Journal of Memory and Language 62 (2010) 311–331

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Adding generalization to statistical learning: The induction of phonotactics from continuous speech

Frans Adriaans *, René Kager

Utrecht Institute of Linguistics OTS, Utrecht University, The Netherlands

ARTICLE INFO

Article history:
Received 17 October 2008
revision received 16 October 2009
Available online 29 December 2009

Keywords:
Phonotactics
Speech segmentation
Statistical learning
Phonological features
Constraint induction
Computational modeling

ABSTRACT

Emerging phonotactic knowledge facilitates the development of the mental lexicon, as demonstrated by studies showing that infants use the phonotactic patterns of their native language to extract words from continuous speech. The present study provides a computational account of how infants might induce phonotactics from their immediate language environment, which consists of unsegmented speech. Our model, StaGe, implements two learning mechanisms that are available to infant language learners: statistical learning and generalization. StaGe constructs phonotactic generalizations on the basis of statistically learned biphone constraints. In a series of computer simulations, we show that such generalizations improve the segmentation performance of the learner, as compared to models that rely solely on statistical learning. Our study thus provides an explicit proposal for a combined role of statistical learning and generalization in the induction of phonotactics by infants. Furthermore, our simulations demonstrate a previously unexplored potential role for phonotactic generalizations in speech segmentation.

During the second half of the first year of life, infants are starting to build up a vocabulary of words. One of the major tasks that infants face in learning the words of their native language is to extract sound sequences from the speech stream to which meaning should be attached at some later point in time. Since the speech input to the infant typically contains no audible pauses between words, word learning crucially involves the segmentation of utterances of continuous speech into discrete, word-sized units. Several types of cues have been demonstrated to guide infants' search for words in continuous speech, thereby facilitating the development of the mental lexicon.

Infants aged between 6 and 9 months are sensitive to native language patterns of stressed and unstressed syllables (Jusczyk, Cutler, & Redanz, 1993; Morgan & Saffran, 1995), while 7.5-month-old infants use metrical patterns as a cue for segmenting word-like units from speech (Jusczyk, Houston, & Newsome, 1999). Fine-grained acoustic cues are attended to by infants ranging in age between 5 months (co-articulation between segments: Fowler, Best, & McRoberts, 1990) and 10.5 months (context-sensitive allophones: Jusczyk, Hohne, & Bauman, 1999), while such cues are used for speech segmentation by infants varying in age between 8 months (co-articulation: Johnson & Jusczyk, 2001) and 10.5 months (context-sensitive allophones: Jusczyk et al., 1999). In addition to metrical and acoustic cues, 9-month-old infants are sensitive to the phonotactic patterns of the native language (Friederici & Wessels, 1993; Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; Jusczyk, Luce, & Charles-Luce, 1994), and use phonotactic cues for segmentation (Mattys, Jusczyk, Luce, & Morgan, 1999; Mattys & Jusczyk, 2001).

Jusczyk, Friederici et al. (1993) show that 9-month-old infants listen longer to words from their native language than to nonnative words, which typically violate the phonotactic constraints of the native language. Similar results were found for phonotactically legal versus illegal nonwords (Friederici & Wessels, 1993). Infants' sensitivity to

* Corresponding author. Address: Utrecht Institute of Linguistics OTS, Janskerkhof 13, 3512 BL Utrecht, The Netherlands. Fax: +31 30 253 6406.
E-mail address: f.w.adriaans@uu.nl (F. Adriaans).

0749-596X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2009.11.007

phonotactics appears to be more fine-grained than the ability to distinguish between legal and illegal sound sequences. Jusczyk et al. (1994) show that 9-month-old infants are sensitive to phonotactic probabilities. Infants listened longer to lists of high-probability nonwords than to lists of low-probability nonwords. In contrast, 6-month-old infants did not show this sensitivity. This finding indicates that, in the period between the first 6 and 9 months of life, infants start to learn about the distributions of sound patterns of their native language from experience with speech input. Mattys and Jusczyk (2001) show that 9-month-old infants are able to use such probabilistic phonotactic cues to segment words from continuous speech. Infants listened longer to words which were embedded in between-word consonant clusters (i.e., consonant clusters that occur more frequently between words than within words) than to words which were embedded in within-word consonant clusters (i.e., consonant clusters that occur more frequently within words than between words).

Although these studies clearly indicate a role for phonotactics in the segmentation of continuous speech, they leave open the question of how infants might have learned the phonotactic patterns of their native language. If infants have probabilistic knowledge of phonotactics when they are only beginning to develop a mental lexicon, then it appears that they are able to induce phonotactic patterns from their immediate language environment, which consists of continuous speech. Moreover, if infants use phonotactics to bootstrap into word learning, then the infant *has to* learn phonotactic constraints from continuous speech. Similar views have been adopted in various studies that argue that segmentation cues need to be derived from unsegmented input (e.g., Brent & Cartwright, 1996; Cairns, Shillcock, Chater, & Levy, 1997; Perruchet & Vinter, 1998; Swingley, 2005).

Phonotactics is typically defined as a set of constraints on the sound structure of words. Indeed, such constraints have been shown to affect speech segmentation (McQueen, 1998; Weber & Cutler, 2006). However, this does not imply that knowledge of words is required for the learning of phonotactic constraints. The crucial question, then, is *how* infants learn phonotactic constraints from continuous speech. The present study addresses this question from a computational angle by investigating how infant learning mechanisms might interact in the learning of phonotactic constraints from continuous speech.

Learning mechanisms

Saffran, Aslin, and Newport (1996) showed that infants are sensitive to transitional probabilities of adjacent syllables when learning artificial words from a stream of continuous speech. Since a low transitional probability between syllables indicates a likely word boundary, statistical learning helps in segmenting word-like units from continuous speech input. Several sources of evidence indicate that statistical learning may also be involved in infants' learning of phonotactics. First, 6- and 8-month-olds are able to discriminate between phonetic categories (Maye, Werker, & Gerken, 2002). While infants may not yet have acquired the full segment inventory, the ability to perceive at least some phonetic categories is a prerequisite for the learning of co-occurrence probabilities of such categories. Second, 9-month-old infants are sensitive to native language probabilistic phonotactics (Jusczyk et al., 1994; Mattys & Jusczyk, 2001), which shows that infants are capable of learning segment co-occurrence probabilities. This ability has also been demonstrated by White, Peperkamp, Kirk, and Morgan (2008), who found that 8.5-month-old infants' responses in learning voicing alternations were driven by segment transitional probabilities.

There appears to be a role for a second learning mechanism in phonotactic acquisition, requiring a different form of computation. This learning mechanism allows the infant to generalize over the observed input to new, unobserved instances through the construction of abstract representations. The capacity of infants to construct such generalizations has been demonstrated for the learning of phonetic categories (e.g., Werker & Tees, 1984; Maye et al., 2002; Maye, Weiss, & Aslin, 2008), phonotactic patterns (e.g., Chambers, Onishi, & Fisher, 2003; Saffran & Thiessen, 2003), and artificial grammars (e.g., Gomez & Gerken, 1999; Marcus, Vijayan, Rao, & Vishton, 1999).

Saffran and Thiessen (2003) showed that 9-month-old infants can induce phonotactic patterns that are more general than the occurrence patterns of the specific phonological segments to which they were exposed. During a pattern induction phase, the infant was familiarized with the phonotactic regularity. Familiarization was followed by a segmentation phase, in which infants could segment novel words from a continuous speech stream by employing the phonotactic pattern to which they had been familiarized. Finally, the test phase served to determine whether the infant indeed was able to distinguish novel test items which conformed to the phonotactic pattern from test items which did not conform to the pattern. Infants acquired a phonotactic generalization about the positional restrictions on voiced and voiceless stops after a brief training period. In contrast, they could not learn patterns of segments which were not phonetically similar. These results indicate that there is more to phonotactic acquisition than the learning of constraints on the co-occurrences of specific segments. It seems that infants are able to abstract over the similarity between segments to construct phonotactic generalizations.

Evidence for infants' sensitivity to phonetic similarity (voicing, manner of articulation, etc.) has been reported in various studies (e.g., Jusczyk, Goodman, & Baumann, 1999; White et al., 2008; Maye et al., 2008) and phonetic similarity thus appears to play a role in generalization. Little is known, however, about how phonotactic generalizations are represented by infants. Specifically, the question is whether such representations involve abstract features. Several recent studies have indeed argued that infants form generalizations that are abstract at the level of the feature, both in the discrimination of sound contrasts (Maye et al., 2008), and in the learning of phonotactic patterns (Cristià & Seidl, 2008; Seidl & Buckley, 2005). While such features could either be innate or learned, it should be noted that abstract phonological features have acoustic and perceptual correlates in the speech signal (e.g., Stevens, 2002), and attempts

have been made to induce abstract features from raw acoustic input data (Lin & Mielke, 2008).

In sum, statistical learning allows the learner to accumulate frequency data over observed input. A second mechanism, generalization, allows the learner to abstract away from the observed input, leading to the formation of categories, patterns, and grammars. While the importance of both learning mechanisms for language acquisition has been widely acknowledged (Gomez & Gerken, 1999; Marcus et al., 1999; Toro, Nespor, Mehler, & Bonatti, 2008; White et al., 2008), surprisingly little is known about how these two mechanisms, statistical learning and generalization, interact. For example, how can statistical learning provide a basis for the construction of generalizations? How do generalizations affect the probabilistic knowledge of the learner? Explicit descriptions of such interactions would greatly enhance our understanding of infant language acquisition.

In addition, while infants' capacity to use probabilistic phonotactics in speech segmentation has been demonstrated (Mattys & Jusczyk, 2001), it is not clear whether infants also use phonotactic generalizations to discover words in continuous speech. Although the study by Saffran and Thiessen explores this possibility, the authors themselves mention that the infants may have applied the patterns that were induced during familiarization to the test items directly, i.e. regardless of the segmentation task (Saffran & Thiessen, 2003, p. 487). Infants' use of phonotactic generalizations in speech segmentation thus remains to be demonstrated.

The goal of the present paper is two-fold. First, we aim to give a computational account of the induction of phonotactics from continuous speech, using the mechanisms of statistical learning and generalization. Second, since the role of phonotactic generalizations in speech segmentation has not been addressed in infant studies, we take a first step in determining the potential use of such generalizations. Through computational modeling, we explore whether adding a generalization mechanism to statistical learning would improve the learner's ability to detect word boundaries in continuous speech. We thereby hope to offer insight into the learning mechanisms and segmentation strategies that play a role in infant phonotactic learning.

Models of speech segmentation and phonotactic learning

Computational models of speech segmentation are typically trained and tested on transcribed utterances of continuous speech. The task of the model is either to learn a lexicon directly, through the extraction of word-like units, or to learn to predict when a word boundary should be inserted in the speech stream. Here we will discuss the latter task, focussing on models that learn phonotactics in order to detect word boundaries in continuous speech. (For more general overviews of computational models of segmentation, see Brent, 1999b; Batchelder, 2002.) Various segmentation models have been proposed that make use of either phonotactics based on utterance boundaries (Brent & Cartwright, 1996), or phonotactics based on sequence probabilities (e.g., Cairns et al., 1997).

Brent and Cartwright (1996) propose that phonotactic constraints can be learned through inspection of consonant clusters that appear at utterance boundaries. The model is based on the observation that clusters at the edges of utterances are necessarily also allowed at the edges of words. Their segmentation model evaluates candidate utterance parsings using a function based on Minimum Representation Length (MRL). The phonotactic constraints act as a filter to eliminate utterance parsings which would produce phonotactically ill-formed words. A disadvantage of this approach is that it assumes categorical phonotactics: a cluster is either allowed or not, based on whether it occurs at least once at an utterance boundary. As a consequence, the approach is rather vulnerable to occurrences of illegal clusters at utterance edges (which may occur as a result of acoustic reductions). In general, categorical phonotactics fails to make a prediction in cases of ambiguous clusters in the speech stream, since such clusters have multiple phonotactically legal interpretations. Moreover, infants have been demonstrated to use phonotactic *probabilities* to segment speech (Mattys & Jusczyk, 2001).

Models that rely on probabilistic phonotactic cues either explicitly implement segment co-occurrence probabilities (e.g., Brent, 1999a; Cairns et al., 1997), or use neural networks to learn statistical dependencies (Cairns et al., 1997; Christiansen, Allen, & Seidenberg, 1998; Elman, 1990). Two different interpretations exist with respect to how co-occurrence probabilities affect speech segmentation (Rytting, 2004). Saffran, Newport, and Aslin (1996) suggest that word boundaries are hypothesized at troughs in transitional probability. That is, a word boundary is inserted when the probability of a bigram is lower than those of its neighboring bigrams. This *trough-based* segmentation strategy thus interprets bigram probabilities using the context in which the bigram occurs. Computational models have shown this interpretation to be effective in segmentation (e.g., Brent, 1999a). A trough-based approach, however, is not capable of extracting unigram words, since such words would require two adjacent local minima (Rytting, 2004; Yang, 2004). The implication is that the learner is unable to discover monosyllabic words in case of syllable-based statistical learning (Yang, 2004), or single-phoneme words in case of segment-based statistical learning (Rytting, 2004).

Studies addressing the role of probabilistic phonotactics in infant speech segmentation (Mattys & Jusczyk, 2001) show that the probability of a bigram can also affect speech segmentation directly, i.e. regardless of neighboring bigrams. The interpretation of probabilities in isolation has been modeled either by inserting boundaries at points of low probability (Cairns et al., 1997; Rytting, 2004), or by clustering at points of high-probability (Swingley, 2005). In both cases the learner needs to use a threshold on the probabilities in order to determine when a boundary should be inserted, or when a cluster should be formed. This *threshold-based* segmentation strategy gives rise to a new problem for the learner: how can the learner determine what this threshold should be? Although the exact value of a statistical threshold remains an open issue, we will argue that a classification of bigrams into functionally distinct categories can be derived from the statistical distribution.

Our study complements previous modeling efforts by adding a generalization component to the statistical learning of phonotactic constraints. The existence of such general ('abstract') constraints is widely accepted in the field of phonology. However, the link to psycholinguistically motivated learning mechanisms, such as statistical learning and generalization, has not been made, and the learning of phonotactic constraints from continuous speech has not been explored. In fact, constraints in linguistic frameworks, such as Optimality Theory (Prince & Smolensky, 1993), are typically assumed to be innate.

Recent work in phonology, however, has focussed on the induction of phonotactic constraints, either from articulatory experience (Hayes, 1999), or from statistical regularities in the lexicon (Hayes & Wilson, 2008). Hayes and Wilson (2008) propose a model in which phonotactic constraints are selected from a space of possible constraints. Constraints are selected according to their accuracy (with respect to the lexicon of the language) and their generality. In this model all logically possible generalizations are *a priori* represented as candidate constraints before the learner has processed any input. In contrast, as we will describe in the next section, our model needs no such *a priori* representations, since it generalizes over statistically learned biphone constraints. That is, the model gradually builds up more general constraints through the processing of input data. Our generalization algorithm is in two respects similar to Minimal Generalization, a model for the learning of past tense inflections (Albright & Hayes, 2002, 2003). First, generalizations are constructed on the basis of similarities in the input. Second, similarity is quantified in terms of shared phonological features. As a result, generalizations affect natural classes, rather than individual segments.

An important difference with earlier models of generalization is that our model is unsupervised. While supervised models evaluate the accuracy of generalizations with respect to the lexicon, we assume that our learner has not yet acquired a lexicon (or that the learner's lexicon is too small to support the learning of phonotactic constraints). In fact, we assume that the learner uses phonotactic generalizations, learned from continuous speech, as a source of knowledge to extract words from the speech stream. Therefore, the learner has no way of determining how good, or useful, the resulting generalizations will be.

Finally, our experiments complement previous computational studies in that our segmentation simulations involve fairly accurate representations of spoken language. While segmentation studies typically use orthographic transcriptions of child-directed speech that are transformed into canonical transcriptions using a phonemic dictionary, the speech transcriptions used in our simulations have been transcribed to a level which includes variations that are typically found in the pronunciation of natural speech, such as acoustic reductions and assimilations.

The model

Modeling phonotactic constraints for speech segmentation requires that we formalize both a learning model, accounting for the constraints, and a segmentation model, explaining how the constraints are used to predict the locations of word boundaries in the speech stream. Before we describe the learning model, StaGe (statistical learning and generalization),[1] we formalize the interpretation of constraints in speech segmentation. Note that no formal models of speech segmentation have been proposed which use abstract constraints.

The OT segmentation model

We propose a modified version of Optimality Theory (OT, Prince & Smolensky, 1993) for regulating interactions between phonotactic constraints in speech segmentation. Optimality Theory is based on the idea that linguistic well-formedness is a relative notion, as no form can possibly meet all demands made by conflicting constraints. The optimal form is one which best satisfies a constraint set, taking into account the relative strengths of the constraints, which is defined by a strict ranking. In order to select the optimal form, a set of candidate forms is first generated. This candidate set contains all logically possible outputs for a given input. All candidates are evaluated by the highest-ranked constraint. Candidates that violate the constraint are eliminated; remaining candidates are passed on to the next highest-ranked constraint. This recursive assessment process goes on until only one candidate remains. This is the optimal form. The optimal form thus incurs minimal violations of the highest-ranked constraints, while taking any number of violations of lower-ranked constraints for granted. This principle of constraint interaction is known as *strict domination*.

The version of OT that we will adopt here retains the assumptions of constraint violability and strict domination, but is otherwise quite different. Whereas OT learners have the task of learning the appropriate ranking for an *a priori* given set of constraints, the task of our learner is: (i) to learn the constraints themselves, as well as (ii) to rank these constraints. Crucially, our version does not employ a universal constraint set (CON) that is given to the learner. Rather, constraints are induced by employing the mechanisms of statistical learning and generalization (cf., Hayes, 1999; Hayes & Wilson, 2008).

The constraint ranking mechanism in our model is also fundamentally different from mechanisms employed in earlier OT learners (in particular, the Constraint Demotion Algorithm; Tesar & Smolensky, 2000, and the Gradual Learning Algorithm; Boersma & Hayes, 2001). Rather than providing the learner with feedback about optimal forms, i.e. segmented utterances, our model assumes unsupervised constraint ranking, since the input to the learner consists exclusively of unsegmented utterances. Each constraint is accompanied by a numerical ranking value, which is inferred by the learner from the statistical distribution, and which expresses the strength of the constraint (cf., Boersma & Hayes, 2001; Boersma, Escudero, & Hayes, 2003).

Finally, note that, although our segmentation model is based on strict constraint domination, one could conceive

[1] The model can be downloaded from: http://www.hum.uu.nl/medewerkers/f.w.adriaans/resources/.

of other constraint interaction mechanisms (such as constraint weighting in Harmonic Grammar; Legendre, Miyata, & Smolensky, 1990) to segment the speech stream. That is, the numerical values of the induced constraints are not committed to the specific interpretation of strict domination. The issue of how the choice of constraint interaction mechanism would affect performance of the model remains open. For the current study, strict domination at least offers a useful mechanism to regulate the interaction between constraints in speech segmentation.

The OT segmentation model is illustrated in Fig. 1. The learner processes utterances of continuous speech through a biphone window. That is, for each biphone in the utterance, the learner needs to decide whether a boundary should be inserted or not. Input to the OT segmentation model consists of biphones, either presented to the model in isolation (*xy*-sequences; see Experiments 1, 2, 4), or embedded in context (*wxyz*-sequences; see Experiment 3). In the latter case, the learner is allowed to inspect not just the current biphone under consideration (*xy*), but also its immediate neighbors (*wx* and *yz*) upon making a segmentation decision for the current biphone.

Segmentation candidates are generated which are possible interpretations of the input sequence. This is implemented using a very simple version of OT's candidate generator (GEN). Each candidate contains either a word boundary at one of the possible boundary locations, or no boundary at all. For biphones in isolation, two candidates are generated: *xy* and *x.y*. For biphones in context, the candidates are: *wxyz*, *w.xyz*, *wx.yz*, and *wxy.z*. Candidates are evaluated using the constraint set, which contains induced, numerically ranked constraints. A boundary is inserted into the speech stream whenever the constraint set favors segmentation of the current biphone under inspection. For the case of biphones in isolation, this means that a boundary is inserted whenever *x.y* is the optimal candidate (i.e., it is preferred over *xy*). For the case of biphones embedded in context, a boundary is inserted whenever *wx.yz* is the optimal candidate (i.e., it is preferred over *wxyz*, *w.xyz*, and *wxy.z*). If multiple candidates remain active after inspection of the constraint set, a winner is chosen at random.

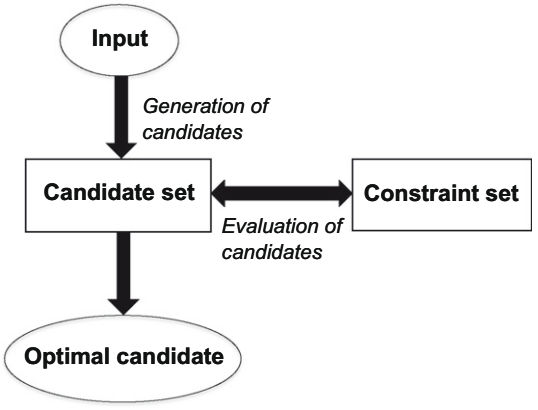

Fig. 1. The OT segmentation model. See main text for details.

The learning model: StaGe

StaGe (statistical learning and generalization) learns specific and abstract phonotactic constraints, as well as ranking values for those constraints, from continuous speech. The model keeps track of biphone probabilities in the input (statistical learning). Biphone probabilities trigger the induction of segment-specific constraints whenever the probabilities reach a specified threshold. These thresholds capture the distinction between high- and low-probability phonotactics. The learner interprets low-probability biphones as likely positions for word boundaries. Conversely, the learner interprets high-probability biphones as unlikely positions for word boundaries. The learner thus infers the likelihood of word boundaries from segment co-occurrence probabilities in continuous speech; a process which we call Frequency-Driven Constraint Induction.

The learner constructs generalizations whenever phonologically similar biphone constraints (of the same phonotactic category, i.e. 'high' or 'low' probability) appear in the constraint set. Similarities are quantified as the number of shared values for phonological features. In case of a single-feature difference between constraints, the learner abstracts over this feature, and adds the generalization to the constraint set; a process which we call Single-Feature Abstraction. The abstract constraints affect sequences of natural classes, rather than sequences of specific segments. In addition to learning constraints, the learner infers ranking values from the statistical distribution. These ranking values determine the strength of the constraint with respect to other constraints in the constraint set.

The general architecture of the model is illustrated in Fig. 2. Below we discuss the three components of StaGe (i.e, statistical learning, Frequency-Driven Constraint Induction, and Single-Feature Abstraction) in more detail.

Statistical learning

Following many psycholinguistic studies StaGe implements a statistical learning mechanism. In our case statistical learning expresses how likely it is that two adjacent segments co-occur in continuous speech. The most well-known formula implementing such statistical dependencies is transitional probability (e.g., Newport and Aslin, 2004; Saffran, Newport et al., 1996):

$$TP(xy) = \frac{Prob(xy)}{\Sigma Prob(xY)} \quad (1)$$

where *Y* can be any segment following *x*. However, as several authors have noted (e.g., Aslin, Saffran, & Newport, 1998; Perruchet & Peereman, 2004), there exists a variety of formulas that can be used to model statistical learning. StaGe implements a slightly different measure of co-occurrence probability, the observed/expected (*O*/*E*) ratio, which has been previously used in studies of phonotactics (e.g., Frisch, Pierrehumbert, & Broe, 2004):

$$\frac{O(xy)}{E(xy)} = \frac{Prob(xy)}{\Sigma Prob(xY) * \Sigma Prob(Xy)} \quad (2)$$

where *Y* can be any segment following *x*, and *X* can be any segment preceding *y*. A third, closely related measure is

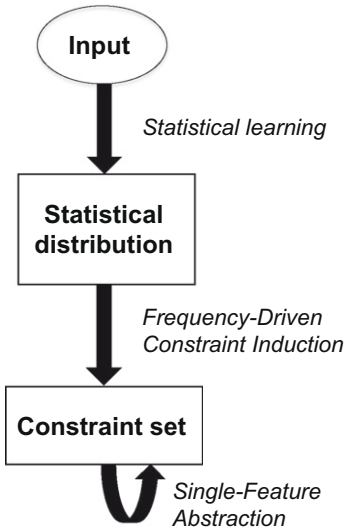

Fig. 2. The architecture of StaGe. Input to the learning model consists of utterances of unsegmented speech. Through the processing of biphones in continuous speech, the learner builds up a statistical distribution (Statistical learning) from which biphone constraints are induced which either favor or restrict the occurrence of a biphone (Frequency-Driven Constraint Induction). Generalizations are constructed whenever phonologically similar constraints appear in the constraint set (Single-Feature Abstraction).

mutual information (MI), which corresponds to the log_2 value of the observed/expected ratio, and which has been used in several computational segmentation studies (Brent, 1999a; Rytting, 2004; Swingley, 2005). The difference between transitional probability and observed/expected ratio (or MI) lies in the directionality of the dependency (Brent, 1999a). Transitional probability expresses a distribution over all elements ($y \in Y$) that *follow* a certain element *x*. In contrast, observed/expected ratio is a bidirectional dependency, expressing the likeliness of two elements *co-occuring*.

Frequency-Driven Constraint Induction

StaGe classifies probabilities obtained through statistical learning into three distinct categories: 'low-probability', 'high probability', and 'neutral probability'. Such a categorization follows from the statistical distribution. The *O*/*E* ratio, for example, expresses whether a biphone is underrepresented or overrepresented in continuous speech. That is, biphones occur either more or less often than would be expected on the basis of the occurrence frequencies of the individual segments.

Biphones that occur more often than expected are considered 'high probability' biphones (overrepresentations) by the learner. In speech segmentation, the learner tries to keep high-probability biphones intact. This is done through the induction of a phonotactic constraint, which states that no boundary should be inserted (a 'contiguity' constraint, cf. McCarthy & Prince, 1995):

Contig-IO(*xy*) 'Sequence *xy* should be preserved.' (3)

In contrast, 'low-probability' biphones (underrepresentations) are likely to contain word boundaries. For this phonotactic category, the learner induces a constraint that favors the insertion of a word boundary (a 'markedness' constraint, stating which sequences are not allowed in the language):

$^{*}xy$ 'Sequence *xy* should not occur.' (4)

Markedness and contiguity constraints thus represent two opposing forces in speech segmentation, both of which are derived directly from the statistical distribution. Contiguity constraints exert pressure towards preservation of high-probability biphones, whereas markedness constraints exert pressure towards the segmentation of low-probability biphones. When interpreted within the framework of OT, a contiguity (or markedness) constraint says that the insertion of a word boundary should be avoided (or enforced), *whenever possible*. That is, unless some other, higher-ranked constraint would favor otherwise.

The strength of a constraint is expressed by its expected frequency ($E(xy)$). Since the expected frequency of a biphone is defined as the product of individual segment probabilities, the strength of a constraint is in fact determined by the frequencies of its segment constituents. If two phonemes, in spite of their high frequencies in isolation, seldom occur in conjunction, then there is a strong constraint restricting their co-occurrence (that is, a highly-ranked markedness constraint). Similarly, if two frequent segments occur much more often than expected, then there is a strong constraint favoring the co-occurrence of these phonemes (that is, a highly-ranked contiguity constraint). The ranking value (r) of a markedness or contiguity constraint is thus based on the same statistical measure:

$$r = E(xy) = \Sigma Prob(xY) * \Sigma Prob(Xy) \quad (5)$$

The third category concerns biphones of 'neutral' probability, whose observed frequency is equal to their expected frequency. Such biphones provide the learner with no phonotactic information (which is reflected in the corresponding mutual information value, which is zero in these cases). Therefore, on the basis of the statistical distribution the learner has no reason to induce either type of constraint for such biphones. In a real-life setting, however, the observed frequency will never *exactly* match the expected frequency.

We propose a classification of biphones on the basis of their statistical values as illustrated in Table 1. StaGe induces contiguity constraints for biphones whose observed frequency is *substantially* higher than their expected frequency. In contrast, markedness constraints are induced for biphones with a much lower observed frequency than expected. Finally, no constraints are induced for biphones which carry little probabilistic information.

The decision of when exactly to induce a constraint can be modeled by setting thresholds on the probabilities. We introduce two parameters in the model: a threshold for the induction of markedness constraints (t_M), and a threshold for the induction of contiguity constraints (t_C). Introducing these parameters raises an important issue: how do these thresholds become available to the learner? Are they fixed values, possibly due to biological factors? Or can they be induced from the statistical distribution? Since there is currently no way of resolving this issue, we set the thresh-

Table 1
Classification of biphones according to their statistical values.

Phonotactic category	Observed/expected ratio	Mutual information	Interpretation	Constraint
Low	$O(xy) \ll E(xy)$	$MI(xy) \ll 0$	Pressure towards segmentation	$^*xy$
High	$O(xy) \gg E(xy)$	$MI(xy) \gg 0$	Pressure towards contiguity	Contig-io(xy)
Neutral	$O(xy) \approx E(xy)$	$MI(xy) \approx 0$	No pressure	–

olds manually. As a first estimation, we model the notion 'substantially' as a factor-two deviation from the expected value. That is, a markedness constraint is induced whenever the observed frequency is less than half of the expected frequency ($t_M = 0.5$). A contiguity constraint is induced whenever observed is more than twice the expected frequency ($t_C = 2.0$). A wide range of possible threshold values is tested in Experiment 2.

Single-Feature Abstraction

The categorization of biphones into markedness and contiguity constraints provides the learner with a basis for the construction of phonotactic generalizations. Such generalizations state which classes of sound sequences should in general be segmented, or be kept intact, respectively. The learner constructs a generalization whenever phonologically similar constraints of the same category arise in the constraint set. Generalizations are added to the constraint set, while keeping existing constraints intact.

In modeling the unsupervised learning of phonotactic generalizations, a very basic measure of similarity is implemented, adopting the notion of 'constraint neighbors' (Hayes, 1999). Two constraints are said to be neighbors when they have different values for one single-feature only. We adopt the following set of features: *syllabic, consonantal, approximant, sonorant, continuant, nasal, voice, place, anterior, lateral* for consonants, and *high, low, back, round, long, tense, nasalized* for vowels. For example, the constraints Contig-IO(pl) and Contig-IO(bl) are neighbors, since they have a single-feature difference (*voice* in the first segment). Generalization consists of abstracting over these differences: a new constraint is created where this feature has been neutralized. That is, only shared feature values remain. The resulting constraint (e.g., Contig-IO($x \in \{p, b\}; y \in \{l\}$)) affects a sequence of natural classes, rather than a sequence of individual segments. This more general constraint is added to the constraint set. The algorithm is recursive: the existence of another phonologically similar biphone constraint that is a neighbor of this abstract constraint would trigger a new generalization. In this case, the feature difference between the abstract constraint (Contig-IO($x \in \{p, b\}; y \in \{l\}$)) and the biphone constraint (e.g., Contig-IO(br)) is assessed as the total number of different feature values for features that have not been neutralized in the generalization (e.g., *voice* has been neutralized in position *x*, therefore only the single-feature difference between /l/ and /r/ is taken into account in computing similarity between the two constraints). Abstraction over the single-feature difference creates an even more general constraint (Contig-IO($x \in \{p, b\}; y \in \{l, r\}$)), which is again added to the constraint set. This constraint states that any sequence of /p/ or /b/, followed by /l/ or /r/ (i.e., /pl/, /pr/, /bl/, /br/) should not be broken up by a word boundary. The model thus creates constraints that have a wider scope than the specific constraints that cause the construction of the generalization. In the current example, /pr/ is included in the generalization, while there is no specific constraint affecting this biphone. No more new generalizations are created if there are no more biphone constraints from the same constraint class (markedness or contiguity) within a single-feature difference from this constraint.

Generalizations are ranked according to the expected frequencies of the biphone constraints that support the generalization, averaged over the total number of biphones that are affected by the generalization. For example, the contiguity constraints Contig-IO(pl), Contig-IO(bl), and Contig-IO(br) support the generalization Contig-IO($x \in \{p, b\}; y \in \{l, r\}$) (in addition to the less general Contig-IO($x \in \{p, b\}; y \in \{l\}$) and Contig-IO($x \in \{b\}; y \in \{l, r\}$)). While this abstract constraint is based on three statistically induced constraints, it affects a total of four biphones: /pl/, /bl/, /br/, /pr/. In this hypothetical example, the fourth biphone, /pr/, does not support the generalization. This is because the learner did not assign this biphone to the contiguity category. More formally, it did not pass the statistical threshold for contiguity constraints (t_C), meaning that it was either assigned to the low-probability category (i.e., *pr), or to the neutral probability category (in which case no specific constraint was induced for /pr/). Therefore, the ranking value of the generalization Contig-IO($x \in \{p, b\}; y \in \{l, r\}$) is the summed ranking values (i.e., expected frequencies) of Contig-IO(pl), Contig-IO(bl), and Contig-IO(br), divided by 4. Note that the generalization would have been given a higher ranking value by the learner if there would have been a contiguity constraint Contig-IO(pr). In that case, the ranking value would be calculated as the summed values of Contig-IO(pl), Contig-IO(bl), Contig-IO(br), *and* Contig-IO(pr), divided by 4. Generalizations with stronger statistical support in the constraint category are thus ranked higher than generalizations with weaker support.

The ranking of constraints based on statistical support within constraint categories is crucial, since it ensures that statistically induced constraints will generally be ranked higher than abstracted constraints. This has two important consequences. The first concerns resolving conflicts between constraints. A conflict arises whenever a biphone is affected by both a markedness constraint and a contiguity constraint. In such a case, the markedness constraint favors segmentation of the biphone, whereas the contiguity constraint favors keeping the biphone intact. If two constraints are in conflict, the highest-ranked constraint determines the outcome (assuming OT's strict

domination). StaGe's constraint ranking allows the learner to represent exceptions to phonotactic regularities, since such exceptions (i.e., specific constraints) are likely to be ranked higher than the regularity (i.e., the abstract constraint).

The second, related consequence concerns constraining the generalization mechanism. StaGe's Single-Feature Abstraction is unsupervised. Since all phonological similarities (with recursive single-feature differences) within a constraint category result in new constraints (which are simply added to the constraint set without any further consideration), the model is likely to overgeneralize. However, overly general constraints will have little statistical support in the data (i.e., relatively few biphone-specific constraints will support them). Since the numerical ranking of the constraints is based on exactly this statistical support, overly general constraints will end up at the bottom of the constraint hierarchy. Thus, general constraints like *CC, Contig-IO(CC), *CV, etc., are likely to be added to the constraint set, but their impact on segmentation will be minimal due to their low ranking values. Note that specific constraints are not *by definition* ranked higher than more general ones. Specifically, biphone constraints that are made up of low-frequency segments are likely to be outranked by a generalization, since such biphones have low expected frequencies. The numerical ranking values, which are inferred by the learner in an unsupervised fashion, thus resolve conflicts between markedness and contiguity constraints, while constraining generalization at the same time.

As a consequence of generalization, the 'neutral' probability biphones are now likely to be pulled into the class of either markedness or contiguity constraints. In fact, this may be the main advantage of generalization for the learner: the middle part of the statistical distribution, where the observed frequency approximates the expected frequency, consists of values that are neither high nor low (the 'neutral' category in Table 1). Biphones in such a 'gray' area carry little probabilistic information. Hence, on the basis of statistical learning alone, the learner would have to make a guess whether or not to segment such a biphone, or would have to inspect the values of neighboring biphones in order to estimate the likelihood of a word boundary. Alternatively, when our model encounters a statistically neutral biphone during segmentation, it can use a more general constraint to determine whether the biphone should be segmented or not. The advantage for the learner is thus that biphones for which no reliable statistical information is available can still be reliably segmented (or be kept intact) due to similarity to other biphones.

An example: the segmentation of plosive-liquid sequences

We illustrate how StaGe builds up a constraint set, and how this constraint set is used in speech segmentation using an example of plosive-liquid sequences in Dutch. The example is based on a simulation in which we apply StaGe to consonant clusters (CC biphones) in transcribed utterances of unsegmented speech in the Spoken Dutch Corpus (Goddijn & Binnenpoorte, 2003). We present the model with the problem of predicting word boundaries in the following hypothetical Dutch utterance:

dɑt lɛx ɪk zo trʏx brurtjə

('I'll put that right back, little brother') (6)

To the learning infant this sounds like:

dɑtlɛxɪkzotrʏxbrurtjə (7)

The model processes utterances through a biphone window, and is faced with the task of deciding whether or not a boundary should be inserted for each biphone in the utterance. At the onset of learning, the learner knows nothing, and he/she would have to insert boundaries at random. We focus on the plosive-liquid sequences in the utterance, representing undecided segmentations as '?':

dɑ t?l ɛxɪkzo t?r ʏx b?r urtjə (8)

Through statistical learning the learner builds up a distribution from which constraints are derived. The learner induces a phonotactic constraint whenever a biphone passes the threshold for markedness constraints ($t_M = 0.5$), or the threshold for contiguity constraints ($t_C = 2.0$). For example, the learner discovers that /br/ is overrepresented in the input (i.e., $\frac{O(br)}{E(br)} > 2.0$), and induces Contig-IO(br) with ranking value $r = E(br) = 344.50$. (For simplicity, we assume static ranking values here. However, since ranking values are derived from expected frequencies, ranking values may change due to changes in the statistical distribution.) Fig. 3 shows how this specific constraint affects segmentation of the plosive-liquid sequences in the utterance, using the OT segmentation model. The learner decides that no boundary should be inserted into /br/. With respect to the other sequences, the learner remains ignorant:

dɑ t?l ɛxɪkzo t?r ʏx br urtjə (9)

As a result of multiple specific plosive-liquid overrepresentations in the statistical distribution, the learner induces multiple similar contiguity constraints: Contig-IO(pl), Contig-IO(pr), contig-IO(bl), Contig-IO(dr). Through Single-Feature Abstraction the learner infers a general constraint, Contig-IO($x \in \{p, b, t, d\}; y \in \{l, r\}$) (in addition to less generalized versions of the constraint, which are not shown in this example). On the basis of statistical support (the specific contiguity constraints for /br/, /pl/, /pr/, /bl/, and /dr/), the learner calculates a ranking value

Input: tl, tr, br		Contig-IO(br) ($r = 344.50$)
? tl		
? t.l		
? tr		
? t.r		
☞ br		
b.r		*

Fig. 3. An OT tableau showing segmentation using a single, specific constraint. The upper left cell contains the input. Segmentation candidates for the input are listed in the first column. The upper row shows the induced constraint set, with corresponding ranking values (r) in parentheses. Ranking is irrelevant at this point, since there is only one constraint. The star '*' indicates a violation of a constraint by a segmentation candidate. The index finger '☞' indicates the optimal candidate.

for this abstract constraint. Since the constraint affects a total number of eight biphones, the ranking value is equal to the summed ranking values of the 5 contiguity constraints, divided by 8. In this case, the strength of the constraint is $r = 360.11$. The effect of the generalization is illustrated in Fig. 4. The generalization has substantial strength: it is ranked slightly higher than the specific constraint for /br/. However, since the constraints do not make conflicting predictions, their respective ranking has no effect on segmentation.

Generalization is both helpful and potentially harmful in this case. As a consequence of generalization, the learner is able to make a decision for all plosive-liquid sequences in the utterance. That is, no more undecided segmentations ('?') remain:

dɑ|tl|ɛxɪkzo|tr|ʏx|br|urtjə (10)

The generalization helps the learner, since it correctly predicts that /tr/, which is statistically neutral in continuous speech and has no specific constraint affecting it, should be kept intact. However, the constraint at the same time overgeneralizes with respect to /tl/ and /dl/. While plosive-liquid sequences are in general well-formed, these specific cases typically do not occur within Dutch words.

Since the learner is keeping track of both high and low probabilities in the statistical distribution, the learner also induces markedness constraints, stating which sequences should not occur in the language. For example, the learner induces the constraint *tl, since /tl/ passes the threshold for markedness constraints $\left(\frac{O(tl)}{E(tl)} < 0.5\right)$. The ranking value of *tl is high, due to its high expected frequency (i.e., /t/ and /l/ are frequent phonemes). Therefore, *tl ends up at the top of the constraint hierarchy (see Fig. 5).

Note that the learner has learned an exception to a generalization: the learner will not insert boundaries into plosive-liquid sequences, *unless* it concerns the specific sequence /tl/. In sum, the model has correctly inferred that /br/ should not contain a boundary (due to the induction of a specific constraint, and confirmed by a generalization). In addition, the model has learned that /tr/ should not contain a boundary (due to the abstract constraint). And finally, the model has correctly inferred that /tl/ is an exception to the generalization, and that /tl/ should therefore be broken up by a word boundary. The learner predicts the correct segmentation for all the plosive-sequences it was presented with:

dɑ|t.l|ɛxɪkzo|tr|ʏx|br|urtjə (11)

We conducted a series of computer simulations to test the performance of StaGe in detecting word boundaries in transcriptions of continuous speech. More specifically, we hypothesized that StaGe, which induces phonotactic constraints with various degrees of generality, would outperform purely statistical approaches to speech segmentation, such as transitional probability. While it is an open issue whether human infants use phonotactic generalizations in speech segmentation, better performance by our model would demonstrate a potential role for such generalizations.

It should be stressed that the goal of these simulations is not to obtain perfect segmentations, but to address the effect of the learning of abstract, natural class constraints compared to learning without generalization. A more complete model of infant speech segmentation would involve the integration of multiple cues (e.g., phonotactics, metri-

Input: tl, tr, br		Contig-IO($x \in$ {p,b,t,d}; $y \in$ {l,r}) ($r = 360.11$)	Contig-IO(br) ($r = 344.50$)
☞	tl		
	t.l	*	
☞	tr		
	t.r	*	
☞	br		
	b.r	*	*

Fig. 4. An OT tableau showing segmentation using an abstract constraint. The upper left cell contains the input. Segmentation candidates for the input are listed in the first column. The upper row shows the induced constraint set, with corresponding ranking values (r) in parentheses. The star '*' indicates a violation of a constraint by a segmentation candidate. Constraints are ranked in a strict domination, shown from left to right. Violation of a higher-ranked constraint eliminates a candidate. Ranking is irrelevant here, since the constraints are not in conflict. The index finger '☞' indicates the optimal candidate.

Input: tl, tr, br		*tl ($r = 2690.98$)	Contig-IO($x \in$ {p,b,t,d}; $y \in$ {l,r}) ($r = 360.11$)	Contig-IO(br) ($r = 344.50$)
	tl	*		
☞	t.l		*	
☞	tr			
	t.r		*	
☞	br			
	b.r		*	*

Fig. 5. An OT tableau showing interaction between specific and abstract constraints. The upper left cell contains the input. Segmentation candidates for the input are listed in the first column. The upper row shows the induced constraint set, with corresponding ranking values (r) in parentheses. The star '*' indicates a violation of a constraint by a segmentation candidate. Constraints are ranked in a strict domination, shown from left to right. Violation of a higher-ranked constraint eliminates a candidate. Ranking is relevant here, since the constraints are in conflict with respect to /tl/. The index finger '☞' indicates the optimal candidate.

cal cues, fine-grained acoustic detail). For the current purposes, we focus on the contribution of phonotactics to speech segmentation.

Experiment 1

The goal of the first experiment is to assess whether infants would benefit from phonotactic generalizations in speech segmentation. We address this question in a computer simulation of speech segmentation, allowing us to compare the segmentation performance of models that vary in their assumptions about infant learning capacities. The crucial comparison is between models that are solely based on biphone probabilities, and StaGe, which relies on both statistical learning and generalization.

Method

Materials

The models are tested on their ability to detect word boundaries in broad phonetic transcriptions of the Spoken Dutch Corpus (*Corpus Gesproken Nederlands*, CGN). To create representations of continuous speech, all word boundaries were removed from the transcribed utterances. The 'core' corpus, about 10% of the total corpus, contains a fairly large sample (78,080 utterances, 660,424 words) of high quality transcriptions of spoken Dutch. These broad phonetic transcriptions are the result of automatic transcription procedures, which were subsequently checked and corrected manually (Goddijn & Binnenpoorte, 2003). Due to this procedure, variations in the pronunciation of natural speech are preserved in the transcriptions to a large extent. For example, the word *natuurlijk* ('naturally') occurs in various phonemic realizations. The Spoken Dutch Corpus contains the following realizations for *natuurlijk*:

$$\text{nətyləkˈ}(86), \text{nɑtyləkˈ}(80), \text{nətyrləkˈ}(70), \text{nətykˈ}(68), \text{ntykˈ}(57), \text{natyrləkˈ}(56), \text{nɑtyrləkˈ}(55), \text{nɑtykˈ}(54), \text{tykˈ}(43), \text{natyləkˈ}(40), \text{nətyləgˈ}(29), \text{nətygˈ}(28), \ldots \quad (12)$$

This example illustrates some of the variability that is found in pronunciations of natural speech. In fact, the canonical transcription /natyrlək/ is not the most frequent realization of *natuurlijk* in natural speech. The combination of its size and level of transcription accuracy makes the Spoken Dutch Corpus fairly representative of spoken Dutch. In contrast, previous computational segmentation studies typically used orthographic transcriptions of child-directed speech that were transformed into canonical transcriptions using a phonemic dictionary. Different realizations of words are lost in such a transcription procedure. Our study thus complements previous modeling efforts by investigating the performance of segmentation models in a setting that includes natural variability in the pronunciation of connected speech.

Procedure

The models are tested on novel data (i.e., data that was not used to train the model). To further increase the generalizability of our results, we use 10-fold cross-validation (see e.g., Mitchell, 1997). Each utterance in the corpus is randomly assigned to one out of 10 disjunct sets, such that each set contains approximately 10% of the data points (i.e., biphones) in the corpus. With this random partition, we ran 10 simulations for each model. In each simulation one of the 10 sets (i.e., 10% of the corpus) is used as test set and the remaining nine sets (90% of the corpus) are used as training set. This procedure gives a more reliable estimate of a model's performance than a single randomly chosen test set.

The models are trained on the unsegmented utterances in the training set. The models are then given the task of predicting word boundaries in the test set. Output of the models thus consists of a hypothesized segmentation of the test set. The models are evaluated on their ability to detect word boundaries in the test utterances. The following metrics are used to evaluate segmentation performance:

Hit rate (H):

$$H = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (13)$$

False alarm rate (F):

$$F = \frac{FalsePositives}{FalsePositives + TrueNegatives} \quad (14)$$

d-prime (d'):

$$d' = z(H) - z(F) \quad (15)$$

The hit rate measures the number of word boundaries that the model actually detects. The false alarm rate measures the number of boundaries that are incorrectly placed. The learner should maximize the number of hits, while minimizing the number of false alarms. The d' score (see e.g., MacMillan & Creelman, 2005) reflects how well the model distinguishes hits from false alarms. The following 'dumb' segmentation strategies would therefore each result in a d' score of zero: (a) inserting a boundary into every biphone, (b) not inserting any boundaries and (c) randomly inserting boundaries. These metrics have some advantages over evaluation metrics borrowed from the information retrieval (IR) literature, i.e. recall, precision, and F-score, since such metrics do not necessarily assign low scores to random models. Specifically, random models (in which no learning takes place) will obtain high precision scores whenever there are many potential boundaries to be found (Fawcett, 2006 for a related discussion).

Note that the corpus transcriptions do not specify the exact location of a word boundary in cases of cross-word boundary phonological processes, such as assimilations, deletions, degeminations and glide insertions. For example, *kan nog* ('can still') is often transcribed as /kɑnɔx/. In such cases, it is unclear whether a segment (in this case, /n/) should go with the word to the left or to the right. We treat these phonemes as belonging to the onset of the following word, rather than to the coda of the preceding word.

Segmentation models

Five models are compared: a random baseline; a statistical learning model based on transitional probabilities (TP); a statistical learning model based on observed/expected ratios (O/E); StaGe's Frequency-Driven Constraint Induction (FDCI; i.e., excluding Single-Feature Abstraction); and the complete StaGe model (i.e., including Single-Feature Abstraction).

No learning takes place in the random baseline, and, hence, boundaries are inserted at random. More specifically, for each biphone in the test utterance a random decision is made whether to keep the biphone intact or to break up the biphone through the insertion of a word boundary. The two statistical models (TP, O/E) serve to illustrate the segmentation performance of a learner that does not induce constraints, nor constructs generalizations. These models use segment-based probabilities directly to predict word boundaries in the speech stream.

In addition to evaluating the complete model, we evaluate the performance of StaGe's Frequency-Driven Constraint Induction (FDCI) separately. This is done to provide a clearer picture of the added value of generalization in the model. That is, we are interested in the contribution of both the statistically induced constraints, and the abstract, natural class-based constraints to the segmentation performance of the model.

In the current experiment, phonotactic knowledge is applied to the segmentation of biphones in isolation (i.e., not considering the context of the biphone). For the statistical learning models, we use a threshold-based segmentation strategy (Cairns et al., 1997; Rytting, 2004; Swingley, 2005). A segmentation threshold can be derived from the properties of the statistical distribution. Since a biphone occurs either more or less often than expected, we set the threshold for observed/expected ratios at $O/E = 1.0$. That is, if observed is less than expected, a boundary is inserted. The TP segmentation threshold is less straightforward. For every segment *x*, transitional probability defines a distribution over possible successors *y for this segment*. Transitional probability thus defines a collection of multiple statistical distributions, each requiring their own segmentation thresholds. We define TP thresholds in the same way as we did for O/E ratio: the threshold is the value that would be expected if all segments were equally likely to co-occur. For example, if a segment has three possible successors, then each successor is expected to have a probability of 1/3. If the TP of a biphone is lower than this expected value, a boundary is inserted between the two segments. (Since all segment combinations tend to occur in continuous speech, the TP segmentation thresholds can be approximated by a single threshold: $TP = \frac{1}{|X|}$, where $|X|$ is the size of the segment inventory.)

The statistical models insert boundaries whenever observed is smaller than expected, regardless of the size of this deviation. StaGe makes similar assumptions about thresholds: a markedness constraint is induced when observed is *substantially* smaller than expected; a contiguity constraint is induced when observed is *substantially* larger than expected. We have argued that generalization over such statistically induced phonotactic constraints would be valuable to the learner, since the generalizations are likely to affect the segmentation of statistically neutral biphones in a positive way. It thus makes sense to compare StaGe to the threshold-based statistical models. If phonotactic generalizations improve the segmentation performance of the learner, then StaGe should have a better segmentation performance than a statistical model that simply considers all biphone values, and inserts boundaries based on a single segmentation threshold. As a first test for StaGe we set $t_M = 0.5$ ('observed is less than half of expected') and $t_C = 2.0$ ('observed is more than twice expected'). The induced constraints are interpreted in the OT segmentation model (Fig. 1), which takes *xy*-sequences (i.e., biphones) as input, and returns either *xy* or *x.y*, depending on which candidate is optimal.

Results and discussion

Table 2 shows the hit rate, false alarm rate, and d' scores for each model. The table contains the estimated means (obtained through 10-fold cross-validation), as well as the 95% confidence intervals for those means.

The random baseline does not distinguish hits from false alarms at all, which results in a d' score of approximately zero. While the random baseline inserts the largest amount of correct boundaries, it also makes the largest amount of errors. The d' score thus reflects that this model has not learned anything. The two statistical models (TP, O/E) detect a fair amount of word boundaries (about one third of all boundaries in the test set), while keeping the false alarm rate relatively low. The learning performance is illustrated by the d' values, which show a large increase compared to the random baseline. These scores also show that the formula used to implement statistical learning (either TP or O/E) does not have a great impact on segmentation performance. O/E has a higher hit rate than TP, but also has a higher false alarm rate.

If we apply FDCI to the O/E ratios, the performance of the learner worsens. This is not surprising: FDCI reduces the scope of the phonotactic learner to the edges of the statistical distribution. That is, boundaries are inserted if $O/E < 0.5$, and boundaries are *not* inserted if $O/E > 2.0$. For the remaining biphones (with $0.5 \leqslant O/E \leqslant 2.0$; the 'gray' area), the learner has no other option but to insert boundaries at random. Therefore, the scores for FDCI are closer to the random baseline. A look at the performance of the complete model reveals that StaGe outperforms both the random baseline and the two statistical models in distinguishing hits from false alarms. Through generalization over the statistically induced constraints, the learner has widened the scope of its phonotactic knowledge. The result is that statistically neutral biphones are not segmented at random, nor are they segmented on the basis of their unreliable O/E ratios (which are by definition either higher or lower than 1.0). In contrast, those biphones are affected by phonotactic generalizations, which say that they should either be segmented or not due to their phonological similarity to biphones in either the markedness or contiguity category. This strategy results in the best segmentation performance. Compared to its purely statistical counterpart (O/E), StaGe has both a higher hit rate and a lower false

Table 2
Simulation results for Experiment 1 (biphones in isolation).

		Hit rate			False alarm rate			d'		
Model			95% CI			95% CI			95% CI	
Learning	Segmentation	Mean	Lower	Upper	Mean	Lower	Upper	Mean	Lower	Upper
–	Random	0.4994	0.4979	0.5009	0.5004	0.4992	0.5016	−0.0024	−0.0062	0.0015
TP	Threshold-based	0.3126	0.3102	0.3149	0.1069	0.1060	0.1077	0.7547	0.7489	0.7606
O/*E*	Threshold-based	0.3724	0.3699	0.3750	0.1372	0.1354	0.1390	0.7678	0.7592	0.7765
FDCI	OT	0.4774	0.4763	0.4786	0.2701	0.2691	0.2710	0.5560	0.5532	0.5588
StaGe	OT	0.4454	0.4375	0.4533	0.1324	0.1267	0.1382	0.9785	0.9695	0.9874

Note. The displayed scores are the means obtained through 10-fold cross-validation, along with the 95% confidence interval (CI). TP = transitional probability, *O*/*E* = observed/expected ratio, FDCI = Frequency-Driven Constraint Induction, StaGe = statistical learning and generalization.

alarm rate (although the difference between false alarm rates is marginal). This results in a d' score that is substantially and significantly higher than those of the statistical models (which is reflected in the large difference in means, and non-overlapping confidence intervals).

These results show that our model, which employs both statistical learning and generalization, is better at detecting word boundaries in continuous speech. We interpret these findings as evidence for a potential role for phonotactic generalizations in speech segmentation. That is, if infants were to construct generalizations on the basis of statistically learned biphone constraints, they would benefit from such generalizations in the segmentation of continuous speech.

While the segmentation thresholds for the statistical learners make a mathematically sensible distinction between high and low-probability biphones, and there is evidence that biphone probabilities directly affect speech segmentation by infants (Mattys & Jusczyk, 2001), there is currently no psycholinguistic evidence supporting any exact value for these thresholds. Moreover, the exact values of the constraint induction thresholds employed by StaGe ($t_M = 0.5;\ t_C = 2.0$) are rather arbitrary. It is therefore important to consider a wider range of possible threshold values. In Experiment 2 we look at the effects of varying thresholds for both the statistical learning models and StaGe.

Experiment 2

In Experiment 2 we ask to what extent the results of Experiment 1 can be attributed to the specific threshold configurations that were used. Specifically, the current experiment aims at determining whether StaGe's superior performance was due to a single successful threshold configuration, or whether StaGe in general outperforms statistical learners, regardless of the specific thresholds that are used in the model. To this end we do a Receiver Operating Characteristic (ROC) analysis (e.g., Cairns et al., 1997; Fawcett, 2006; MacMillan & Creelman, 2005). Such an analysis is useful for visualizing the performance of a classifier (such as a threshold-based segmentation model), since it portraits the complete performance in a single curve. An ROC curve is obtained by plotting hit rates as a function of false alarm rates over the complete range of possible threshold values. Such a curve thus neutralizes the effect of using a single, specific threshold in a model and gives a more general picture of the model's performance as the threshold is varied.

Method

Materials and procedure

The materials are identical to Experiment 1. The procedure is identical to Experiment 1, with the exception that we only use the first of our 10 cross-validation sets.

Segmentation models

The crucial comparison will again be between purely statistical learning models (TP, *O*/*E*) and StaGe. Rather than defining a single threshold for the statistical learners, all relevant threshold values are considered. Thresholds are derived from the statistical distribution after the models have processed the training set. (Since there are 1541 different biphones in the training set, each with a different probability, there are 1541 relevant threshold values.) A simulation on the test set is conducted for each threshold value using the same segmentation principle as in the previous experiment: if the probability of a biphone is lower than the current threshold, a boundary is inserted.

For StaGe the situation is slightly more complex, due to the fact that StaGe uses two induction thresholds (t_M, t_C). Testing all combinations of values for the two thresholds is not feasible. A range of thresholds for StaGe is tested, based on the assumption that t_M should be smaller than 1.0, and t_C should be larger than 1.0. A baseline configuration can thus be formulated: $t_M = 1.0;\ t_C = 1.0$. In this case, all biphones with $O/E < 1.0$ result in the induction of a markedness constraint, and all biphones with $O/E < 1.0$ result in the induction of a contiguity constraint. As a consequence, the baseline configuration has no 'neutral probability' category. Such a category is introduced by pushing the thresholds away from 1.0 towards the low- and high-probability edges of the statistical distribution. Varying the induction thresholds causes changes in the amount of specific constraints that are induced by the learner, and affects the generalizations that are based on those constraints. For the induction of markedness constraints we consider {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0} as possible values for t_M. Similarly, we use {1.0, 1.11, 1.25, 1.43,

1.67, 2.0, 2.5, 3.33, 5.0, 10.0} as values for t_C (using logarithmic steps). This results in a total of $10 \times 10 = 100$ different configurations. Note that the configuration from Experiment 1 ($t_M = 0.5;\ t_C = 2.0$) is exactly in the middle. We consider thresholds that are both less and more conservative than this configuration.

Results and discussion

The resulting ROC graph is shown in Fig. 6. Random performance in an ROC graph is illustrated by the diagonal line. For each point on this line, the hit rate is equal to the false alarm rate, and the corresponding d' value is zero. d' increases as the hit rate increases and/or the false alarm rate decreases. Perfect performance would be found in the upper left corner of the ROC space (i.e., where the hit rate is 1, and the false alarm rate is 0). In general, the closer a model's scores are to this point, the better its performance is (due to high hit rate, low false alarm rate, or both).

Since StaGe uses O/E ratios for the induction of constraints, we are particularly interested in the difference between the O/E line and the different configurations of StaGe (represented as circles in the graph). The TP performance is included for completeness. It should be noted, however, that TP slightly outperforms O/E for a substantial part of the ROC graph.

The graph shows that most of the 100 configurations of StaGe lie above the performance line of the statistical learning models (most notably O/E). This should be interpreted as follows: if we make a comparison between StaGe and statistical learning, based on model configurations with identical false alarm rates, StaGe has a higher hit rate (and therefore a higher d'). Conversely, for configurations with identical hit rates, StaGe has a lower false alarm rate (and therefore a higher d'). This confirms the superior performance of StaGe that was found in Experiment 1. StaGe tends to outperform statistical models, regardless of the specific thresholds that are used (with some exceptions, which are discussed below).

While the configuration from Experiment 1 ($t_M = 0.5;\ t_C = 2.0$) retrieved about 44% of the word boundaries at a false alarm rate of 13% (resulting in a d' of 0.98), the results in the current experiment show that this performance can be changed without great loss of d'. For example, the model can be made less conservative, boosting the hit rate to 67%, when using the configuration $t_M = 0.4,\ t_C = 2.5$. In this case, the false alarm rate is 30%. While both the hit rate and false alarm rate are higher, the configuration yields a d' that is comparable to the original configuration: $d' = 0.97$. Similarly, the model can be made more conservative (e.g., $t_M = 0.3,\ t_C = 1.43$, hit rate: 0.32, false alarm rate: 0.07, d': 1.00).

Interestingly, the baseline configuration (the solid circle marked with "1"; $t_M = 1.0;\ t_C = 1.0$) is positioned exactly on the curve of the O/E statistical learning model. In fact, the baseline configuration has a performance that is nearly identical to the statistical O/E model from Experiment 1 (with segmentation threshold $O/E = 1.0$). In this case there

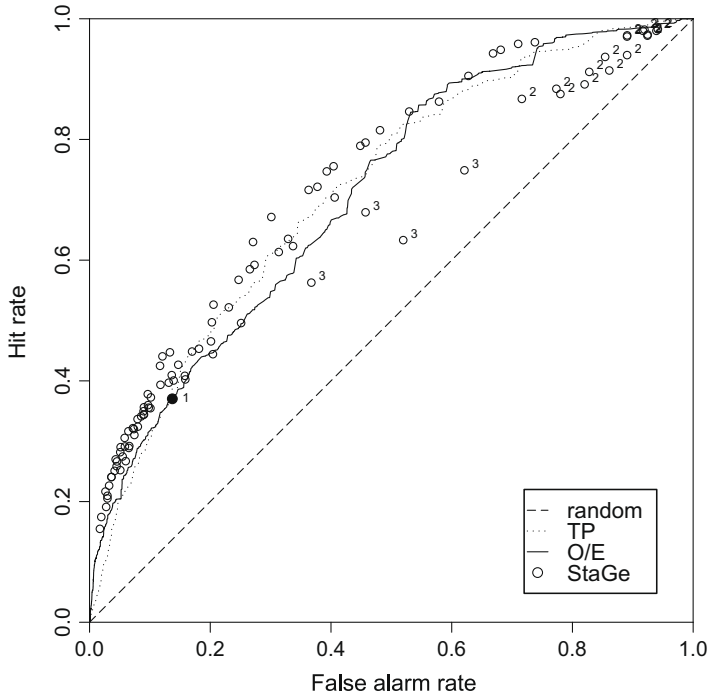

Fig. 6. An ROC (Receiver Operating Characteristic) graph showing the performance of segmentation models over various thresholds. TP = transitional probability, O/E = observed/expected ratio, StaGe = statistical learning and generalization. The StaGe baseline configuration ($t_M, t_C = 1.0$) is the solid circle, marked with"1". Extreme contiguity thresholds ($t_C = 5.0$, or $t_C = 10.0$) are marked with"2", whereas configurations with both extreme markedness and extreme contiguity thresholds ($t_M = 0.1$, or $t_M = 0.2;\ t_C = 5.0$, or $t_C = 10.0$) are indicated with."3".

appears to be neither a positive nor a negative influence of constructing generalizations. This is not surprising: due to the high number of specific constraints in the baseline model, and since specific constraints are typically ranked high in the constraint set, the statistical values tend to overrule the complementary role of phonological similarity. The result is that the model behaves similarly to the purely statistical (O/E) model. Consider, for example, the sequence /tx/ with $\frac{O(tx)}{E(tx)} = 1.0451$. In the baseline configuration, this sequence is kept intact, since the learner induces a highly-ranked constraint CONTIG-IO(tx). The statistical O/E model makes the same prediction: no boundary is inserted, due to an O/E ratio that is slightly higher than 1.0. In contrast, a model with $t_M = 0.5$; $t_C = 2.0$ ignores such biphones because of their neutral probability. As a consequence, the sequence /tx/ is affected by a phonotactic generalization, $^*x \in \{t,d\}; y \in \{k,x\}$, which favors segmentation of the sequence. By pushing the induction thresholds to the edges of the statistical distribution, a smaller role is attributed to probability-based segmentation, and a larger role is attributed to phonological similarity to biphones at the edges of the statistical distribution. This is indeed helpful: inspection of the segmented version of the corpus reveals 4418 occurrences (85.6%) of /t.x/, against only 746 occurrences (14.4%) of /tx/.

Fig. 6 also shows that there are cases in which STAGE performs worse than the statistical learning models. In these cases, the model uses thresholds which are too extreme. Specifically, the model fails for configurations in which the threshold for contiguity constraints is high ($t_C = 5.0$ or $t_C = 10.0$, marked with "2" in the graph). In such cases there are too few contiguity constraints to provide counter pressure against markedness constraints. The model therefore inserts too many boundaries, resulting in a high number of errors. Finally, the worst scores are obtained for configurations that, in addition to an extreme contiguity threshold, employ an extreme markedness threshold ($t_M = 0.1$ or $t_M = 0.2$; $t_C = 5.0$ or $t_C = 10.0$, marked with "3" in the graph). There are not enough constraints for successful segmentation in this case, and the model's performance gets closer to random performance.

The current analysis shows that the superior performance of STAGE, compared to purely statistical models, is stable for a wide range of thresholds. This is an important finding, since there is currently no human data supporting any specific value for these thresholds. The findings here indicate that the learner would benefit from generalization for any combination of thresholds, except when both thresholds are set to 1.0 (in which case generalization has no effect), or when extreme thresholds are used (in which case there are too few constraints).

A possible criticism of these experiments is that we did not use statistical learning as it was originally proposed. We argued for the use of statistical thresholds in applying probabilistic phonotactics directly to the speech segmentation problem. In contrast, the original work on statistical learning by Saffran, Newport et al. (1996) proposes that word boundaries are inserted at *troughs* in transitional probability. In Experiment 3, we ran a series of simulations in which the learning models make use of context in the segmentation of continuous speech.

Experiment 3

In Experiment 3, we test the same learning models as in Experiment 1, but use a different segmentation strategy. Rather than considering biphones in isolation, we allow the learner to use the immediate context of the biphone. That is, for each biphone *xy*, the learner also inspects its neighboring biphones *wx* and *yz*.

Method

Materials and procedure

The materials and procedure are identical to Experiment 1. The same training and test sets are used.

Segmentation models

For the statistical models (TP and O/E), we use the trough-based segmentation strategy, as described in Brent (1999a), which is a formalization of the original proposal by Saffran, Newport et al. (1996). Whenever the statistical value (either TP or O/E) of the biphone under consideration (*xy*) is lower than the statistical values of its adjacent neighbors, i.e. one biphone to the left (*wx*) and one to the right (*yz*), a boundary is inserted into the biphone *xy*. Note that trough-based segmentation is 'threshold-free', since it only considers relative values of biphones.

We again include a simulation using Frequency-Driven Constraint Induction (FDCI) (i.e., without applying generalization) to show how much of the STAGE's performance is due to statistical constraint induction, and how much is due to feature-based abstraction. For FDCI and the complete version of STAGE we use the original threshold configuration with thresholds $t_M = 0.5$ and $t_C = 2.0$. In this experiment we ask whether this configuration is better at detecting word boundaries in continuous speech than the trough-based segmentation models. The input to the OT segmentation model is the same as for the trough-based model, namely *wxyz* sequences. Segmentation candidates for these sequences are *wxyz*, *w*.xyz, *wx*.*yz*, and *wxy*.*z* (see Fig. 1). The model inserts a word boundary into a biphone whenever *wx*.*yz* is optimal.

Note that the segmentation models tested here do not consider the initial and final biphones of an utterance as potential boundary positions, since the current segmentation setting requires neighboring biphones on both sides. No boundaries are therefore inserted in these biphones. The size of this bias is reflected in the random baseline, which uses the same *wxyz*-window, but makes random decisions with respect to the segmentation of *xy*.

Results and discussion

Table 3 shows the estimated means, and 95% confidence intervals, of the hit rates, false alarm rates, and d' scores for each model.

In this experiment the random baseline performs slightly above chance, due to the bias of not inserting boundaries at utterance-initial and utterance-final biphones. In general, using context has a positive impact on segmentation: the performance of all models has in-

Table 3
Simulation results for Experiment 3 (biphones embedded in context).

		Hit rate			False alarm rate			d'		
Model			95% CI			95% CI			95% CI	
Learning	Segmentation	Mean	Lower	Upper	Mean	Lower	Upper	Mean	Lower	Upper
–	Random	0.4900	0.4885	0.4915	0.4580	0.4575	0.4586	0.0803	0.0765	0.0840
TP	Trough-based	0.6109	0.6093	0.6125	0.2242	0.2235	0.2249	1.0399	1.0346	1.0452
O/*E*	Trough-based	0.5943	0.5930	0.5955	0.2143	0.2138	0.2149	1.0301	1.0258	1.0344
FDCI	OT	0.3700	0.3684	0.3716	0.1478	0.1471	0.1484	0.7142	0.7096	0.7188
StaGe	OT	0.4135	0.4062	0.4207	0.0913	0.0882	0.0945	1.1142	1.1081	1.1203

Note. The displayed scores are the means obtained through 10-fold cross-validation, along with the 95% confidence interval (CI). TP = transitional probability, *O*/*E* = observed/expected ratio, FDCI = Frequency-Driven Constraint Induction, StaGe = statistical learning and generalization.

creased compared to the performance found in Experiment 1, where biphones were considered in isolation. As in Experiment 1, FDCI by itself is not able to account for the superior performance. By adding Single-Feature Abstraction to the frequency-driven induction of constraints, the model achieves a performance that is better than that of both statistical models (as measured by d'). While the difference in d' is smaller than in Experiment 1, the difference is significant (due to non-overlapping confidence intervals). As in Experiment 1, the formula used to implement statistical learning (TP vs. *O*/*E*) does not seem to have a substantial impact on the segmentation results. Note that in this experiment the statistical models have a higher hit rate than StaGe. However, this coincides with a much higher false alarm rate. In contrast, StaGe is more conservative: it places fewer, but more reliable word boundaries than the models based on statistical learning. The net result, as measured by d', is that our model is better at distinguishing hits from false alarms.

Note that, although the infant should eventually learn to detect all word boundaries, the relatively small amount of hits detected by the model does not necessarily pose a large problem for the learning infant for two reasons. First, phonotactics is merely one out of several segmentation cues. Some of the boundaries that are not detected by our model might therefore still be detected by other segmentation cues. Second, the high d' score of our model is mainly the result of a low false alarm rate. Our model thus makes relatively few errors. Such an undersegmentation strategy may in fact result in accurate proto-words, i.e. chunks of speech which are larger than words, but to which meaning can easily be attributed (e.g. 'thisis.thedoggy'). In contrast, oversegmentation (e.g. 'this.is.the.-do.ggy') results in inaccurate lexical entries to which no meaning can be attributed. The tendency towards undersegmentation, rather than oversegmentation, is supported by developmental studies (e.g., Peters, 1983). See Appendix A for a selection of marked-up utterances from the Spoken Dutch Corpus which exemplify this undersegmentation behavior.

To get an impression of how robust the results in the current experiment are, we again tested a range of threshold values for StaGe. Because of the computational cost involved in running this type of simulation, we restrict ourselves to a smaller range of thresholds, using only the first of our 10 cross-validation sets. We tested a total of nine different configurations (three thresholds for each constraint category: $t_M = 0.4, 0.5, 0.6$; $t_C = 1.67, 2.0, 2.5$). Out of these nine configurations, one configuration ($t_M = 0.6$, $t_C = 1.67$) performed worse than the statistical learning models (Hit rate: 0.3712; false alarm rate: 0.1013; d': 0.9455). The best performance was obtained using $t_M = 0.4$ and $t_C = 2.5$ (Hit rate: 0.5849; false alarm rate: 0.1506; d': 1.2483). It thus appears that, in the current setting, better performance can be obtained by pushing the thresholds further towards the low- and high-probability edges.

The results of Experiment 3 are similar to the results of Experiments 1 and 2, and therefore provide additional support for our hypothesis that learners benefit from generalizations in the segmentation of continuous speech. Regardless of whether the learner employs a segmentation strategy that considers biphones in isolation, or whether the learner exploits the context of neighboring biphones, in both cases StaGe outperforms models that rely solely on biphone probabilities. We interpret these findings as evidence that the combined strengths of statistical learning and generalization provide the learner with more reliable cues for detecting word boundaries in continuous speech than statistical learning alone (i.e. without generalization).

In Experiments 1–3 StaGe was tested at the end point of learning, i.e. after the model had processed the complete training set. In Experiment 4 we look at how the segmentation performance of the model develops as a function of the amount of input that has been processed by the model.

Experiment 4

The current experiment serves to illustrate developmental properties of the model. That is, given the mechanisms of statistical learning and generalization, how does the model's segmentation behavior change as more input is given to the learner? It should be stressed that the model's trajectory should not be taken literally as a time course of infant phonotactic learning. Several unresolved issues (discussed below) complicate such a comparison. Nevertheless, the experiment allows us to better understand StaGe's learning behavior. In particular, we are interested in whether the model has reached stable segmentation performance after processing the training set. In addition, we look at which constraints are learned when by the model.

Modeling development

The current model works on the assumption that the segment inventory and feature specifications have been established prior to phonotactic learning. It is, however, likely that the processes of segmental acquisition and phonotactic acquisition will, at least partially, overlap during development. Since StaGe makes no prediction regarding the development of the speech sounds themselves (segments, features), and since there exists no corpus documenting such a development, we will assume a static, adult-like segment inventory here.

StaGe models infant phonotactic learning as a combined effort of statistical learning and generalization. Both mechanisms have been shown to be available to 9-month-old infants (biphone probabilities: Mattys & Jusczyk, 2001; similarity-based generalization: Saffran & Thiessen, 2003). StaGe can therefore be thought of as modeling phonotactic learning in infants around this age. Unfortunately, developmental data regarding the *exact* ages (or input quantities) at which each of these mechanisms become active in phonotactic learning are currently lacking. In the current experiment both mechanisms will be assumed to be used from the start. StaGe could in principle, however, start with statistical learning and only start making generalizations after the model has accumulated a certain critical amount of input data.

Another issue with respect to modeling development concerns the model's use of memory (Brent, 1999b). The current implementation of the model assumes a perfect memory: All biphones that are encountered in the input are stored, and are used for constraint induction. Hence, phonotactic constraints are derived from accumulated statistical information about biphone occurrences. While the model's processing of the input is incremental, the perfect memory assumption obviously is a simplification of the learning problem.

Finally, the current set of simulations relies on static, manually-set thresholds for the induction of markedness and contiguity constraints. We use StaGe in the same form as in Experiment 1 (segmentation of biphones in isolation; $t_M = 0.5$, $t_C = 2.0$). The difference with the previous experiments is that we test the performance of the model at various intermediate steps, rather than only testing the model at the end point of learning.

Method

Materials and procedure

The materials are identical to Experiment 1. We use the first of our 10 cross-validation sets. The procedure is different. Rather than presenting the complete training set to the model at once, we present the model with input in a stepwise fashion. The total training set (=100%) contains about 2 million data points (biphones). Starting with an empty training set, the set is filled with training utterances which are added to the set in random order. The model is trained repeatedly after having processed specified percentages of the training set, using logarithmic steps. For each intermediate training set, the model's performance (hit rate, false alarm rate) on the test set is measured. As a consequence, the model's performance progresses from random segmentation (at 0%) to the performance reported in Experiment 1 (100%).

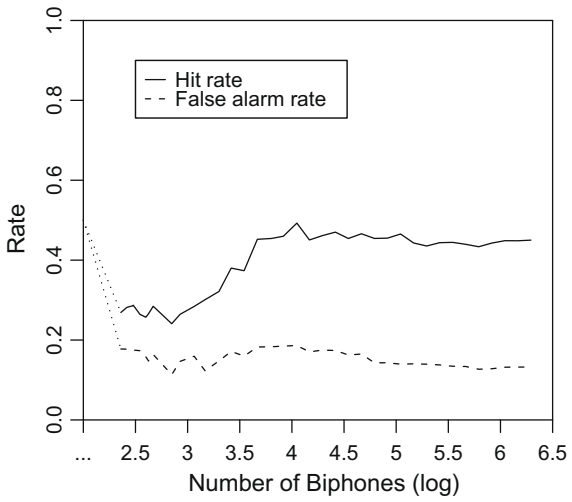

Fig. 7. The development of StaGe, measuring segmentation performance on the test set as a function of input quantity in the training set (i.e. the number of biphones that the learner has processed). Initially, the learner segments at random (hit rate and false alarm rate are 0.5). The learner starts by inducing contiguity constraints (reducing the number of boundaries that are posited) and induces markedness constraints only after a substantial amount of input has been processed. Learning where *not* to put boundaries thus precedes the insertion of word boundaries.

Results and discussion

The developmental trajectory of StaGe is shown in Fig. 7, where the hit rate and false alarm rate are plotted as a function of the input quantity (measured as the number of biphones on a log_{10} scale). The model's segmentation performance appears to become stable after processing ±15,000 biphone tokens ($log_{10} \approx 4.2$), although the false alarm rate still decreases slightly after this point. Starting from random segmentation in the initial state, the model shows effects of undersegmentation after processing only a minimal amount of input: both the hit rate and the false alarm rate drop substantially. Interestingly, the false alarm rate stays low throughout the whole trajectory. The hit rate starts increasing after a substantial amount of input (±650 biphones; $log_{10} \approx 2.8$) has been processed, and becomes relatively stable at ±15,000 biphones.

To explain the segmentation behavior of the model we consider the constraints that are induced at the various developmental stages. The first constraints to emerge in the model are contiguity constraints. This is caused by overrepresentations in the statistical distribution at this point: Since only a small amount of the total number biphones has been processed, all biphones that do occur in this smaller set are likely to have high *O*/*E* ratios. The model tends to induce contiguity generalizations that affect CV and VC biphones (e.g., Contig-IO($x \in \{t, d, s, z\}$; $y \in \{ɑ, ə\}$)), and thereby prevents insertion of boundaries into such biphones. The segmentation of CC and VV sequences is left to

random segmentation (and a relatively small number of specific contiguity constraints). Among the high-ranked specific constraints are several Dutch function words (e.g., CONTIG-IO(ɪn), 'in'; CONTIG-IO(də), 'the'), as well as transitions between such words (e.g., CONTIG-IO(nd)). As a consequence, function words tend to be glued together (e.g., /ɪndə/). This type of undersegmentation continues to exist throughout the learning trajectory, and appears to be a general property of the model.

As the distribution becomes more refined, statistical underrepresentations start appearing, and more biphones fall into the markedness category. The growing number of markedness constraints causes the hit rate to increase. Some of the first specific markedness constraints are *əə, *tn, *nn, *eə, and *mt. Generalizations start appearing after processing about 1500 biphones ($log_{10} \approx 3.2$), such as $^{*}x \in \{ə\}; y \in \{ɪ, ɛ, i, ə\}$, and $^{*}x \in \{n\}; y \in \{l, r\}$. The early-acquired markedness constraints seem to affect other types of biphones (CC, VV) than the early-acquired contiguity constraints (CV, VC). While the model ultimately learns a mixture of markedness and contiguity constraints, affecting all types of biphones, this distinction is a general property of the model.

Taking the modeling simplifications for granted, some of the findings here are supported by developmental studies (e.g., undersegmentation; Peters, 1983). Conversely, new findings that follow from the computational model could provide a basis for future experimental testing in developmental research.

General discussion

In this paper we have proposed a computational model for the induction of phonotactic knowledge from continuous speech. The model, STAGE, implements two learning mechanisms that have been shown to be accessible to infant language learners. The first mechanism, statistical learning, allows the learner to accumulate data and draw inferences about the probability of occurrence of such data. The second mechanism, generalization, allows the learner to abstract away from the observed input and construct knowledge that generalizes to unobserved data, and to probabilistically neutral data. We integrate these mechanisms into a single computational model, thereby providing an explicit, and testable, proposal of how these two mechanisms might interact in infants' learning of phonotactics.

In addition, we investigated the potential role of phonotactic generalizations in speech segmentation. We hypothesized that learners would benefit from constructing phonotactic generalizations in the segmentation of continuous speech. This hypothesis was confirmed in a series of computer simulations, which demonstrated that STAGE, which acknowledges a role both for statistical learning and for generalization, and which uses a modified version of OT to regulate interactions between constraints, was better at detecting word boundaries in continuous speech data than models that rely solely on biphone probabilities. Specifically, the generalizations seem to positively affect the segmentation of biphones whose phonotactic probability cannot reliably be classified as being either high or low.

STAGE relies on the edges of the statistical distribution, rather than on the whole distribution. The model employs statistical thresholds to filter out biphones that cannot be reliably classified as being of either high or low probability. Successful generalization crucially depends on this categorization: Experiment 2 showed that generalization has no effect if all biphones are taken into account in the generalization process. In addition, the experiment shows that generalization fails when the thresholds are set to extreme values. STAGE thus provides an explicit description of how generalization relies on statistical learning: statistical learning provides a basis for generalization. This basis is constructed through the use of thresholds on the values that are obtained through statistical learning.

Several properties of STAGE's Frequency-Driven Constraint Induction are worth noticing. First, it should be stressed that the learner does not process or count any word boundaries for the induction of phonotactic constraints. It has been argued that phoneme pairs typically occur either only within words or only across word boundaries, and that keeping track of such statistics provides the learner with a highly accurate segmentation cue (Hockema, 2006). However, such a counting strategy requires that the learner is equipped with the ability to detect word boundaries *a priori*. This is a form of supervised learning that is not representative of the learning problem of the infant, who is confronted with unsegmented speech input. In contrast, STAGE draws inferences about the likelihood of boundaries based on probabilistic information about the occurrences of segment pairs in unsegmented speech. The model categorizes biphones without explicitly taking any boundary information into account, and thus represents a case of unsupervised learning.

Second, while earlier segmentation models employed either boundary detection strategies (Cairns et al., 1997; Rytting, 2004) or clustering strategies (Swingley, 2005), STAGE integrates these approaches by acknowledging a role for both edges of the statistical distribution. A low-probability biphone is likely to contain a word boundary, which is reflected in the induction of a markedness constraint. Conversely, a high-probability biphone is interpreted as a contiguous cluster, reflecting the low probability of a boundary breaking up such a biphone.

Finally, the assumption of functionally distinct phonotactic categories is also what sets our approach apart from earlier constraint induction models (Hayes, 1999; Hayes & Wilson, 2008). Whereas these models induce only markedness constraints, penalizing sequences which are ill-formed in the language, STAGE uses both ends of a statistical distribution, acknowledging that other sequences are *well*-formed. While ill-formedness exerts pressure towards segmentation, well-formedness provides counter pressure *against* segmentation. This distinction provides a basis for the construction of phonotactic generalizations, while at the same time providing counter pressure against overgeneralization.

By adding generalization to the statistical learning of phonotactic constraints, the model achieves two things. First, through generalization over observed data, the learner has acquired abstract knowledge. While most theories of abstract linguistic knowledge assume abstract represen-

tations to be innate, our model shows that it is possible to derive abstract phonotactic constraints from observed data. This finding is in line with recent studies (e.g., Hayes & Wilson, 2008) that aim at minimizing the role of Universal Grammar in language acquisition, while still acknowledging the existence of abstract representations and constraints. Second, the interaction of markedness and contiguity constraints, with varying degrees of generality, and the use of strict constraint domination in speech segmentation, allows the learner to capture generalizations, as well as exceptions to these generalizations. Similar to earlier work by Albright and Hayes (2003), our model thus makes no principled distinction between 'exceptions' and 'regularities'. Both are modeled in a single formal framework.

Although it is not known whether infants actually do learn phonotactic generalizations from continuous speech, and use such generalizations in speech segmentation, our study provides indirect support for such a strategy. Our simulations show that infants would benefit from such an approach in the segmentation of continuous speech. Of course, it remains to be determined by experimental testing whether infants actually exploit the benefits of both statistical learning and generalization in a way that is predicted by our model. Nevertheless, the psychological plausibility of StaGe is based on evidence for the learning mechanisms that it combines. Infants' sensitivity to the co-occurrence probabilities of segment pairs has been demonstrated (Jusczyk et al., 1994; Mattys & Jusczyk, 2001; White et al., 2008). In addition, infants' capacity to abstract over linguistic input in order to construct phonotactic generalizations has been demonstrated (Chambers et al., 2003; Saffran & Thiessen, 2003). A recent series of artificial grammar learning experiments by Finley and Badecker (2009) provides further evidence for the role of feature-based generalizations in phonological learning. There is thus evidence for both statistical learning and feature-based generalization in phonotactic learning.

In addition, StaGe is compatible with available evidence about infants' representational units. Our generalization algorithm implements phonological features to express the similarity between segments. Although evidence for the psychological reality of such abstract phonological features is limited, infants have been shown to be sensitive to dimensions of acoustic similarity (Jusczyk, Goodman et al., 1999; Saffran and Thiessen, 2003; White et al., 2008). Furthermore, several studies suggest that abstract phonological features may constrain infant phonotactic learning (Cristià & Seidl, 2008; Seidl & Buckley, 2005).

Given these findings, we believe that StaGe makes reasonable assumptions about the mechanisms and representations that are involved in phonotactic learning by infants. However, the model is not committed to these representations *per se*. The statistical learning component of the model, Frequency-Driven Constraint Induction, could be applied to other units, such as syllables or allophones. The generalization mechanism, Single-Feature Abstraction, could, in principle, work with different types of features. It remains to be seen how differences in assumptions about the representational units that are processed by the model would affect the speech segmentation performance of the model.

An important property of StaGe is that the model is unsupervised. That is, unlike previous models of phonotactic learning (e.g., Hayes & Wilson, 2008; Pierrehumbert, 2003), the model does not receive any feedback from segmented utterances or word forms in the lexicon. The learner induces phonotactic constraints from its immediate language environment, which consists of unsegmented speech. We thereby provide a computational account of phonotactic learning during the very first stages of lexical acquisition. In fact, through the induction of phonotactics from unsegmented speech, the learner is able to bootstrap into word learning. Alternatively, one might argue that infants rely on segmentation cues other than phonotactics to learn their first words, and then derive phonotactics from a proto-lexicon. Such a view raises the new question of how such other segmentation cues would be learned. In general, knowledge of words cannot be a prerequisite for the learning of segmentation cues, since words are the *result* of segmentation (e.g., Brent & Cartwright, 1996; Swingley, 2005). If these cues are to be used to bootstrap into word learning, then such cues need to be learned from unsegmented speech input. We therefore argue that at least some knowledge of phonotactics comes before the infant starts to build up a vocabulary of words. This knowledge is learned from continuous speech using a combination of statistical learning and generalization. An interesting open issue is what the effect of the lexicon will be on the child's acquired phonotactic knowledge when her vocabulary reaches a substantial size.

Another possible area of future exploration would be to investigate whether the combination of statistical learning and generalization, as proposed in our model, could also be applied to the induction of other linguistic segmentation cues. Swingley (2005) suggests that statistical clustering of syllable *n*-grams could serve as a basis to bootstrap into the Metrical Segmentation Strategy (Cutler & Norris, 1988). While it is unclear what the exact generalization mechanism would look like in this case, the general view that statistical learning serves as a basis for generalization is in accordance with the predictions of StaGe. A similar view has been proposed for a different type of phonological learning by infants: White et al. (2008) propose that learning phonological alternations requires two different forms of computation. Infants first learn about dependencies between specific sounds in the input, and then group similar sounds that occur in complementary distribution into a single phonemic category.

A final issue concerns the model's use of memory. The model derives phonotactic constraints from accumulated statistical information about biphone occurrences. The model thus assumes a perfect memory, which is a simplification of the learning problem. A more psychologically motivated memory implementation could be obtained by adding memory decay to the model: biphones that were encountered a long time ago should be "forgotten". A similar approach has been proposed by Perruchet and Vinter (1998), who argue for the implementation of laws of associative learning and memory, such as temporal proximity, in segmentation models.

To conclude, the mechanisms used by StaGe have received much attention in the psycholinguistic literature,

and thus appear to be available to infant language learners. StaGe provides a computational account of how statistical learning and generalization might interact in the induction of phonotactics from continuous speech. The combined strengths of statistical learning and generalization provide the learner with a more reliable cue for detecting word boundaries in continuous speech than statistical learning alone. Our computational study thus demonstrates a potential role for phonotactic generalizations in speech segmentation.

Acknowledgments

This work was supported by the Netherlands Organization for Scientific Research (NWO grant 277-70-001 to René Kager). We are grateful to Natalie Boll-Avetisyan, Walter Daelemans, Bruce Hayes, Marijn Koolen, Tom Lentz, Jacques Mehler, Hugo Quené, and three anonymous reviewers for providing valuable comments on earlier drafts of this paper. Parts of this work were presented at the University of Michigan (ExpOT, 2007), University College London (MLCS07), University of Toulouse-le Mirail (OCP5), University of Wellington (LabPhon11), Ludwig-Maximilians-University, Munich (CCSC, 2008), and Boston University (BUCLD33), as well as at informal meetings at the University of Amsterdam and the International School of Advanced Studies (SISSA), Trieste. We would like to thank these audiences for helpful discussion.

Appendix A. Examples of segmentation output (Experiment 3)

Orthography:	*Ik zou natuurlijk idioot zijn als ik ja zou zeggen he?* ('I would of course be an idiot if I would say yes')
Transcription:	ɪk zɑu natylək idijot sɛin ɑls ɪk ja zɑu zɛɣə hɛ
StaGe:	ɪk zɑunatylək idijot sɛin ɑlsɪk jazɑuzɛɣə hɛ
O/*E*:	ɪk zɑu natyl ək idijo tsɛin ɑl sɪk ja zɑu zɛ ɣə hɛ
TP:	ɪk zɑunat ylək idij ots ɛin ɑls ɪk ja zɑuzɛ ɣə hɛ
Orthography:	*Toch een heel bekend standaardwerk.* ('Just a well-known standard work')
Transcription:	tɔx ən hel bəkɛnt stɑndard wɛrək
StaGe:	tɔxən helbəkɛnt stɑndard wɛrək
O/*E*:	tɔxə nhel bək ɛn tst ɑnd ard wɛr ək
TP:	tɔxən hel bək ɛnt st ɑnd ard wɛ rək
Orthography:	*Liefde is de enige manier om je tegen de dood te verzetten.* ('Love is the only way to resist death')
Transcription:	livdə ɪs ət enəɣə manir ɔm jə teɣə də dot tə vərzɛtə
StaGe:	liv də ɪsətenəɣəmanirɔm jəteɣədədot təvər zɛtə
O/*E*:	liv də ɪsət en əɣə ma ni rɔm jət eɣə dədot təvər zɛ tə
TP:	liv də ɪsət enə ɣə manir ɔm jət eɣə dəd ot tə vər zɛ tə
Orthography:	*En ik heb ook in ieder geval uh ja een paar collega's waarmee ik heel goed daarmee zou kunnen samenwerken.* ('And I also have in any case uh yes a couple of colleagues with whom I might very well collaborate')
Transcription:	ɛn ɪk hɛp ok ən idə xəfal ə ja əm par kɔlexas wame ɪk hel xut dame zɑu kʉnə saməwɛrəkə
StaGe:	ɛnɪk hɛpokənidəxəfaləja əm parkɔlexas wame ɪk hel xut damezɑukʉnəs aməwɛrəkə
O/*E*:	ɛn ɪk hɛp ok əni də xə fɑlə ja əmp ar kɔ le xa swa me ɪk hel xut dame zɑu kʉn əs am əwɛr əkə
TP:	ɛn ɪk hɛp okən idə xə fɑ lə ja əm par kɔ le xas wa me ɪk hel xut damez ɑuk ʉnəs amə wɛ rə kə
Orthography:	*Vond ik heel erg boeiend.* ('I found (it) very interesting indeed')
Transcription:	fɔnd ɪk hel ɛrəx bujənt
StaGe:	fɔndɪk helɛrəx bujənt
O/*E*:	fɔ nd ɪk hel ɛr əx bujənt
TP:	fɔnd ɪk hel ɛrəx bujənt
Orthography:	*Ik weet nog niet precies hoe ik zal gaan.* ('I don't know yet precisely how I will go')
Transcription:	k wet nɔɣ nit prəsis hu wɪk sɑl xan
StaGe:	kwet nɔɣ nit prəsis hu wɪk sɑlxan
O/*E*:	kwet nɔɣ nit pr əsis hu wɪks ɑl xan
TP:	kwet nɔɣ nit prəs is hu wɪks ɑl xan
Orthography:	*Maar in ieder geval in die film heeft ie wat langer haar.* ('But in any case in this film his hair is somewhat longer')
Transcription:	ma ɪn i fɑl ɪn di fɪlm heft i wɑt lɑŋə har
StaGe:	ma ɪnifɑlɪndifɪlm hef ti wɑt lɑŋə har
O/*E*:	ma ɪni fɑl ɪn difɪl mhef tiwɑt lɑŋ əh ar
TP:	ma ɪni fɑl ɪndi fɪlm he ft iwɑt lɑ ŋə har
Orthography:	*Een paar jaar geleden heeft ze haar restaurant verkocht.* ('A few years ago she sold her restaurant')
Transcription:	əm pa ja xəledən heft sə ha rɛsturɑ̃: fərkɔxt
StaGe:	əm pajaxəledən heft sə harɛsturɑ̃:fər kɔxt

(*continued on next page*)

Appendix A (*continued*)

O/*E*:	əmp aja xə le də nhef ts əh arɛ stu rɑ̃:f ər kɔxt
TP:	əm paja xə ledən he ftsə har Est ur ɑ̃:fər kɔxt
Orthography:	*Binnen in de vuurtoren zit een groot dier in een schommelstoel.* ('Inside the lighthouse a large animal is sitting in a rocking chair')
Transcription:	bɪnən ɪn də vyrtorən zɪt əŋ xrod dir ɪn ən sxɔmɔstul
StaGe:	bɪnən ɪndəvyrtorən zɪ təŋ xrod dirɪn ən sxɔmɔ stul
O/*E*:	bɪn ən ɪndəv yrt or ənzɪ tə ŋx rod di rɪn ən sx ɔmɔ st ul
TP:	bɪn ən ɪn dəv yrt orən zɪ tə ŋxrod dir ɪn ən sx ɔm ɔst ul
Orthography:	*Die horizon kan toch ook naar ons komen.* ('That horizon may also come to us')
Transcription:	di horizɔn kɑn tɔx ok nar ɔns komə
StaGe:	dihorizɔn kɑn tɔxok narɔns komə
O/*E*:	di hor iz ɔn kɑn tɔx ok nar ɔns komə
TP:	di hor izɔn kɑnt ɔx ok nar ɔns komə

References

Albright, A., & Hayes, B. (2002). Modeling English past tense intuitions with Minimal Generalization. In M. Maxwell (Ed.), *Proceedings of the sixth meeting of the ACL special interest group in computational phonology* (pp. 58–69). Philadelphia: Association for Computational Linguistics.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition, 90*, 119–161.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324.

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition, 83*, 167–206.

Boersma, P., Escudero, P., & Hayes, R. (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. In *Proceedings of the 15th international congress of phonetic sciences* (pp. 1013–1016).

Boersma, P., & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry, 32*(1), 45–86.

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

Brent, M. R. (1999b). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences, 3*, 294–301.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*, 93–125.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology, 33*, 111–153.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition, 87*, B69–B77.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*, 221–268.

Cristià, A., & Seidl, A. (2008). Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development, 4*, 203–227.

Cutler, A., & Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 113–121.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874.

Finley, S., & Badecker, W. (2009). Artificial language learning and feature-based generalization. *Journal of Memory and Language, 61*, 423–437.

Fowler, C., Best, C., & McRoberts, G. (1990). Young infants' perception of liquid coarticulatory influences on following stop consonants. *Perception & Psychophysics, 48*, 559–570.

Friederici, A. D., & Wessels, J. M. I. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & Psychophysics, 54*, 287–295.

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory, 22*, 179–228.

Goddijn, 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).

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109–135.

Hayes, B. (1999). Phonetically driven phonology: The role of Optimality Theory and inductive grounding. In M. Darnell, E. Moravscik, M. Noonan, F. J. Newmeyer, & K. M. Wheatley (Eds.), *Functionalism and formalism in linguistics* (pp. 243–285). Amsterdam: John Benjamins.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry, 39*, 379–440.

Hockema, S. A. (2006). Finding words in speech: An investigation of American English. *Language Learning and Development, 2*, 119–146.

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*, 548–567.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development, 64*, 675–687.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language, 32*, 402–420.

Jusczyk, P. W., Goodman, M. B., & Baumann, A. (1999). Nine-month-olds' attention to sound similarities in syllables. *Journal of Memory and Language, 40*, 62–82.

Jusczyk, P. W., Hohne, E., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics, 61*, 1465–1476.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*, 159–207.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*, 630–645.

Legendre, G., Miyata, Y., & Smolensky, P. (1990). Harmonic Grammar: A formal multi-level connectionist theory of linguistic wellformedness: Theoretical foundations. In *Proceedings of the twelfth annual conference of the cognitive science society* (pp. 388–395). Cambridge, MA: Lawrence Erlbaum.

Lin, Y., & Mielke, J. (2008). Discovering place and manner features: What can be learned from acoustic and articulatory data? In *Proceedings of the 31st Penn Linguistics Colloquium* (pp. 241–254). Philadelphia, PA: University of Pennsylvania.

MacMillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Marcus, G., Vijayan, S., Rao, S. B., & Vishton, P. (1999). Rule learning by seven-month old infants. *Science, 283*, 77–80.

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78*, 91–121.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38*, 465–494.

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

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*, B101–B111.

McCarthy, J. J., & Prince, A. S. (1995). Faithfulness and reduplicative identity. *Papers in Optimality Theory: University of Massachusetts*

occasional papers (Vol. 18, pp. 249–384). Amherst, MA: Graduate Linguistics Student Association.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language, 39*, 21–46.

Mitchell, T. M. (1997). *Machine learning*. New York, NY: McGraw-Hill.

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development, 66*, 911–936.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*, 127–162.

Perruchet, P., & Peereman, R. (2004). The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics, 17*, 97–119.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*, 246–263.

Peters, A. M. (1983). *The units of language acquisition*. Cambridge, UK: Cambridge University Press.

Pierrehumbert, J. B. (2003). Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics*. Cambridge, MA: The MIT Press.

Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint interaction in generative grammar (technical report)*. New Brunswick, NJ: Rutgers University Center for Cognitive Science, Rutgers University.

Rytting, C. A. (2004). Segment predictability as a cue in word segmentation: Application to modern Greek. In *Current themes in computational phonology and morphology: Seventh meeting of the ACL special interest group on computational phonology (SIGPHON)* (pp. 78–85). Barcelona: Association for Computational Linguistics.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.

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.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology, 39*, 484–494.

Seidl, A., & Buckley, E. (2005). On the learning of arbitrary phonological rules. *Language Learning and Development, 1*, 289–316.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America, 111*, 1872–1891.

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

Tesar, B., & Smolensky, P. (2000). *Learnability in Optimality Theory*. Cambridge, MA: The MIT Press.

Toro, J. M., Nespor, M., Mehler, J., & Bonatti, L. L. (2008). Finding words and rules in a speech stream: Functional differences between vowels and consonants. *Psychological Science, 19*, 137–144.

Weber, A., & Cutler, A. (2006). First-language phonotactics in second-language listening. *Journal of the Acoustical Society of America, 119*, 597–607.

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

White, K. S., Peperkamp, S., Kirk, C., & Morgan, J. L. (2008). Rapid acquisition of phonological alternations by infants. *Cognition, 107*, 238–265.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences, 8*, 451–456.