will be ruled out, although *sd* has an O/E value above 0.5 (0.5096). $[sd]$ is most often at a word boundary (247 word-internal occurrences vs. 2525 across a word boundary), so including it in the constraint is beneficial to the model.

However, in other cases, StaGe overgeneralizes in forming constraints. Biphone $[st]$ is also ruled out by the above constraint. Further, this biphone is most often word internal (17255 word internal vs. 3566 across a word boundary). However, $[st]$ has a high O/E value - 2.6240 - and so forms a contiguity constraint $\text{Contig-}[s][t]$. As will be seen, ranking resolves the problem of overgeneralization.

In the event that a biphone is not included in any constraint, it is segmented by chance (half of the time), a property of threshold-based segmentation models.

3.1.4 Ranking constraints and evaluating word boundaries

Constraint ranking generally takes place in a supervised way: the learner has access to the optimal form, and demotes or promotes constraints until the optimal form is always the form derived from the constraint ranking (Tesar & Smolensky, 1998; Hayes & Wilson, 2008). This strategy, however, is not available to the StaGe learner, since word-learning is taking place - the optimal form cannot be known. An alternative,

unsupervised strategy is to rank constraints based on distributional criteria, to which the learner does have access.

Constraints are ranked based on their expected values. The higher the expected value of a constraint, the higher its ranking. Generalized constraints, which do not have a single expected value, are ranked based on the average of the expected values of the biphones in the constraint. Thus, each constraint receives a number, and then the numbers are ranked in order from highest to lowest, giving the overall constraint ranking.

Because specific constraints will typically have higher expected values than generalized constraints, which are given a ranking based on the average expected values of their component biphones, StaGe’s tendency to overgeneralize will be checked. Continuing the example in the previous section, [st] has a high enough O/E value to generate a specific contiguity constraint, Contig-[s][t], with O/E value 2.6240 and ranking value Contig-IO([s][t]) 7505.1849. This constraint ranks higher than the highest ranking general constraint that includes *st : *[fsvz][st], with ranking value 1331.1245.

3.1.5 Summary

To summarize, StaGe uses statistical learning to induce phonotactic constraints of two kinds: markedness constraints (segment the biphone) and contiguity constraints (do not segment the biphone). These constraints are ranked according to the expected frequency of the relevant biphone(s). This ranking is then used to evaluate biphones in the test corpus, making the decision whether or not the biphones should have a word boundary placed between them. The O/E learner is the statistical learning component of StaGe; it makes word boundary decisions based on the O/E value of a biphone: if $O/E < 1$, the biphone is segmented, and if the $O/E > 1$, it is kept intact.

Results (hit rate, false alarm rate, and d') for the O/E and StaGe models are given in the table below; standard deviation is shown in parentheses:

Model	hit rate	false alarm rate	d'
O/E	0.3734 (0.0098)	0.1345 (0.0146)	0.794
StaGe	0.4477 (0.0261)	0.1316 (0.0169)	1.000

TABLE 3.2: Results with full Contrast

For further details see [Adriaans & Kager \(2010\)](#).

3.2 Experiment 2a: Place

Experiment 1 showed that neutralizing place of articulation in stops and nasals had the least effect on performance. Here I examine the effects of neutralizing the place information in nasals and stops in the input to an *unsupervised* learner, comparing the hit and false alarm rates of the StaGe and O/E models to a control condition in which place is not manipulated, the question being: if a contrast is neutralized, will the unsupervised learner be devastated (d' approaching 0.0)?

Materials

The model used is StaGe and the O/E component of the StaGe model (Adriaans & Kager, 2010). The model is tested on the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN), which is a series of utterances. While it is a phonemically transcribed corpus, variation between speakers is present in the input. Also, assimilation across word boundaries is present, making it a more detailed, realistic corpus than many. For further details of the corpus, see the section A

Method

Inputs were tested using ten-fold-cross-validation: the corpus was split into ten subparts, and for any given trial one of these subparts served as the test, while the other nine parts served as the training set. In this way, for any given utterance, it served as the test exactly once, and was used in training for nine other trials.

The input was neutralized as follows:

m,n,ŋ > n
p,t,k > t
b,d,g > d

TABLE 3.3: Neutralizations in the input

The StaGe model reports the hit and false alarm rates. StaGe also reports the constraints learned and their ranking, decisions for each biphone (ie, whether or not it is segmented), and which constraint determined the decision.

Results and discussion

Model	hit rate	false alarm rate	d'
O/E	0.3636 (0.0102)	0.1400 (0.0142)	0.722
StaGe	0.4616 (0.0157)	0.1569 (0.0126)	0.894
supervised	0.5362	0.0644	1.6550

TABLE 3.4: Results with place neutralization

The mean for the hit rates and false alarm rates across the ten trials (standard deviation in parentheses), and the d' for the O/E model, the StaGe model, and, for comparison, the supervised model from Experiment 1 are reported in Table 3.4.

Clearly, in neither case does d' approach 0.0, though both are worse than the idealized, supervised model. Indeed, it is interesting to compare the results for the two models when full contrast is present; the results for the O/E and StaGe models with full contrast is repeated below:

Model	hit rate	false alarm rate	d'
O/E	0.3734 (0.0098)	0.1345 (0.0146)	0.794
StaGe	0.4477 (0.0261)	0.1316 (0.0169)	1.000

TABLE 3.5: Results with full Contrast

Not only does d' not approach 0.0, the performance is not so different from when the model has full contrast. This is especially surprising because Dutch has a potentially important constraint reliant on place: nasal-stop sequences must agree in place of articulation. Based on this, it might be expected that, given place contrast, the learner would find that nasal-stop sequences should be segmented if they do not share place of articulation, but should remain intact if they do. For the most part, this is the case; nasal-stop sequences are almost never word-internal if they are not homo-organic.

Unsurprisingly, in the neutralization condition there is a failure to discover this constraint: in fact, there is a failure to discover any constraint. Nasal-stop biphones are segmented by chance, which is what StaGe does if it has no constraint for a particular biphone. In the control condition, however, all nasal-stop biphones are segmented, with the exception of [nd], which has a high O/E value and forms a contiguity constraint, as well as a high expected value, meaning the constraint is highly ranked². While nasal-stop biphones that do not share place are segmented as might be expected, even those agreeing in place also are segmented, due to highly ranking constraints like *[mn][Sfkpstx] and *[Nmn][GSZbdfghkpstvxz], which overgeneralize. There is no contiguity constraint that can save homo-organic nasal-stops, since they do not reach high enough O/E values to form a constraint - the threshold for StaGe is too high.

3.2.1 Why some constraints and biphones do not matter as much as others

Since [nd] is by far the most frequent of any nasal-stop cluster that agrees in place ([nd] occurs almost as often alone as all other homo-organic nasal-stop clusters combined - 27,522 vs. 30,936 occurrences, meaning that [nd] alone accounts for about half of homo-organic nasal+stop clusters), StaGe is not as hurt as it might be by failing to discover a contiguity constraint for the other nasal-stops agreeing in place. Thus, segmenting at random approximates performance in the control condition for this kind of cluster, in the sense that each model gets about half "correct." However, in the control condition, StaGe segments all nasal-stop clusters, excepting [nd], whereas in the neutralization

²Other homo-organic nasal+stops also have high O/E values (most above 2.0), but low expected values, causing their specific contiguity constraints to be ranked below general markedness constraints; hence, they are segmented; the exception is nt, but its O/E value is too low to form a contiguity constraint

condition, it only segments half, whether they disagree in place or not. This means the hit rate ought to go down, because many nasal+stop clusters are *not* homo-organic, and presumably should be segmented. That it does not is due to nasal+voiceless stop clusters being relatively infrequent.

Up to this point, it has been assumed that segmenting homo-organic nasal+stop clusters would make the false alarm rate rise quite a bit; again, that it does not could be related to nasal+stop clusters not being overwhelmingly frequent. However, additionally, it is assumed that these clusters are more often than not word-internal. However, as it turns out, while they are more often word internal than nasal+stop clusters that differ in place of articulation, they are not as often word internal as might be expected.

nasal+stop	total (observed)	expected	O/E	word internal	word boundary
mp	2293	1006.6654	2.2778	1132	1161
mb	3815	1089.7970	3.5007	839	2976
nt	18923	13158.2548	1.4381	15264	3659
nd	27522	9104.6512	3.0229	12333	15181
ŋk	5240	601.2880	8.7146	3035	2205
ŋg	665	69.9456	9.5074	523	142

TABLE 3.6: Homo-organic nasal+stop cluster frequency across a word boundary

Most of the homo-organic nasal+stop clusters in fact occur about as often across a word boundary as word-internally. Thus, segmenting them should make the false alarm rate rise, but also the hit rate. It is not expected that *d'* would change very much whether homo-organic clusters are segmented or not, which is in fact the result.

The O/E model treats nasal-stop clusters differently. In the O/E model, in the control condition, the homo-organic nasal-stop clusters are not segmented, while those which differ in place are segmented. In the neutralization condition, the O/E learner segments all nasal-voiceless stop clusters while leaving intact all nasal+voiced stop clusters - likely due to the high O/E value that [nd] has, and the much greater frequency of [nd] than any other nasal+voiced stop.

Since [nd] is much more frequent than either [mb] or [ŋg], it is the behavior of [nd] that may be expected to have the most effect for the hit and false alarm rates. [nd], as seen in the table, is more often at a word boundary than word internal, but only slightly so. Thus, not segmenting [nd] would make the hit rate and false alarm rate fall about equally, as compared with segmenting it. However, [nd] is kept intact in the control condition as well, so overall little change between the two might be expected. If anything, hit rate would fall, since nasal+stop clusters that differ in place would no longer be segmented. This is apparent when all nasal+voiceless stop clusters are considered.

nasal+stop	total (observed)	expected	O/E	word internal	word boundary
mt	1879	5703.9323	0.3294	1320	559
mk	273	2405.6425	0.1135	78	195
md	2199	3946.7479	0.5572	873	1326
mg	8	279.8396	0.0286	1	7
np	653	2322.2505	0.2812	33	620
nk	1282	5549.5148	0.2310	57	1255
nb	1233	2514.0246	0.4904	87	1146
ng	38	645.5548	0.0589	3	35
np	52	251.6150	0.2067	25	27
nt	540	1425.6922	0.3788	277	263
nb	187	272.3937	0.6865	51	136
nd	579	986.4857	0.5869	51	528

TABLE 3.7: Non-homo-organic nasal+stop cluster frequency across a word boundary

Clearly, non-homo-organic nasal+stop clusters are more often across a word boundary than word internal. Equally clear, however, is that the frequency of any of these clusters is so small, that it is hardly expected to make a difference what the model does with them.

In the case of nasal+voiceless stop clusters, [nt] is the most frequent. Unlike [nd], [nt] is more often word internal than across a word boundary. Ironically, [nt] would be the better target to not segment. Segmenting it should cause the false alarm rate to increase, while hit rate would not be expected to increase very much (given that other nasal+voiceless stop clusters are less frequent). In fact, the O/E model has little overall change in either hit or false alarm rate. What is perhaps most interesting, however, is that what seems like an important cue to contiguity - place assimilation in stops and nasals - turns out to not be very helpful, since homo-organic nasal+stop clusters are across a word boundary about equally as often as they are word internal, and since non-homo-organic nasal+stop clusters are rare.

Another potentially important constraint which could be relevant is that banning geminates. In the control case, geminates are always segmented, as their O/E values tend to be low. For the most part, no single geminate is very frequent, but collectively they are fairly frequent in an unsegmented corpus (ie, without word boundaries). A table of nasal and stop geminates and their frequencies is given in Table 3.8.

geminate	total (observed)	O/E value	at word boundary	not at word boundary
pp	103	0.2405	97	6
bb	194	0.3879	191	3
tt	9138	0.6706	9118	20
dd	4456	0.6613	4449	7
kk	950	0.3857	945	5
gg	24	0.7176	23	1
mm	325	0.1648	236	89
nn	446	0.0388	440	6
ŋŋ	1	0.0081	1	0
total	15637		15500	137

TABLE 3.8: Plosive and nasal geminate frequencies

In the neutralization condition, [TT], [DD], [TD], [DT] and [NN] are segmented by the O/E model, where T stands for voiceless stops, D stands for voiced stops, and N stands for nasals. This means that all stop clusters and all nasal clusters will be segmented. However, all stop-stop and nasal-nasal clusters are segmented by both StaGe and the O/E model in the control condition anyway. While not illegal - stop-stop clusters may occur word-medially - these clusters are rare, and occur far more often at the word boundary than word-internally.

For a full list of biphones which are segmented differently in each condition for StaGe vs. the control, along with their frequencies word-internally vs. at a word boundary, see the appendix. Interestingly, the biphone which is most frequent among these is ən; this happens to be the plural morpheme in Dutch. In the control condition, where n is in contrast with m and ŋ, ən is kept intact, since ən occurs very often (as the plural morpheme). In the control condition, it is segmented at chance. This could well explain what rise there is in false alarm rate, since ən is more often word internal than at a word boundary, and is also very frequent.

Overall, while still a small change, StaGe has more of a change in hit rate, in false alarm rate, and in d' than does the O/E model. Note that the d' change (Table 3.4) for the StaGe model, between the control and neutralization conditions, is less than the change in d' between the StaGe model and the O/E model in the control condition (a difference of 0.106 vs. 0.206). However, the change in hit rate and false alarm rate is greater between conditions than between models. Nonetheless, StaGe retains its superior d' (although a higher false alarm rate). That there is a greater *change* in performance for StaGe could be due to a threshold effect; that is, the change is less in the O/E model because it is already doing so poorly.

3.3 Experiment 2b: Voice

Experiment 2a showed that neutralizing place of articulation in the input to the unsupervised learners StaGe and the O/E learner has little effect on performance, as might be expected based on Experiment 1. Experiment 2b examines the effects of neutralizing the voice information in obstruents, comparing the hit and false alarm rates of the StaGe and O/E models to a control condition. The question is two-fold: a) will the performance of either model be devastated when a contrast is neutralized? and b) will the change in performance be greater than in Experiment 2a, where place is neutralized? Voice neutralization had a greater effect on performance for the supervised learner, so it is likely that it will result in a greater effect on performance for the unsupervised learners. Further, if this is the case, then there is further support for the correlation between the usefulness of a contrast and its perceptual salience, since voice was found experimentally to be more salient than place of articulation (Cutler et al., 2004; Miller & Nicely, 1955).

Voice as a contrast might well be expected to serve as a good cue for word boundaries. In careful speech, adjacent obstruents must agree in voicing, and word-final obstruents may not be voiced. Accordingly, adjacent obstruents which do not agree in voice - specifically, a -voice obstruent followed by a +voice obstruent - may not be word internal, while adjacent obstruents which do agree in voice might be expected to be word internal most of the time.

Materials

Same as for experiment 2a.

Method

The method was the same as for Experiment 2a, except that the contrast neutralized was voice rather than place of articulation. Voice in all obstruents were neutralized; hence, the following neutralizations:

Results and discussion

The hit and false alarm rates (with standard deviation in parentheses), and the d' for the StaGe and O/E model are shown in Table 3.10. For comparison, results for the control are again shown in Table 3.11.

p,b > p
t,d > t
k,g > k
f,v > f
s,z > s
ʃ,ʒ > ʃ
x,ɣ > x

TABLE 3.9: Neutralizations in the input

Model	hit rate	false alarm rate	d'
StaGe	0.5458 (0.0379)	0.1830 (0.0212)	1.041
O/E	0.4059 (0.0159)	0.1406 (0.0152)	0.852

TABLE 3.10: Results with voice neutralization

Model	hit rate	false alarm rate	d'
StaGe	0.4477 (0.0261)	0.1316 (0.0169)	1.000
O/E	0.3734 (0.0098)	0.1345 (0.0146)	0.794

TABLE 3.11: Results with full contrast

Again, the model is not devastated by the reduction in contrast in the input. Indeed, interestingly, d' actually increases as compared with the corresponding model when trained with full contrast in the input. The trend is the same for both the StaGe and O/E model.

As said, d' for the StaGe model is higher, although false alarm rate is also higher. Since the false alarm rate rises quite a bit, the increase in d' is entirely due to a higher hit rate. In the O/E model, hit and false alarm rates change only slightly more than in Experiment 2a, hit rate again in the positive direction. d' changes more for the O/E model than for the StaGe model, despite a smaller change in hit and false alarm rate. Again, the change in d' for the O/E model is less than the change for the same model in Experiment 2a, although the change in hit and false alarm rate is higher. To focus only on the change in d' masks what is going on at a more basic level; the greater change in d' replicates the result of Experiment 1, and supports the correlation between perceptual salience and usefulness in word segmentation.

In short, while d' changes less than in Experiment 2a, hit and false alarm rates in both models change more. Why does voice neutralization result in such a large increase in the hit and false alarm rates?

3.3.1 Exploring the segmentation decisions

One reason that the voice neutralization condition could result in a higher hit rate is that, while a difference in voice may cue a word boundary, as already noted, adjacent obstruents in general are more often than not at a word boundary. Adjacent obstruents differing in voice are always bad word-internally, while word-internal adjacent obstruents are legal provided they agree in voice. From a classic phonological description of Dutch, one would expect the model to form a constraint banning an obstruent-obstruent sequence that does not agree in voice, but to not have a constraint against obstruent-obstruent sequences in general, since they are allowable. However, adjacent obstruents are in general under-represented word-internally, and even if they are legal, they are improbable.

Further, while not obligatory in Dutch, on examination of the corpus, voicing assimilation across a word boundary is common. Thus, while a difference in voicing cues a word boundary, agreement in voicing does not cue contiguity. This makes it less surprising that there are many cases of constraints against adjacent obstruents, even when these obstruents do agree in voice. It also means that the model may not increase its false alarm rate very much by segmenting all or nearly all obstruent clusters.

Table 3.12 shows these under-represented biphones (with O/E value less than 1.0), their word internal count, word boundary count, and their O/E values.

In fact, all of these obstruent biphones have O/E values less than 0.5, which means that in the O/E model they will be segmented, but that also in the StaGe model, there will be a segmentation constraint for each of these biphones. Further, with a few exceptions, these biphones are more often than not at a word boundary. With generalization, StaGe may be expected to make constraints against adjacent obstruents in general, and not only against adjacent obstruents not agreeing in voice, even in the control condition. Thus, while none of the biphones above are very frequent, the effect of generalization will mean that all obstruent clusters will be segmented, unless a contiguity constraint is formed for a particular biphone and outranks the general markedness constraint.

Since most obstruent-obstruent biphones occur infrequently, to get a clearer picture it make sense to focus only on those that occur most often. These will have the biggest effect. Adjacent obstruents which occur more than 500 times include: sp, st, sx, xt, ts, tf, ks, sf, ps, kf, ft, px, pt, px, pf, xf, kt, fs, kx, xs, sk, fx, tk, tp, kk, xk, gb, gd, zb, db, zd, dd, bd, vd, sd, sb, tb, xd, kd, td, and fd. Most of these occur over 1000 times. A table is given below showing these most frequent obstruent biphones, their segmentation rates, and how often they occur word-internally vs. at a word boundary (in the corpus with correct word boundaries).

biphone	total	word internal	word boundary	percent word boundary	O/E
ʒb	3	0	3	1.00	0.3124
ʒd	7	0	7	1.00	0.2013
vg	2	0	2	1.00	0.0226
bg	15	1	14	0.933	0.1141
dv	33	3	30	0.909	0.0264
tp	1059	129	930	0.878	0.4279
tk	2811	373	2438	0.867	0.4753
kp	317	57	260	0.820	0.3037
xp	271	52	219	0.808	0.3095
dg	15	3	12	0.800	0.0315
gv	27	6	21	0.778	0.3051
bz	37	9	28	0.757	0.0823
dy	20	5	15	0.750	0.0226
pk	270	70	200	0.741	0.2587
fk	289	80	209	0.723	0.1830
vb	141	42	99	0.702	0.4091
ʃk	20	7	13	0.650	0.1851
zg	13	5	8	0.615	0.1126
fp	114	47	67	0.588	0.1725
xk	624	283	341	0.546	0.2982
ʃp	10	5	5	0.500	0.2212
dz	252	135	117	0.464	0.1547
by	9	5	4	0.444	0.0368
ʃt	47	30	17	0.362	0.1834
vd	557	366	191	0.343	0.4463
bv	28	23	5	0.179	0.0812

TABLE 3.12: Obstruent biphones agreeing in voice

Table 3.13 shows the O/E values for the control condition and neutralization condition for each biphone. Boldface indicates a value that has changed such that it is above 1.0 in the neutralization condition, and will no longer be segmented (at least by the O/E model), while italics indicate a value that has changed such that it is below 1.0 in the neutralization condition.

	total	word internal	word boundary	% word boundary	O/E (control)	O/E (neutralization)
sp	3692	3279	413	0.1119	2.6364	1.4676
st	20821	17255	3566	0.1713	2.624	1.7593
sx	6663	4689	1974	0.2963	2.3738	2.3738
xt	9438	8635	803	0.0851	1.9025	1.2314
ts	14223	6931	7292	0.5127	1.7925	1.0678
tf	6282	569	5713	0.9094	1.6777	0.9968
ks	4942	2657	2285	0.4624	1.4768	1.3304
sf	2940	463	2477	0.8425	1.3877	1.3877
ps	1883	1009	874	0.4642	1.3446	0.6529
kf	2117	45	2072	0.9787	1.3405	1.2127
ft	3938	3279	659	0.1673	1.0517	0.7103
tx	5164	746	4418	0.8555	1.041	0.6211
pt	2563	1283	1280	0.4994	1.0355	0.4509
px	886	489	397	0.4481	1.012	0.4996
pf	641	189	452	0.7051	0.97	0.4781
xf	1173	177	996	0.8491	0.8856	0.8856
kt	5115	3571	1544	0.3019	0.8648	0.7096
fs	1778	709	1069	0.6012	0.8392	0.8392
kx	1538	67	1471	0.9564	0.7351	0.6636
xs	1989	895	1094	0.5500	0.7086	0.7086
sk	1757	834	923	0.5253	0.525	0.4751
fx	674	423	251	0.3724	0.5088	0.5088
tk	2811	373	2438	0.8673	0.4753	0.2609
tp	1059	129	930	0.8782	0.4279	0.4611
kk	951	5	946	0.9947	0.3812	0.3156
xk	624	283	341	0.5465	0.2982	0.2685
gb	1391	171	1220	0.8771	10.5816	0.7825
gd	1745	108	1637	0.9381	3.6654	0.7096
zb	675	155	520	0.7704	1.5011	0.7976
db	1927	263	1664	0.8635	1.0394	0.4611
zd	1543	196	1347	0.8730	0.9475	0.4116
dd	4462	7	4455	0.9984	0.6646	0.3921
bd	948	181	767	0.8091	0.5113	0.4509
vd	557	366	191	0.3429	0.4463	0.1943
sd	2798	247	2551	0.9117	0.5096	1.7593
sb	588	156	432	0.7347	0.3879	1.4676
tb	832	48	784	0.9423	0.3105	0.4611
xd	898	351	547	0.6091	0.2616	1.2314
kd	1012	89	923	0.9121	0.2473	0.7096
td	2051	10	2041	0.9951	0.2114	0.3921
fd	562	224	338	0.6014	0.2169	0.7103

TABLE 3.13: Frequent obstruent biphones and their segmentation rate

Note that no voiced+voiceless obstruent pairs occur more than 500 times (or even more than 50 times): this is not surprising, since this pair is entirely illegal in Dutch - a voiceless+voiced pair is possible, since the voiceless obstruent may be word-final, and assimilation across a word boundary is optional. There is no case that would allow for obstruents to disagree in voice in this way. Strangely, voiceless+voiced obstruent biphones are not always at a word boundary (though they most often are). Examining particular cases show that voiced+voiceless obstruent pairs occur in long obstruent/sonorant clusters, like [dadlɡk], where only the final obstruent is devoiced. Word-internal voiceless+voiced biphones occur where the voiced portion occurs seemingly as an offset of a voiceless phoneme, due to assimilation with a following vowel. For example, satdaxmərɣə, zakdɪŋ.

In the StaGe model, many of these O/E values do not fall below the 0.5 threshold, so these will not form a constraint per se, although with generalization, all appear in a segmentation constraint. In fact, even in the control condition, most of these obstruent biphones are segmented, mainly because they fall into neutral territory and thus appear only in generalized constraints; it happens that most of these general segmentation constraints outrank any relevant general contiguity constraints that could be formed, since related obstruent clusters tend to have low O/E values. However, crucially, [sp], [st], and [sx] are not segmented, because the contiguity constraints formed in these cases - Contig-IO([s][pt]) and Contig-IO([s][x]) - outrank more general segmentation constraints such as *[Sfɪkpstx][bɪfpstvz], which contains both *sp and *st. These three biphones have some of the lowest rates of actual word boundaries, as well as being some of the most numerous of the obstruent biphones. [st] alone occurs 20,821 times, more often than any other single obstruent-obstruent biphone. Hit rate inevitably rises when segmentation rate increases; even these over-represented biphones are sometimes at a word boundary rather than word internal. However, this primarily explains the higher false alarm rate of the neutralization condition, which segments these biphones, but does not entirely explain the higher hit rate, especially since hit rate rises more than false alarm rate.

On further examination of biphones whose segmentation differs between conditions, another class of biphones emerges as a candidate to explain the rise in hit rate. In the control condition, vowel+voiceless obstruent sequences are often kept intact. However, when voice is neutralized, these sequences are segmented. When these are segmented, false alarm rate increases, but hit rate would increase even more, since overall vowel+obstruent is more often than not at a word boundary.

Table 3.14 and Table 3.15 show vowel+obstruent biphones which are segmented differently between conditions.

Biphone	Neutralized	Control	Total Count	Word Internal Count	Word Boundary Count
ɪd	1	0	1259	1141	118
ɣg	0	1	102	101	1
ɛf	1	0	839	741	98
ɛz	1	0	306	229	77
e:b	1	0	1	1	0
œyp	1	0	51	50	1
ət	1	0	20020	14458	5562
yd	1	0	645	302	343
ɣʒ	0	1	24	2	22
ʌup	1	0	68	25	43
ez	1	0	1344	1204	140
ih	1	0	1000	52	948
ɛt	1	0	6988	6891	97
əs	1	0	16212	9400	6812
œyb	1	0	37	31	6
ap	1	0	949	600	349
ʌf	1	0	1	1	0
əf	1	0	7975	2179	5796
ep	1	0	483	379	104
əd	1	0	20763	6343	14420
əb	1	0	8765	2715	6050
ɛs	1	0	2538	2357	181
yp	1	0	118	96	22
yb	1	0	280	143	137
od	1	0	2238	1144	1094

TABLE 3.14: Differently segmented vowel+obstruent biphones

Biphone	Neutralized	Control	Total Count	Word Internal Count	Word Boundary Count
ɔg	0	1	80	79	1
ɣf	0	1	93	43	50
ɣv	1	0	7	6	1
ɣb	1	0	108	104	4
ɾy	0	1	357	348	9
oz	1	0	573	366	207
av	0	1	340	265	75
es	1	0	2540	2271	269
ub	1	0	168	87	81
ʌub	1	0	203	41	162
os	1	0	1292	1005	287
yt	1	0	801	662	139
ɔf	0	1	47	41	6
ab	0	1	601	337	264
et	1	0	5005	4657	348
ɛv	1	0	113	73	40
əp	1	0	4011	1210	2801
ab	1	0	849	342	507
ed	1	0	4476	3489	987
yh	1	0	239	40	199
ɔ:d	1	0	4	4	0
up	1	0	488	477	11
ad	1	0	5494	1818	3676
at	1	0	8341	7444	897
əz	1	0	8201	3302	4899
əv	1	0	6036	1849	4187
uh	1	0	192	21	171
ɛd	1	0	2122	1782	340
ot	1	0	2823	2628	195

TABLE 3.15: Differently segmented vowel+obstruent biphones

This is especially likely because, for the O/E learner, many of the vowel+obstruent sequences are not segmented in either the control or in the neutralized condition. This would explain why neither false alarm rate nor hit rate change as much for the O/E model as for the StaGe model.

For a full list of biphones whose segmentation differs between conditions, see the appendix.

3.4 Comparing the StaGe and O/E models

In both Experiment 2a and 2b, d' is consistently higher for the StaGe model. However, overall performance *changes* more for StaGe. As mentioned in the discussion of Experiment 2a, this could be a threshold effect; since the O/E model already does so poorly,

change in performance is minimal.

However, although d' is consistently higher for StaGe, in both Experiment 2a and 2b neutralization results in a jump in false alarm rate. This is especially the case in Experiment 2b, where false alarm rate goes up to 0.1830 (from 0.1316 in the control). As discussed in Experiment 1, undersegmentation may be preferable to a rise in false alarm rate, meaning that a high d' does not compensate for a high false alarm rate. In this case, the O/E model may be said to have superior performance in Experiments 2a and 2b, since the false alarm rate is lower.

The O/E model has a lower false alarm rate primarily, it seems, because overall there is less change between when the input has a neutralized contrast and when it does not. As noted, this is because many biphones are not segmented in either condition. It seems that these results may at least be partly due to the thresholds chosen for each model, rather than any other property. Neutralizing contrasts results in a large number of biphones' O/E values falling into the neutral area between 0.5 and 2.0; as a result, in StaGe the segmentation decisions of these biphones will be based on generalized constraints, since the specific constraint is prevented from being formed. This results in overgeneralization and oversegmentation. In contrast, in the O/E model, there are no generalized constraints other than the ones due to neutralization (eg, *VD where V is neutralized f,v and D is neutralized t,d). Likewise, there is no neutral area; all biphones are segmented based on whether the O/E value is below 1.0 or not. While [Adriaans & Kager \(2010\)](#) show that the superior performance of StaGe does not depend on the exact threshold values chosen, this may not be the case here. As this is not the focus of this thesis, I will not go into this further, but it should be kept in mind when considering the performance of the two models.

Chapter 4

General Discussion and Conclusions

I have approached the problem of word segmentation from the point of view of markedness by perceptual salience. I asked whether, broadly, perceptual salience of a contrast correlates with its usefulness as a cue for word boundaries. This does indeed seem to be the case: place of articulation, which is less perceptually salient both cross-linguistically and independently of context, seems to be less helpful for segmentation than other contrasts in Dutch. This extends to voice contrasts, which, while apparently more useful than place of articulation, are less useful than manner contrasts continuant and nasal, and like place of articulation are not categorizable from the acoustic signal (Lin, 2005). Further, it looks as though place of articulation in nasals, which is less perceptually salient than place of articulation in stops or in fricatives, is likewise less helpful than either of these for word segmentation.

Additionally, the contrasts that are more perceptually salient - continuant and nasal - are more useful than either place or voice for biphone (phonotactic) word boundary cues. Precision and d' were both lower when these contrasts were removed from the input. While the relation between perceptual salience and usefulness in word boundary cues is less clear in Experiment 1b, overall the relation holds, with fricative place of articulation being an exception.

Thus, all in all these tests are in support of the notion that there is a correlation between perceptual salience of a contrast and its usefulness in phonotactic cues for word boundaries.

These conclusions also bring up a question. Given that place of articulation is less helpful than other contrasts for finding word boundaries, it is surprising that it should be such a common contrast cross-linguistically.

Full answers to this question are beyond the scope of this thesis. However, I will make a couple suggestions. Place of articulation may be a common contrast for articulatory reasons. Altering place of articulation is relatively easy, regardless of the other features of a natural class. For example, compared with nasal contrasts, which are very difficult to produce in fricatives, place of articulation contrasts may be combined more easily with other features. The other reason place of articulation may be so common is because it introduces many more types to a phoneme inventory, given that it is not a binary contrast. A phoneme inventory which relied only on continuant, sonorant, nasal, and voicing distinction would have far fewer possibilities, particularly given that not all theoretically possible contrasts are practically producible. That place of articulation neutralization in fricatives was more devastating to the learner than in either stops or nasals may be a compromise between perceptual ease and the pressure to have more types.

In the case studies in Experiments 2a and 2b, it was found that in the case of unsupervised learning, neutralization of the least perceptually salient contrasts did not devastate the learner. Given the results of Experiment 1a, this was to be expected. Further, between the latter two experiments, the same trend was found as in Experiment 1a; neutralization of place of articulation in nasals and stops had less of an effect on the learner than did voice neutralization.

While the effect on the StaGe learner was greater than that on the O/E learner, this may be largely because the O/E learner already does so poorly (ie, a threshold effect); it may also be due to the threshold values chosen for StaGe. Regardless, the same overall trend is observed: greater effect on performance for voice neutralization than for place neutralization. This suggests that the effect will hold for unsupervised learners with similar learning mechanisms, and depends less on the specific model.

Thus, all results point to the same conclusion: there is a correlation between perceptual salience of a contrast and its usefulness in word segmentation in Dutch.

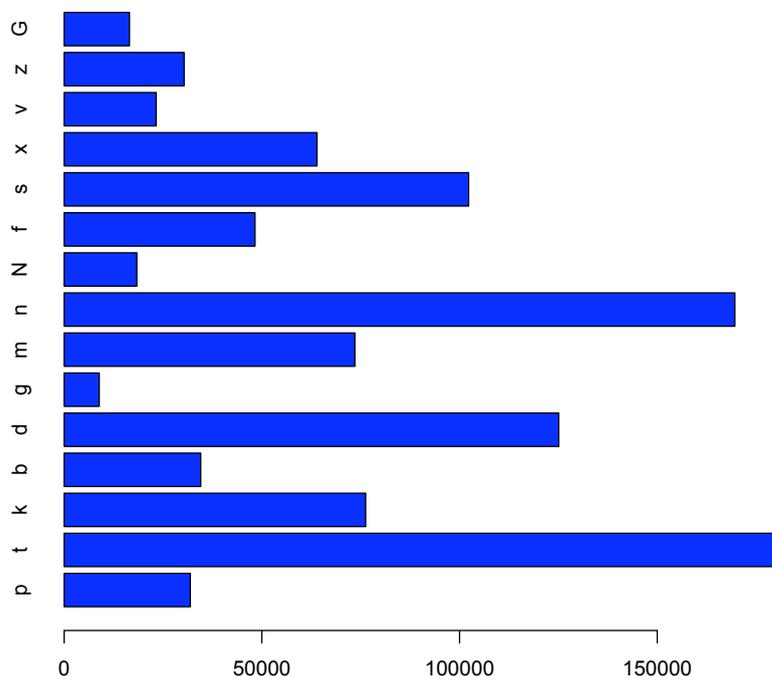
4.1 Frequency Distributions in Dutch

Putting perceptual salience aside, it is of interest to consider the relative distributions of contrasts in Dutch. Having noted in the discussion of Experiment 2a that [nd] was so much more frequent than any other nasal+stop sequence, and seen how as a result it was

the treatment of [nd] that mattered more for the hit and false alarm rates than any other nasal+stop sequence, it would not be surprising if relative frequencies of single phones could make similar predictions. That is, it could be that the effect on the hit and false alarm rates of the learner as a whole will depend on the frequencies of phones within a class being neutralized. The prediction is that if frequencies within a class are similar, the effect will be larger; if one phone dominates the others in its class, the effect will be smaller. This is true provided that the same asymmetry holds throughout relevant classes; for example, if coronals are more frequent in all manners of articulation. This is then also true regardless of the *n* in an *n*-phone model; hence, if phone (ie, unigram) *x* is overall much more frequent than other members in its class {*x*, *y*, *z*}, then the *n*phones that it is in will have more weight than the *n*phones containing *y* or *z*, since they will be overall more frequent.

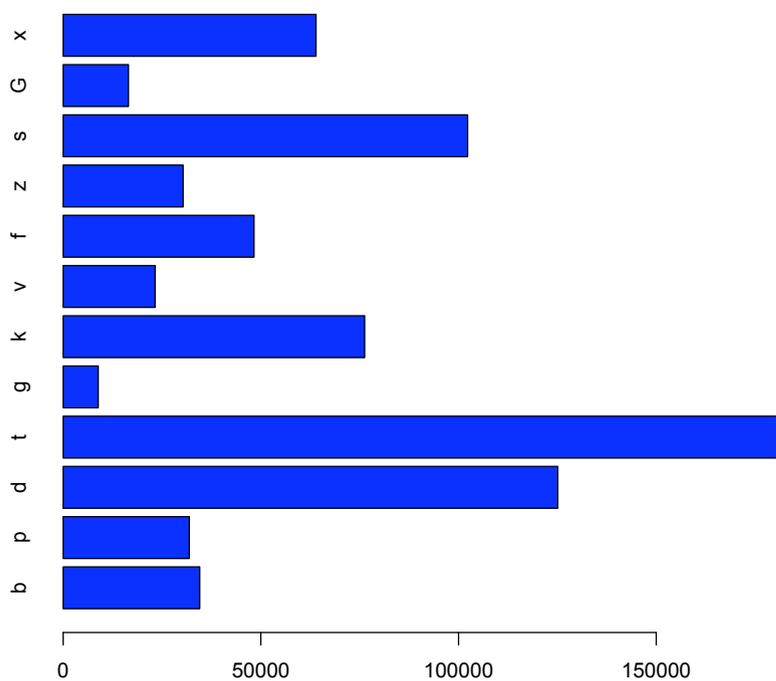
As an example of this, consider the biphones [pb], [pp], [pd], [pt], [pk], [pg]. Of these, [pt] and [pd] are the most frequent, by virtue of [t] and [d] being vastly more frequent than *p*, *k* or *b*, *g*, respectively (both may be expected to have low O/E values, however, since expected values are much higher than observed values). Accordingly, the segmentation decisions for biphones [pt] and [pd] will matter more than for the others; if place of articulation is neutralized, and all TT and all TD are treated the same, little change is expected, since [pp] and [pk] are less frequent than [pt], and likewise [pp] and [pg] are less frequent than [pd]. Thus, it can be seen how a unigram phone can dominate segmentation decisions, even though the calculations are performed on biphones, not unigrams.

Looking at the frequencies of phones in different classes, a pattern emerges.

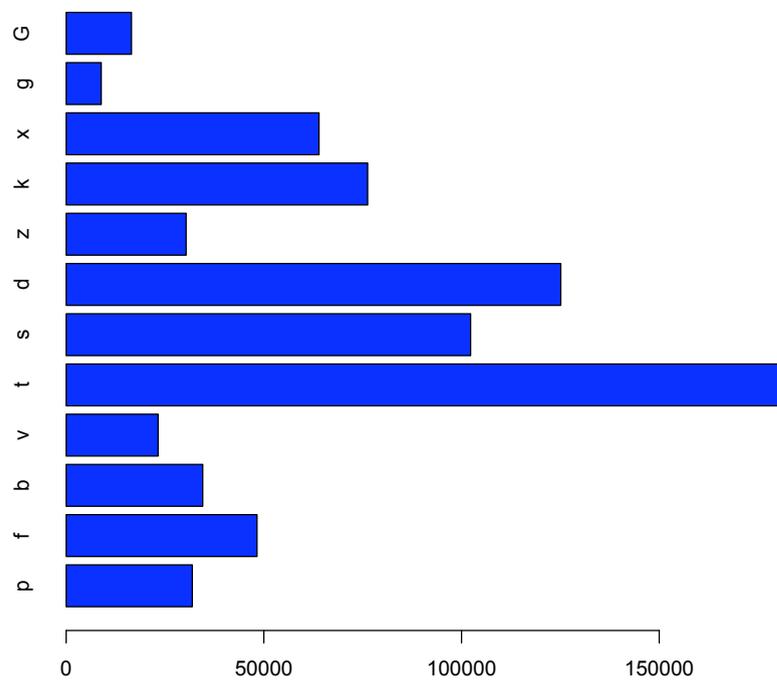
Comparison of Phone Frequencies by Place Contrast

In the case of place contrasts, coronal is far more frequent than either labial or dorsal. It is perhaps for this reason that lack of place contrast has a smaller effect - less change in hit rate, less change in false alarm rate - than voice. Voice, in contrast, with the exception of *k* vs. *g* and *x* vs. *ɣ* (which are essentially allophonic in Dutch in any case), has a smaller difference in frequencies.

Comparison of Phone Frequencies by Voice Contrast

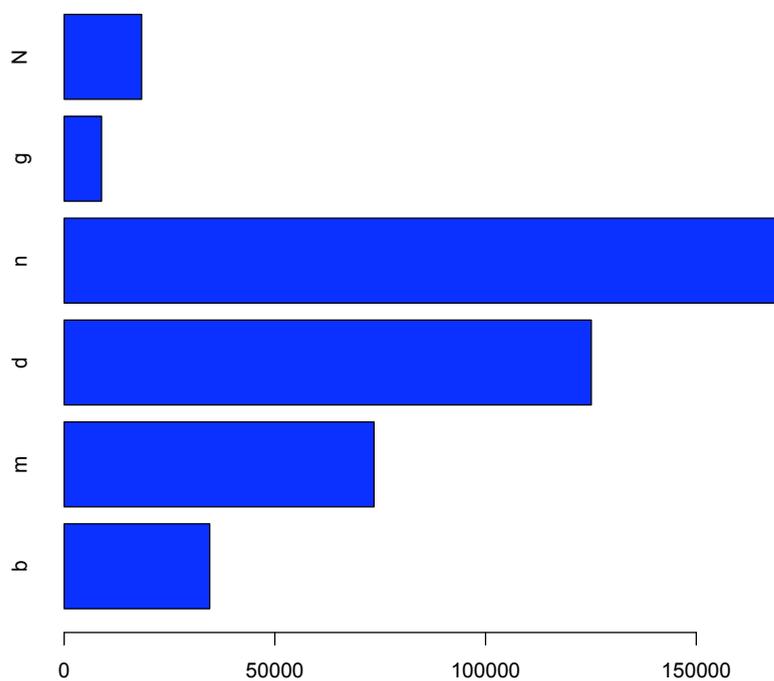


Voice contrasts do not have quite the same degree of asymmetry; b and p are approximately equally distributed, and t, while certainly more frequent than d, is not more frequent on the scale of coronals vs. labials and dorsals. The difference between k and g may be explained by the fact that k and g are not contrastive in Dutch, g appearing only as an allophone of k; the same may be said for the difference between x and ɣ. Among fricatives, the voiceless phone of each pair is more frequent; however, the trend of voiceless being much more frequent than voiced does not hold across voiced-voiceless pairs in the way that coronal being far more frequent than labial or dorsal holds for place of articulation contrasts.

Comparison of Phone Frequencies by Continuant Contrast

Discounting the g-y frequency difference, stops tend to be somewhat more frequent than fricatives. Again, however, the frequency differences are not on the scale of the coronal dominance of place of articulation.

Comparison of Phone Frequencies by Nasal Contrast



As for the oral/nasal stop contrast, nasal stops are more frequent than the voiced stops, but less frequent than oral stops altogether. Nonetheless, unlike for place contrasts, nasals are not entirely overwhelmed by their oral counterparts.

There are several ways to explain this pattern. It is possible that the lesser use of the place contrast is due to coronals being present in a few high frequency words in Dutch. The top ten most frequent words are shown in Table 4.1, all with frequency greater than 5,000:

word	frequency
fan	5519
is	6497
tə	7550
m	7608
jə	7898
ɪk	8402
di	8632
dat	8806
ən	8836
ət	9155

TABLE 4.1: Frequent Dutch Words

Sure enough, most of these words contain coronals. It could be, then, that the high frequency of coronals is due to just a few words that contain them. However, type frequencies - the frequency of a phone in the lexicon - are not likely to matter to an infant who does not yet have a lexicon, as would be the case for an infant just learning to segment speech.

It is not surprising that frequency of phones, both in isolation and in ngrams, should affect the importance of that phone and the ngrams containing it as a cue in segmentation. While these frequencies are probably a factor in why neutralization of some contrasts has less effect than others, they do not explain why place contrasts are so skewed toward a particular value in the first place. It could be an accidental fact of Dutch, or, alternatively, given that place contrasts are difficult to perceive, this may be a cross-linguistic tendency. A number of studies, as discussed in Chapter 1, suggest that it is the latter (Mielke, 2003a; Steriade, 2001; Ohala, 1981).

4.2 Perception as a driving force of phonology

The conclusion that there is a correlation between the perceptual salience of a contrast and the usefulness of that contrast in word segmentation would fit in well with the literature on perception as a driving force for phonological change and/or alternations.

Perception has long been argued to shape phonological grammar. That perceptual salience can drive alternations and changes either synchronically (Steriade, 2001), diachronically (Mielke, 2003a; Ohala, 1981), or both, suggests that the more salient a contrast is perceptually, the more likely it is to be crucial for the organization of the language. Less salient contrasts are more likely to be weeded out over time. This predicts that the contrasts that are most crucial for the phonological organization of a language are those that are easiest to perceive. That place of articulation should be less useful for cueing word boundaries would fit in well with this literature.

The alternative point of view, of course, is that contrasts are perceptually salient entirely because they are contrastive. Degree of salience depends not on any inherent quality of a contrast, but entirely on whether it is present or absent in a language. This viewpoint is not born out by experimental study, which shows that certain contrasts, such as place of articulation, are more susceptible to misperception under noise, even when they are present in a language (Cutler et al., 2004; Hura et al., 1992; Miller & Nicely, 1955). A study by Mielke (2003b) tests this directly, by testing the perceptual salience of h in pre-vocalic vs. pre-consonantal position for English speakers (who have h pre-vocalically), French speakers (who have no h), and Turkish and Arabic speakers (who have h in many

environments). He found that for all speakers, [h] was significantly more perceptible when either preceding or following a sonorant than when either preceding or following a voiceless obstruent. Further, the difference between these two environments was greater when [h] was before a consonant than when after a consonant. While English and French speakers were overall less sensitive to [h] than Turkish and Arabic speakers, within a group, the environments that gave the most or least trouble for speakers was the same for each group. Hence, all groups, regardless of native language, showed the same pattern of perceptual difficulty (ie, the same environments caused difficulty for all groups), but differed in the degree of difficulty, depending on the native language phonotactics.

In short, while native phonotactics influence speech perception, inherent acoustic salience may also influence the structure of a phonemic inventory, a phonological grammar, and perhaps also frequency of particular phones within a language. If this is correct, one would expect the results found in this thesis.

4.3 Conclusion and Future Directions

I have concluded that there is a correlation between perceptual salience of a contrast and the usefulness of that contrast in word segmentation in Dutch. Experiment 1 shows this to be the case, by training a supervised learner on input containing various neutralizations and then comparing performance. Experiment 2 repeats the experiment, but instead with the unsupervised learners StaGe and the O/E models, to see how the results compare to Experiment 1 in a more realistic model of infant learning. Again, neutralization of the place of articulation contrast results in a smaller change in performance than the more salient voice contrast.

While this may be the case for Dutch, this need not be true for other languages. Hence, the most readily apparent way to make this conclusion more robust is to test other languages in like manner. It would be especially interesting to test languages which do not allow the complex consonant clusters that are allowed in Dutch (ie, "CV" languages), as it is possible that these languages do not provide the same phonotactic cues that a language such as Dutch provides. Further, if all consonants are always preceded and followed by a vowel, the relative perceptual salience of different contrasts may reach a threshold, since vowels maximize the perceptibility of consonantal features.

The implications of a correlation between perceptual salience and usefulness in word segmentation are broad. As already mentioned in Section 4.2, it fits in well with the literature on perception as a driving force of the phonological organization of a language. The other implication of the results of these experiments is that word segmentation is not

dependent on perfectly transmitted input, nor on the learner having a fully adult perception of phones. The learner can get a good head start on learning word boundaries, and bootstrap into word learning, without a full-fledged, adult-like system of phonological contrasts. This is especially true if the first contrasts the infant learns are the more perceptually salient ones, which is likely the case anyway. In short, while it is often supposed that language acquisition is a complex process that miraculously occurs in the absence of much information, here is shown that a lot can be learned with even less information than is generally assumed.

Appendix A

Terminology

Below is some of the terminology as used in this paper, as well as the abbreviations.

biphone A pair of phones. For example, [bp]

contrast Contrast is a central concept to phonology. Whether or not an audible difference cues a difference in meaning is the standard test for whether or not it qualifies as a distinctive feature in a particular language - that is, whether or not a phonetic difference is phonemic. Here, I use the term somewhat more loosely, since infants may or may not understand the concept of phonemes.

Corpus Gesproken Nederlands / Spoken Dutch Corpus The Spoken Dutch Corpus (CGN) (Oostdijk, 1999) is a phonemically transcribed spoken corpus of Dutch. About 10 percent of the corpus (the Flemish portion of the corpus was excluded) was used in the simulations in Experiments 1, 2, and 3, as was the case for the simulations done in Adriaans & Kager (2010). The portion used contained 78,080 utterances, 660,424 words, a fairly large sample size. The corpus was obtained by first using automatic transcription procedures, followed by correcting the corpus by hand to include the variation and some of the detail present in the corresponding speech (Oostdijk, 1999). Due to this method of transcription, the corpus contains more variation than most corpora, including assimilations across word boundaries, some epenthetic stops, and variations in word pronunciation.

For the purposes of accurately representing the word segmentation task, the word boundaries, represented in the corpus by spaces, were removed, creating a representation of continuous speech. Utterance boundaries, which contain audible pauses in speech, remained in the continuous speech version of the corpus, represented by a newline. Thus, two utterances as represented in the corpus would look like

mjam@rtIsweI@n@@@nOp@vIAgw@tIsweI@nIENt@
fert@xsEntimet

The corresponding, segmented utterances would look like

mja m@r t Is weI @n @ @n Op@vIAg w@ t Is weI @n IENt@
fert@x sEntimet

with the spaces representing word boundaries, and newlines representing utterance boundaries.

Expected Expected is the expected frequency of a sequence. It is calculated from $E = f([xY]) * f([Xy]) * n$

Observed Observed is the frequency count of a sequence

Observed/Expected (O/E) model The statistical component of the StaGe model.

It learns the O/E values of biphones in a training phase, and then segments these biphones based on the threshold 1.0. A biphone with an O/E value < 1.0 will be segmented, while a biphone with an O/E value > 1.0 will be left intact.

Observed/Expected (O/E) value A bi-directional probability statistic that expresses the probability of a sequence co-occurring. It is calculated from the observed value divided by the expected value.

phone A generic term for a segment. While based on the idea of phonetic categories, unlike phoneme, or phonetic category, it has no theoretical implications.

Statistical learning and Generalization (StaGe) model A threshold-based learning model that combines the statistical learning component with the generalization component, as described in [Adriaans & Kager \(2010\)](#). Phonotactic learning consists of forming constraints, with thresholds 0.5 and 2.0 O/E value for forming segmentation and contiguity constraints respectively. Constraints are ranked according to the average of the expected values of the biphones contained in them. Segmentation decisions are made based on the ranking of the constraints, as in Optimality Theory.

Transitional Probability (TP) A uni-directional probability statistic that expresses the probability of y given x preceding y .

utterance An utterance, in contrast to words, has an audible pause at an utterance boundary. The corpus used does contain utterance boundaries, separated by a newline.

word segmentation Word segmentation involves finding the boundaries in continuous speech between phrases and words.

Appendix B

Differently Segmented Biphones

Experiment 1: Differently Segmented Biphones

Biphone	Neutralized	Control	Total Count	Word Internal Count	Word Boundary Count
ne	0.5	1	7866	7079	787
tr	0.5	1	5428	4806	622
ki	0	1	545	449	96
eb	0	1	450	208	242
ɛf	1	0	839	741	98
ɣn	0.5	0	2774	2634	140
my	0	1	343	322	21
gl	0.5	1	101	32	69
ŋe	0.5	1	24	1	23
nu	0	1	607	561	46
ɣʒ	0	1	24	2	22
az	1	0	1878	1690	188
ŋa	0.5	1	91	5	86
ij	0.5	0	69	66	3
br	0.5	1	2123	2113	10
ŋh	0.5	1	216	6	210
øt	0	1	99	90	9
ən	0.5	0	30556	23988	6568
ort	0	1	9	9	0
em	0.5	0	3015	2040	975
dʌu	1	0	29	23	6

Biphone	Neutralized	Control	Total Count	Word Internal Count	Word Boundary Count
mo	0.5	1	1919	1745	174
aŋ	0.5	0	531	520	11
eŋ	0.5	0	324	322	2
gr	0.5	1	100	79	21
mv	0.5	1	237	22	215
mt	0.5	1	1879	1320	559
ky	0	1	292	232	60
gi	0	1	25	10	15
ŋɔ	0.5	1	227	14	213
ms	0.5	1	1147	822	325
ŋo	0.5	1	100	1	99
mi	0	1	856	811	45
mɪ	0.5	1	2471	1933	538
ŋs	0.5	1	751	632	119
mã:	0.5	1	5	5	0
øm	1	0	27	16	11
ʏs	0	1	4541	4407	134
ng	0.5	1	38	3	35
ʏɑ	0	1	289	199	90
ʏf	0	1	93	43	50
ʏp	0	1	215	211	4

Biphone	Neutralized	Control	Total Count	Word Internal Count	Word Boundary Count
nø	0.5	1	228	161	67
na	0	1	6839	6054	785
mh	0.5	1	1349	975	374
mʌu	0	1	53	46	7
ɲɪ	0.5	1	407	86	321
um	0.5	0	787	533	254
no	0.5	1	3096	1652	1444
nh	0.5	1	4289	293	3996
øn	1	0	128	120	8
mə	0.5	1	10685	9540	1145
nœy	0.5	1	540	178	362
mø	0.5	1	28	25	3
ɲʃ	0.5	1	209	136	73
ny	0.5	1	292	88	204
in	0.5	0	4636	3782	854
əm	0.5	0	16850	8201	8649
ɲɛ	0.5	1	232	4	228
ɣɲ	0.5	0	139	139	0
ʌum	0.5	0	311	10	301
ɲu	0	1	3	1	2
me	0.5	0	4732	4617	115
om	0.5	0	3181	2587	594
ɲʏ	0.5	1	765	527	238
ɪp	0	1	447	434	13
ɲg	0.5	1	665	523	142
ɛt	0	1	35	34	1
yɲ	0.5	0	2	2	0
ɔ:n	0.5	0	12	11	1
lg	0.5	1	43	30	13
am	0.5	0	4024	2515	1509
na	0.5	1	4635	1895	2740
rm	1	0.5	2454	1055	1399
nk	0.5	1	1282	57	1225
ɔd	0	1	502	354	148
ym	0.5	0	331	202	129
bw	0.5	1	209	21	188
nv	0.5	1	2099	309	1790
nz	0.5	1	3703	1150	2553
mj	0.5	1	583	129	454
ɲa	0	1	75	12	63
ɛin	0.5	0	273	266	7

Biphone	Neutralized	Control	Total Count	Word Internal Count	Word Boundary Count
nr	0.5	1	5453	1998	3455
ɲə	0.5	1	3325	3088	237
mg	0.5	1	8	1	7
ɲei	0.5	1	14	5	9
bɒu	1	0	299	299	0
su	0	1	336	323	13
ɛz	1	0	306	229	77
lm	1	0.5	1921	579	1342
pɒu	1	0	68	63	5
eʒ	1	0	33	28	5
on	0.5	0	4291	3741	550
nɛ:	0.5	1	17	16	1
øʂ	0	1	161	149	12
nb	0.5	1	1233	87	1146
ku	0	1	285	268	17
kɒu	1	0	149	137	12
kh	1	0	2206	264	1942
mz	0.5	1	345	78	267
dw	0.5	1	1086	198	888
kw	0.5	0	3567	1173	2394
ɣt	0	1	242	234	8
ih	1	0	1000	52	948
mɛ	0.5	0	7435	7114	321
əɲ	0.5	0	1082	1074	8
mf	0.5	1	358	41	317
ɲb	0.5	1	187	51	136
pr	0.5	1	4322	4271	51
lb	0	1	109	45	64
rp	0.5	1	681	369	312
nf	0.5	1	3118	456	2662
ɛ:m	0.5	0	2	2	0
ɛ:n	0.5	0	33	33	0
np	0.5	1	653	33	620
ag	1	0	195	160	35
nt	0.5	1	18923	15264	3659
ɲʃ	0.5	1	20	18	2
kl	0.5	1	2580	2296	284
mœy	0.5	1	113	59	54
mp	0.5	1	2293	1132	1161
mw	0.5	1	542	43	499

Biphone	Neutralized	Control	Total Count	Word Internal Count	Word Boundary Count
ɔt	0	1	1643	1608	35
εg	1	0	152	143	9
mε:	0.5	1	11	11	0
my	0.5	1	114	84	30
ɲv	0.5	1	388	10	378
nə	0.5	1	18810	7766	11044
pw	0.5	1	218	24	194
ld	0	1	1259	1141	118
dr	0.5	1	2777	2641	136
pl	0.5	1	1881	1822	59
nɔ	0.5	1	6589	4324	2265
yz	0	1	969	957	12
ɪʒ	0	1	8	3	5
ɔg	0	1	80	79	1
sk	0	1	1757	834	923
nε	0.5	1	4051	1659	2392
ɣv	1	0	7	6	1
ɣb	1	0	108	104	4
ɲk	0.5	1	5240	3035	2205
an	0.5	0	8959	7613	1346
œym	0.5	0	300	295	5
yn	0.5	0	542	270	272
gw	0.5	0	676	32	644
ɲj	0.5	1	63	1	62
ɲœy	0.5	1	32	0	32
oɲ	0.5	0	144	140	4
ɲd	0.5	1	579	51	528
ɔ:m	0.5	0	2	1	1
mεi	0.5	0	2122	2094	28
kr	0.5	1	2414	2295	119
lp	0.5	1	504	245	259
ɲf	0.5	1	418	22	396
ʌun	0.5	0	385	49	336
nẽ	0.5	1	5	5	0
tl	0.5	1	1336	271	1065
εv	1	0	113	73	40
tj	1	0	6359	3365	2994
nd	0.5	0	27522	12333	15189
im	0.5	0	2515	1013	1502
mk	0.5	1	273	78	195
ɲi	0	1	18	0	18
en	0.5	0	6577	5567	1010
pj	1	0	1163	405	758

Biphone	Neutralized	Control	Total Count	Word Internal Count	Word Boundary Count
un	0.5	0	3258	3052	206
ŋt	0.5	1	540	277	263
dl	0.5	1	915	266	649
md	0.5	1	2199	873	1326
tw	0.5	0	8048	2938	5110
nɛi	0.5	1	536	199	337
rg	0.5	1	151	111	40
qm	0.5	0	4	2	2
ŋz	0.5	1	207	73	134
mɔ	0.5	1	1318	955	363
uŋ	0.5	0	77	76	1
ŋp	0.5	1	52	25	27
ɛb	1	0	3233	3155	78
ɣm	0.5	0	437	413	24
dɛi	0	1	143	113	30
ma	0.5	1	5032	4766	266
mb	0.5	1	3815	839	2976
œyx	1	0	100	94	6
ŋy	0	1	3	0	3
ns	0.5	1	9086	5862	3224
øb	0	1	23	16	7
ɛd	1	0	2122	1782	340
ad	1	0	5715	5325	390
bl	0.5	1	2074	2047	27

Experiment 2: Differently Segmented Biphones

Biphone	Neutralized	Control	Total Count	Word Internal Count	Word Boundary Count
su	0	1	336	323	13
sp	1	0	3692	3279	413
rd	1	0	1259	1141	118
yg	0	1	102	101	1
ɛf	1	0	839	741	98
ɛz	1	0	306	229	77
e:b	1	0	1	1	0
œyp	1	0	51	50	1
ət	1	0	20020	14458	5562
yd	1	0	645	302	343
ɣʒ	0	1	24	2	22
rʒ	0.5	0	77	70	7
ku	0	1	285	268	17
ʌup	1	0	68	25	43
kh	1	0	2206	264	1942
ez	1	0	1344	1204	140
xw	0.5	1	1712	678	1034
ŋh	0.5	1	216	6	210
dw	0.5	1	1086	198	888
—z	1	0	101	99	2
st	1	0	20821	17255	3566
bj	0.5	1	155	26	129
ih	1	0	1000	52	948
ɛt	1	0	6988	6891	97
əs	1	0	16212	9400	6812
œyb	1	0	37	31	6
γw	0.5	1	410	65	345
ap	1	0	949	600	349
rp	0.5	1	681	369	312
ʌf	1	0	1	1	0
nt	0	1	18923	15264	3659
əf	1	0	7975	2179	5796
qp	1	0	2	0	2
ky	0	1	292	232	60
tʌu	0	1	94	34	60
ep	1	0	483	379	104
əd	1	0	20763	6343	14420
əb	1	0	8765	2715	6050
lʒ	0	1	8	3	5
ɛs	1	0	2538	2357	181
yp	1	0	118	96	22

Biphone	Neutralized	Control	Total Count	Word Internal Count	Word Boundary Count
yb	1	0	280	143	137
od	1	0	2238	1144	1094
ɔg	0	1	80	79	1
ɣf	0	1	93	43	50
ɣv	1	0	7	6	1
ɣb	1	0	108	104	4
ɪɣ	0	1	357	348	9
gj	0.5	1	74	10	64
mh	0.5	1	1349	975	374
—v	1	0	55	51	4
oz	1	0	573	366	207
av	0	1	340	265	75
es	1	0	2540	2271	269
sw	0.5	1	2124	381	1743
ub	1	0	168	87	81
ʌub	1	0	203	41	162
nh	0.5	1	4289	293	3996
lp	0.5	1	504	245	259
os	1	0	1292	1005	287
yt	1	0	801	662	139
ɔf	0	1	47	41	6
—p	1	0	15	12	3
ab	0	1	601	337	264
et	1	0	5005	4657	348
ɛv	1	0	113	73	40
əp	1	0	4011	1210	2801
tj	0.5	0	6359	3365	2994
ab	1	0	849	342	507
pj	0.5	0	1163	405	758
ed	1	0	4476	3489	987
yh	1	0	239	40	199
ɔ:d	1	0	4	4	0
—f	1	0	43	35	8
tw	0.5	0	8048	2938	5110
up	1	0	488	477	11
hɪ	1	0	212	206	6
rg	0.5	1	151	111	40
ad	1	0	5494	1818	3676
rɪ	0.5	1	74	56	18
sx	1	0	6663	4689	1974
at	1	0	8341	7444	897
—d	1	0	47	21	26

Biphone	Neutralized	Control	Total Count	Word Internal Count	Word Boundary Count
zw	0.5	1	965	351	614
kj	0.5	1	1091	511	580
dɛi	0	1	143	113	30
əz	1	0	8201	3302	4899
əv	1	0	6036	1849	4187
dj	0.5	1	591	299	292
uh	1	0	192	21	171
lk	1	0.5	1453	985	468
ɛd	1	0	2122	1782	340
ot	1	0	2823	2628	195

Bibliography

- Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, *62*(3), 311–331.
- Aslin, R. N., & Newport, E. L. (2004). Learning at a distance i. statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.
- Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, (pp. 321–324).
- Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*, 568–591.
- Blanchard, D., & Heinz, J. (2008). Improving word segmentation by simultaneously learning phonotactics. In *12th CoNLL*, (pp. 65–72).
- Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*, 71–105.
- Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation distributional regularity and phonotactic constraints are useful for segmentation distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*.
- Cairns, P., Shillcock, R. C., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: a bottom up corpusbased approach to speech segmentation. *Cognitive Psychology*, *33*, 111–153.
- Cho, T., McQueen, J., & Fox, E. A. (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics*, *35*, 210–243.

- Coetzee, A. (2005). The obligatory contour principle in the perception of English. the obligatory contour principle in the perception of English. the obligatory contour principle in the perception of English. In S. Frota, M. Vigario, & M. J. Freitas (Eds.) *Prosodies*, (pp. 223–245). Berlin: Mouton de Gruyter.
- Cole, R., & Jakimik, J. (1980). *A Model of Speech Perception*, (pp. 136–163). Lawrence Erlbaum Associates, Hillsdale, NJ.
- Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In *3rd Meeting of the ACL Special Interest Group in Computational Phonology: Proceedings of the Workshop*, (pp. 49–56). Association for Computational Linguistics, Somerset NJ.
- Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation. In G. T. M. Altmann (Ed.) *Cognitive models of speech processing*. Cambridge, MA: MIT Press.
- Cutler, A. (1994). Segmentation problems, rhythmic solutions. *Lingua*, *92*, 81–104.
- Cutler, A., & Carter, D. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*(3-4), 133–142.
- Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human perception and performance*, *14*(1), 113–121.
- Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America*, *116*(6), 3668–3678.
- Daland, R. (2009). *Word Segmentation, Word Recognition, and Word Learning: A Computational Model of First Language Acquisition*. Ph.D. thesis, Northwestern University.
- Daland, R., & Pierrehumbert, J. (submitted). Learnability of diphone-based segmentation. To appear.
- Dupoux, E., Pallier, C., Kakehi, K., & Mehler, J. (2001). New evidence for prelexical phonological processing in word recognition. *Language and Cognitive Processes*, *5*(16), 491–505.
- Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*, 3728–3740.
- Frisch, S. A., Pierrehumbert, J., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory*, *22*, 179–228.

- Goddign, S., & Binnenpoorte, D. (2003). Assessing manually corrected broad phonetic transcriptions in the spoken dutch corpus. In *Proceedings of the 15th International Congress of Phonetic Sciences*, (pp. 1361–1364).
- Goldwater, S. (2007). *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.
- Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112 (2009) 21–54, 21–54.
- Hay, J., & Baayen, H. (2004). Phonotactics, parsing and productivity. *Italian Journal of Linguistics*, 15(1), 99–130.
- Hayes, B. (1999). Phonetically driven phonology: the role of optimality theory and inductive grounding. *Functionalism and Formalism in Linguistics: General papers*, (p. 243).
- Hayes, B., & Wilson, C. (2008). A maximum entropy model phonotactics and phonotactic learning. *Linguistic Inquiry*, 33(3), 379–440.
- Hura, S. L., Lindblom, B., & Diehl, R. L. (1992). On the role of perception in shaping phonological assimilation rules. *Language and Speech*, 35, 59.
- Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44(4), 548–567.
- Jusczyk, P., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in english-learning infants. *Cognitive Psychology*, 39(3-4), 159–207.
- Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29.
- Kohler (1990). Segmental reduction in connected speech in german: Phonological facts and phonetic explanations. In W. J. Hardcastle, & A. Marchal (Eds.) *Speech production and speech modelling*. Dordrecht: Kluwer Academic Publishers,.
- Lin, Y. (2005). *Learning Features and Segments from Waveforms: A Statistical Model of Early Phonological Acquisition*. Ph.D. thesis, University of California Los Angeles.
- Mattys, S., Jusczyk, P., Luce, P., & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38(4), 465–494.
- Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78, 91–121.

- Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(B101-B111).
- McQueen, J. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, *39*, 21–46.
- Mielke, J. (2003a). The diachronic influence of perception: Experimental evidence from turkish. *Proceedings of BLS*, *29*.
- Mielke, J. (2003b). The interplay of speech perception and phonology: Experimental evidence from turkish. *Phonetica*, *60*, 208–229.
- Miller, G., & Nicely, P. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *27*, 338.
- Ohala, J. (1981). The listener as a source of sound change. *Parasession on language and behavior*, (pp. 178–203).
- Oostdijk, N. (1999). Building a corpus of spoken dutch.
- Peters, A. (1983). *The Units of Language Acquisition, Monographs in Applied Psycholinguistics*. Cambridge University Press.
- Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology-Human Perception and Performance*, *20*(2), 421–435.
- Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.
- Smits, R., Warner, N., McQueen, J., & Cutler, A. (2003). Unfolding of phonetic information over time: A database of dutch diphone perception. *The Journal of the Acoustical Society of America*, *113*, 563.
- Steriade, D. (2001). Directional asymmetries in place assimilation. In E. Hume, & K. Johnson (Eds.) *The role of speech perception in phonology*, (pp. 219–250). San Diego: Academic Press.
- Tesar, B., & Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, *29*(2), 229–268.
- Thiessen, E. D., & Saffran, J. R. (2003). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, *3*, 73–100.

-
- Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics*, *27*(3), 352–372.
- Werker, J. F., & Tees, R. C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology*, *37*, 278–286.