COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 37 (2013) 103–124

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2012.01267.x

Learning Phonemes With a Proto-Lexicon

Andrew Martin,[a,b] Sharon Peperkamp,[a] Emmanuel Dupoux[a]

[a]*Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS-ENS-CNRS)*
[b]*Laboratory for Language Development, RIKEN Brain Science Institute*

Received 26 May 2011; received in revised form 28 November 2011; accepted 5 March 2012

Abstract

Before the end of the first year of life, infants begin to lose the ability to perceive distinctions between sounds that are not phonemic in their native language. It is typically assumed that this developmental change reflects the construction of language-specific phoneme categories, but how these categories are learned largely remains a mystery. Peperkamp, Le Calvez, Nadal, and Dupoux (2006) present an algorithm that can discover phonemes using the distributions of allophones as well as the phonetic properties of the allophones and their contexts. We show that a third type of information source, the occurrence of pairs of minimally differing word forms in speech heard by the infant, is also useful for learning phonemic categories and is in fact more reliable than purely distributional information in data containing a large number of allophones. In our model, learners build an approximation of the lexicon consisting of the high-frequency *n*-grams present in their speech input, allowing them to take advantage of top-down lexical information without needing to learn words. This may explain how infants have already begun to exhibit sensitivity to phonemic categories before they have a large receptive lexicon.

Keywords: First language acquisition; Statistical learning; Phonemes; Allophonic rules

1. Introduction

Infants acquire the fundamentals of their language very quickly and without supervision. In particular, during the first year of life, they converge on the set of phonemic categories for their language (Kuhl et al., 2008; Werker & Tees, 1984), they become attuned to its phonotactics (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993), and they begin to extract words from continuous speech (Jusczyk & Aslin, 1995). Despite a wealth of studies

Correspondence should be sent to Andrew Martin, Laboratory for Language Development, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako-shi, Saitama, 351-0198, Japan. E-mail: amartin@brain.riken.jp

documenting these changes (see reviews in Jusczyk, 2000; Kuhl, 2004; Werker & Tees, 1999), very little is known about the *computational mechanisms* involved in this rapid acquisition process.

One reason for this state of affairs is that most computational modeling studies have narrowly focused on one learning sub-problem, assuming the others can be solved independently. For instance, proposed mechanisms for word segmentation have typically assumed that phonemic categories have been acquired beforehand (Brent, 1999; Goldwater, Griffiths, & Johnson, 2009; Venkataraman, 2001). Similarly, models of the phonetic/phonological buildup of phonemic categories assume that these categories have to be constructed or refined through generalization over lexical items (Pierrehumbert, 2003). Of course, this presupposes that a lexicon has been acquired beforehand. Finally, models of grammar induction for learning stress systems or phonotactics make the assumption that both the phonemic categories and the lexicon of underlying forms have already been learned (Dresher & Kaye, 1990; Tesar & Smolensky, 1998). These circularities illustrate what is known as the ''bootstrapping problem,'' which is the apparently insoluble problem that faces infants when they have to acquire several co-dependent levels of linguistic structure. Therefore, while the existing computational studies have allowed us to better understand individual pieces of the puzzle, taken together, they do not coalesce into a coherent theory of early language acquisition. This is especially true when one considers that the experimental data show that infants do not learn phonemic categories, lexical entries, and phonotactic regularities one after the other, but rather, start learning these levels almost simultaneously, between 5 and 9 months of age (see Kuhl, 2004 for a review).

A second reason for our lack of comprehension of this learning process is that most proposed algorithms have only been tested on artificial or simplified language inputs, and only assume that they can *scale up* to more realistic inputs like raw speech; in several instances, however, such assumptions have turned out to be incorrect. For example, it has been claimed that phonetic categories emerge through an unsupervised statistical learning process whereby infants track the modes of the distributions of phonetic exemplars (Maye, Weiss, & Aslin, 2008; Maye, Werker, & Gerken, 2002) or perform perceptual warping of irrelevant phonetic dimensions (Jusczyk, 1993; Kuhl et al., 2008). Modeling studies have found that Self Organizing Maps (Gauthier, Shi, & Xu, 2007a,b; Guenther & Gjaja, 1996), Gaussian mixtures (de Boer & Kuhl, 2003; Mcmurray, Aslin, & Toscano, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007), or Hidden Markov Models (Goldsmith & Xanthos, 2009) can converge on abstract categories for tones, vowels, or consonants in an unsupervised fashion, that is, without any kind of lexical or semantic feedback. However, these clustering algorithms have only been applied to a fragment of the phonology or to individual speech dimensions segmented or extracted by hand (such as F0 contours, F1 and F2, vowel duration, or VOT). Do they scale up to the full complexity of unsegmented speech signals? Varadarajan, Khudanpur, and Dupoux (2008) applied unsupervised clustering (a version of Gaussian modeling using successive state splitting) on 40 h of raw conversational speech and found that the algorithm constructed a network of states that failed to correspond in a one-to-one fashion to phonemic categories. Although the states were sufficient to sustain state-of-the-art speech recognition performance, they did so by encoding a very large

number of heavily context-dependant discrete acoustic (subphonemic) events. Evidently, blind unsupervised clustering on raw speech needs to be supplemented by further processes in order to adequately model the early construction of phonemic categories by infants.

It is important to realize that the failure to derive abstract phonemic categories from the raw signal using unsupervised clustering in Varadarajan et al. (2008) is not a fluke of the particular algorithm they used but reflects a deep property of running speech signals: The acoustic realization of phonemes is massively context-dependent, creating a great deal of overlap between phoneme categories (see Pierrehumbert, 2003). One could argue that gradient coarticulation effects and abstract linguistic rules are fundamentally different, in that the former are the result of universal principles which could be undone by infants without needing any language-specific knowledge. If this were the case, perhaps the wealth of variants discovered by ASR systems could be reduced to a small, orderly set of allophones. There is substantial evidence, however, that even fine-grained coarticulation varies across languages (Beddor, Harnsberger, & Lindemann, 2002; Choi & Keating, 1991; Manuel, 1999; Öhman, 1966), and that perceptual compensation on the part of listeners shows language-specific effects (Fowler, 1981; Fowler & Smith, 1986; Lehiste & Shockey, 1972). The boundary between gradient phonetic effects and discrete phonological rules is thus difficult to draw, especially from the viewpoint of infants who have not yet acquired a system of discrete categories.

Constructing a single context-independent Gaussian model for each abstract phoneme is therefore bound to yield more identification errors when compared to finer-grained models that make use of contextual information. This is well known in the speech recognition literature, where HMM models of abstract phonemes yield worse performance than models of contextual diphone- or triphone-based allophones (Lee, 1988; Makhoul & Schwartz, 1995). Typically, a multi-talker recognition system requires between 2,000 and 12,000 contextual allophones (representing between 50 and 300 allophones per phoneme) in order to achieve good performance (Jelinek, 1998).

Of course, one could propose that infants adopt the same strategy—simply compile a large number of such fine-grained allophonic categories and use them for word learning. But this would be unwise, since massive allophony causes the performance of word segmentation algorithms to drop dramatically (Rytting, Brew, & Fosler-Lussier, 2010). Boruta, Peperkamp, Crabbé, and Dupoux (2011) ran two such algorithms—the MBDP-1 of Brent (1999) and the NGS-u of Venkataraman (2001)—on transcripts of child-directed speech in which contextual allophony has been implemented. They found that when the input contains an average of 20 allophones per phoneme, the performance of both algorithms falls below that of a control algorithm which simply inserts word boundaries at random. In addition, there is empirical evidence that infants do not do this, since before they have a large comprehension lexicon toward the end of their first year of life,[1] they have begun to lose the ability to discriminate within-category distinctions (e.g., Dehaene-Lambertz & Baillet, 1998; Werker & Tees, 1984) and pay less attention to allophonic contrasts compared to phonemic ones (Seidl, Cristia, Bernard, & Onishi, 2009).

This does not mean that fine-grained phonetic information is necessarily disregarded, as speakers could make use of both phonetic detail and abstract categories (Cutler, Eisner,

McQueen, & Norris, 2010; Ramus et al., 2010). Evidence that phonetic detail is represented includes results showing that both infants and adults use allophonic differences for the purposes of word segmentation (Gow & Gordon, 1995; Jusczyk, Hohne, & Bauman, 1999; Nakatani & Dukes, 1977), and that adults are able to remap allophones to abstract phonemes with little training (Kraljik & Samuel, 2005; Norris, McQueen, & Cutler, 2003). Similarly, both adults and infants make use of within-category phonetic variation during lexical access (McMurray & Aslin, 2005; McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008).

In this paper, we revisit the issue of early acquisition of phoneme categories, while attempting to address simultaneously the two problems mentioned above, that is, the circularity problem and the scalability problem. We build our approach on the work described in two recent papers: Peperkamp et al. (2006) and Feldman, Griffiths, and Morgan (2009).

Peperkamp et al. (2006) proposed that infants construct phoneme categories in two steps: Starting with raw signal, they first construct a rather large number of detailed phonetic categories (allophones), and in a second step they cluster these allophones into abstract phonemes. They demonstrated that two potential sources of information, distributional and phonetic, are potentially helpful in performing this clustering step. Distributional information is useful for category formation because variants of a single phoneme occur in distinct contexts, while phonetic information is relevant due to the fact that such variants tend to be acoustically or articulatorily similar to each other and to share phonetic features with their contexts. However, the phonetic part of the study may have created a circularity problem since it assumed the prior acquisition of phonetic features (which may itself require learning the phonological system; see Clements, 2009). In addition, the entire study may have a scalability problem, as it only tested a very small number of allophones in each language (7 allophones of French in Peperkamp et al., 2006; 15 allophones of Japanese in Le Calvez, Peperkamp, & Dupoux, 2007). This falls short of the range of allophonic/coarticulatory variants that are needed to achieve reasonable performance in speech recognition, which are on the order of a few thousand variants. Since we do not know the granularity of the categories constructed by actual infants, in the present study we manipulate the number of variants in a parametric fashion, from a few dozen to a few thousand.

Feldman et al. (2009) showed, using a Bayesian model, that it can be more beneficial to simultaneously learn a lexicon and phoneme categories than to learn phoneme categories alone. While non-intuitive, this result makes an extremely important conceptual point, as it shows that potential circularities between learning at the lexical and sublexical levels can be broken up using appropriate learning algorithms, where bottom-up and top-down information are learned simultaneously and constrain each other (see also Swingley, 2009). However, their study may also rest on a potential circularity problem since the words provided to the model were all segmented by hand (whereas, as we know, segmentation depends on the availability of abstract enough phonemic categories; Boruta et al., 2011). It may also have a scalability problem, as it modeled the acquisition problem with toy examples, consisting of a small number of artificial or phonetically idealized categories and a small number of words. Here, we will expand on Feldman et al.'s (2009) idea of simultaneous learning of phonemic units and words, while using a realistically sized corpus as in Peperkamp et al. (2006), incorporating the full phoneme inventory, phonotactic constraints and lexicon.

Importantly, in order to avoid the circularity problem mentioned above, we will use an approximate proto-lexicon derived in an unsupervised way from the sublexical representation, without any kind of hand segmentation.

In brief, we combine ideas from both Peperkamp et al. (2006) and Feldman et al. (2009). Following the former, we assume that some kind of initial clustering has yielded a set of discrete segment candidates (the allophonic set). The model's task is then to group them into equivalence classes in order to arrive at abstract phonemes (the phonemic set). Our approach is not aimed at producing a realistic simulation or an instantiated theory of phonological acquisition in infants. Rather, we are interested in quantitatively evaluating the usefulness of different kinds of information that are available in infant's input for the purpose of learning phonological categories.

In Experiment 1 we study the scalability of the bottom-up distributional measure used in Peperkamp et al. (2006) and Le Calvez et al. (2007) when the number of allophones is increased beyond an average of two per phoneme. We also implement two types of allophone: those defined by the following context (Experiments 1 and 2), as in Peperkamp et al. (2006) and Le Calvez et al. (2007), and those conditioned by bilateral contexts (Experiment 3), which mimic the triphones used in ASR and allow us to test an even larger number of allophones per phoneme. In Experiments 2 and 3 we implement a new algorithm, incorporating the idea of Feldman et al. (2009) and Swingley (2009) that feedback from a higher level can help the acquisition of a lower level, even if the higher level has not yet been perfectly learned. To assess this quantitatively, we compare the effect of a perfect (supervised) lexicon with that of an approximate proto-lexicon derived by means of a simple unsupervised segmentation mechanism. We conclude by discussing the predictions of this new model regarding the existence and role of approximate word forms during the first year of life.

2. Experiment 1

In this experiment, we examine the performance of Peperkamp et al.'s (2006) algorithm, which uses Kullback–Leibler (KL) divergence as a measure of distance between contexts to detect pairs of allophones that derive from the same phoneme, on corpora with varying numbers of allophones. A pair of segments with high KL divergence (i.e., having dissimilar distributions) is deemed more likely to belong to the same phoneme than one with low KL divergence. The robustness of this algorithm has been demonstrated using pseudo-random artificial corpora, as well as transcribed corpora of French and Japanese infant-directed speech in which nine and fifteen allophones, respectively, had been added to the phoneme inventory (Le Calvez et al., 2007; Peperkamp et al., 2006). It has also been successfully used on the consonants in the TIMIT English database, whose transcriptions include three allophones in addition to the standard phonemic symbols (Kempton & Moore, 2009). In this experiment we greatly increase the number of allophones in the training data in order to determine whether this method can scale up to systems of realistic complexity.

2.1. Method

2.1.1. Corpora

Starting with a corpus consisting of 7.5 million words of spoken Japanese phonemically transcribed by hand (Maekawa, Koiso, Furui, & Isahara, 2000), we created several versions of the corpus in which artificial rules are used to convert each phoneme into several context-dependent allophones, where a context is defined as a following phoneme or utterance boundary. Given that Japanese has 42 phonemes, the maximum number of distinct contexts for a given phoneme is 43.[2] The corpora we created differ in the number of allophones per phoneme that are implemented, ranging from two (each phoneme has one realization occurring in some contexts and another one occurring in all other contexts) to 43 (each phoneme has as many realizations as there are possible distinct contexts).

For each phoneme *p* in a corpus with *n* allophones per phoneme, *n* rules were generated which convert *p* to one of *n* allophones $p_1 \ldots p_n$ depending on context. Which contexts trigger which allophones was determined by randomly partitioning the set of all possible contexts into *n* partitions, one for each rule, and then randomly assigning contexts to each partition. Note that real allophonic contexts are typically grouped into natural classes based on similarity, a property not shared by our random procedure. Finally, the different versions of the corpus were created by applying the rules to the base corpus. Fig. 1 demonstrates the procedure on one utterance for a corpus with two allophones per phoneme (plus symbols (+) represent word boundaries, which are ignored for the purposes of rule application, and the pound symbol (#) represents an utterance boundary). The notation used in Fig. 1 is read as follows: A rule of the form $X \rightarrow Y / \_\_ \{A, B, C\}$ states that phoneme *X* is realized as allophone *Y* when followed by *A*, *B*, or *C*.

Many of the rules that do not apply in this utterance apply elsewhere in the corpus. Some rules, however, never apply, due to phonotactic sequencing constraints. For instance, the rule assigning the allophone [g_1] to /g/ before any of the segments /w, b^j, m, p, z, p^j, d^j, f/ cannot apply because /g/ only occurs before vowels in Japanese. We therefore measure the

a → a_1 / __ {w, v, t, #, g^j, m^j, ʤ, a, k, o:, r, u, b, i, r^j, ʧ, n, k^j, h, f^j, h^j, p, a:, z, p^j, d^j, e:, d}
a → a_2 / __ {m, s, t^j, b^j, e, u:, o, ts, j, ɴ, ʃ, i:, f, g, n^j}

t → t_1 / __ {w, t^j, a, k, u, e, r^j, o, ts, ʧ, k^j, f^j, m, a:, ʃ, z, d^j, g, d, n^j}
t → t_2 / __ {v, #, g^j, m^j, ʤ, h^j, o:, s, t, b^j, r, u:, b, i:, n, j, h, p, ɴ, p^j, i, e:, f}

m → m_1 / __ {t^j, #, g^j, ʤ, h^j, k, s, t, r, u, b, o, ʧ, n, k^j, j, f^j, m, p, a:, ʃ, z, d^j, e:, f, g, n^j}
m → m_2 / __ {w, v, m^j, a, o:, b^j, e, u:, i:, r^j, ts, h, ɴ, p^j, i, d}

g → g_1 / __ {w, b^j, m, p, z, p^j, d^j, f}
g → g_2 / __ {a, v, g^j, m^j, ʤ, h^j, t^j, k, o:, s, t, r, u, e, u:, b, i:, r^j, o, ts, n, #, k^j, j, h, f^j, ʧ, a:, ɴ, ʃ, i, e:, g, d, n^j}

i → i_1 / __ {w, v, #, g^j, k, o:, u, m, p, ʃ, i, g, d}
i → i_2 / __ { ʧ, m^j, ʤ, h^j, a, s, t, b^j, r, e, u:, b, i:, r^j, o, ts, t^j, n, k^j, j, h, f^j, a:, ɴ, z, p^j, d^j, e:, f, n^j}

Base utterance:	a	t	a	m	a	+	g	a	+	i	t	a	i	#
	↓	↓	↓	↓	↓		↓	↓		↓	↓	↓	↓	
Utterance after rule application:	a_1	t_1	a_2	m_2	a_2	+	g_2	a_1	+	i_2	t_1	a_1	i_1	#

Fig. 1. Example of rule application on the utterance *atama ga itai* ‘‘(my) head hurts,’’ using artificial rules which assign each phoneme one of two allophones.

allophonic complexity of a given corpus by referring to the total number of distinct segments actually occurring in the corpus. The utterance in Fig. 1, for example, contains five phonemes (*a*, *t*, *m*, *g*, *i*) but has an allophonic complexity of seven, because after rule application it contains seven unique segments (a_1, a_2, t_1, m_2, g_2, i_1, i_2).

2.1.2. Procedure

We ran the learning algorithm of Peperkamp et al. (2006) on the corpora as follows. For each corpus, we began by constructing a list of the segments occurring in that corpus. We then listed all logically possible pairs of the attested segments. In order to assess the algorithm's performance on each corpus, the remaining list of segment pairs was divided into two types: those that are allophones of the same phoneme (labeled *same*) and those that are allophones of different phonemes (labeled *different*). The task of the algorithm is to assign the correct label to each pair of allophones, given the corpus as input. This is done by computing the symmetrized KL divergence, a measure of the difference between two probability distributions, for each pair (Appendix). Because allophones of the same phoneme occur in complementary sets of environments, the KL values for such pairs should tend to be higher than the KL values of pairs of allophones of different phonemes.

2.2. Results and discussion

The KL measure can be used to label pairs of sounds by selecting a cutoff value and labeling those pairs with a KL higher than the cutoff as *same* and those pairs with a lower KL as *different*. We evaluate classification performance by means of the ρ statistic (Bamber, 1975), which represents the probability that, given one *same* pair and one *different* pair, each chosen at random, KL divergence assigns a higher value to the *same* pair. Chance is thus represented by a ρ of 0.5, and perfect performance (in which there is no overlap between the two categories) by a ρ of 1.0. Table 1 lists ρ values for each of the corpus types in Experiment 1.

These results show that for the corpora with the lowest allophonic complexity, KL divergence is fairly effective at distinguishing *same* from *different*. This makes sense, because there are many possible ways to divide up the set of contexts, meaning that the probability of

Table 1
Performance of KL divergence as a function of allophonic complexity, expressed as ρ-scores

Mean Allophonic Complexity	ρ
79.0	0.852
164.4	0.692
269.4	0.632
425.8	0.592
567.6	0.562
737.2	0.548

Note. All values are averaged over five corpora generated with the same parameters.

two unrelated allophones in a *different* pair happening to have complementary distributions is relatively low. For the corpora with the maximum number of rules, however, the algorithm performs much less well. When every segment has an extremely narrow distribution, complementary distribution is the rule rather than the exception, and so it is no longer a reliable indicator of pairs of allophones derived from the same phoneme. Unless infants are first able to greatly reduce the number of allophones, examining the distributions of individual segments is not a very efficient way to assign allophones to the appropriate phonemic categories.

3. Experiment 2

The solution we propose to the dilemma posed by high allophonic complexity takes advantage of the fact that the linguistic input is composed of words. A pair of phonological rules which change phoneme x into x_1 when followed by y and into x_2 when followed by z will cause most words that end in the phoneme x to occur in two variants: one that ends in x_1 (when the word is followed by y) and one that ends in x_2 (when the word is followed by z). Thus, encountering a pair of word forms that differ only in that one ends in x_1 and the other ends in x_2 is a clue that x_1 and x_2 are allophones of the same phoneme—conversely, never encountering such a word form pair (in a sufficiently large sample) is a clue that x_1 and x_2 are allophones of different phonemes. The Japanese word *atama* ''head,'' for example, appears as $a_1t_1a_2m_2\boldsymbol{a_2}$ before the nominative marker *ga* in the utterance in Fig. 1, but it would appear as $a_1t_1a_2m_2\boldsymbol{a_1}$ before the word *to* ''and.'' The presence of both word forms in the infant's input is evidence that a_1 and a_2 are allophones of the phoneme /a/. This is not an infallible learning strategy, since every language contains minimal pairs, different words which by chance differ only in a single segment (e.g., *kiku* ''listen'' and *kiki* ''crisis'' in Japanese), which could result in allophones of different phonemes being misclassified as belonging to the same phoneme. However, as long as the number of minimal pairs is small relative to the number of word form pairs derived from the same word, the strategy will be effective.

A learner who knows where the word boundaries are could filter out those segment pairs that are not responsible for multiple word forms, before using KL divergence to classify the remaining pairs. Of course, such a word form filter would unrealistically require perfect knowledge of word boundaries, and thus the set of attested word forms, on the part of the learner. Infants who have not discovered word boundaries yet, however, can approximate the set of word forms by compiling a list of high-frequency strings that occur in the input. In this experiment, we implement both a word form filter and an n-gram filter; the latter is identical to the former except that the set of strings used to construct it is made up of the most frequent n-grams occurring in the corpus for a range of values of n.

3.1. Method

3.1.1. Corpora

We used both the same Japanese corpora as in Experiment 1 and Dutch corpora constructed in a similar manner. Dutch was added as a test language to ensure that any

results are not due to specific properties of Japanese. The base corpus was a nine-million-word corpus of spoken Dutch (Corpus Gesproken Nederlands—Spoken Dutch Corpus; Oostdijk, 2000). The orthographic transcriptions in the Dutch corpus were converted to phonemic transcriptions using the pronunciation lexicon supplied with the corpus.[3] Rules triggered by the following context were then implemented as described for Japanese in Experiment 1, with the difference that the maximum number of contexts and hence of allophones per phoneme in Dutch is 51 (50 phonemes plus the utterance boundary).

3.1.2. Procedure

For the word form filter, two segments *A* and *B* were considered potential allophones of the same phoneme if the corpus contained at least one pair of words *XA* and *XB*, where *X* is a string containing at least three segments. Words shorter than four segments were ignored because of the higher probability of minimally differing pairs occurring by chance among these words. Any segment pairs that did not meet these conditions were labeled as *different*; then, KL divergence was calculated for all remaining pairs as described in section 2.1. The *n*-gram filter works in the same way, except that *XA* and *XB* are frequent *n*-grams rather than words. We used the top 10% most frequent strings of lengths 4, 5, 6, 7, and 8 as surrogate word forms (very short strings tend to generate too many false alarms, while very long strings occur too infrequently to be informative).

Otherwise, the procedure is the same as in Experiment 1.

3.2. Results and discussion

As in Experiment 1, we evaluated the algorithms by means of the ρ statistic. For the word form and *n*-gram filters, KL divergence was computed as in Experiment 1, but only for those pairs passed by the filter. All pairs labeled *different* by the filter were assigned a KL value of -1, so that they were lower than the values of all other pairs. Fig. 2 compares the results of KL divergence in combination with the word form filter and the *n*-gram filter to the results of Experiment 1 (KL alone).

As in Experiment 1, the performance of the KL measure alone degrades as allophonic complexity increases, eventually approaching chance level. This degradation appears to be exponential in shape; that is, ρ drops below 0.7 for corpora containing between 200 and 300 unique segments, showing a rapid loss of performance in the presence of moderate allophonic complexity. In sharp contrast, the performance of the algorithm using the word form filter either increases or slowly decreases with allophonic complexity, with ρ remaining above 0.7 even on corpora of maximal allophonic complexity. Finally, the performance of the algorithm using the *n*-gram filter is intermediate, showing only a moderate decrease in performance with allophonic complexity and a ρ higher than 0.65 on corpora of maximal allophonic complexity.

These results attest to the usefulness of top-down lexical information in learning phonemes, even for data that contain a large number of allophones. Crucially, the *n*-gram filter, although not as effective as the word form filter, is substantially more resistant to allophonic complexity than the KL measure alone.

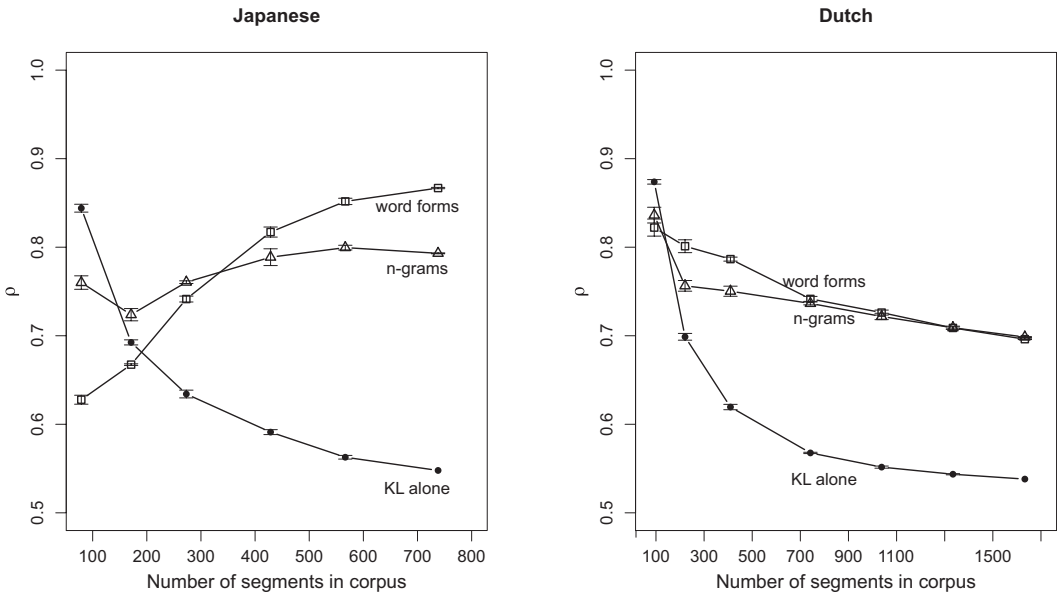

Fig. 2. Performance of allophone clustering (ρ-score) as a function of allophonic complexity measured by the number of following-context allophones in the corpus, for three algorithms (KL alone, KL + word form filter, and KL + *n*-gram filter), on Japanese input (left panel) and Dutch input (right panel). Each point represents the mean performance of the algorithm on five corpora randomly generated using identical parameters. Error bars indicate the standard error across all five corpora.

In order to assess the added value of the KL measure, we compared the *n*-gram filter with KL to the *n*-gram filter with a random measure that simply assigns *same* and *different* labels with a probability of 0.5 each. The extent to which KL contributes to the ρ value above and beyond the contribution made by the filter will be reflected in the size of the difference between the performance of KL and that of the random measure. Table 2 displays this difference (i.e., ρ(filter + KL measure) – ρ(filter + random measure)) for each of the corpora described in Fig. 2.

The table demonstrates that the only corpora for which the KL measure substantially contributes to discriminability are the simplest ones, those averaging fewer than two allophones per phoneme. On all other corpora, KL improves performance either only slightly or not at all. Hence, for all but the simplest corpora, the filter does almost all of the work of increasing discriminability, with little or no contribution made by the KL measure.

In the results presented above, we used *n*-grams ranging from four to eight segments in length. In order to justify this choice of values, and discuss the effects on performance of using *n*-grams of different lengths, we present in Fig. 3 the discrimination performance achieved if individual values for *n* are used instead of the combination of multiple lengths we used.

Several things are clear from this chart. First, using 3-grams alone results in substantially worse performance than any other *n*-gram length. Second, although performance improves

Table 2
Difference in performance of allophone clustering (ρ-score) between KL + n-gram filter and random measure + n-gram filter.

Japanese		Dutch	
Mean Allophonic Complexity	Mean Advantage of KL	Mean Allophonic Complexity	Mean Advantage of KL
79.0	0.223	92.8	0.180
164.4	0.051	220.8	0.051
269.4	0.006	409.6	0.003
425.8	−0.003	741.8	−0.004
567.6	−0.004	1,037.8	−0.002
737.2	0.000	1,334.2	−0.001
		1,633.0	0.000

Note. All values are averaged over five corpora generated with the same parameters.

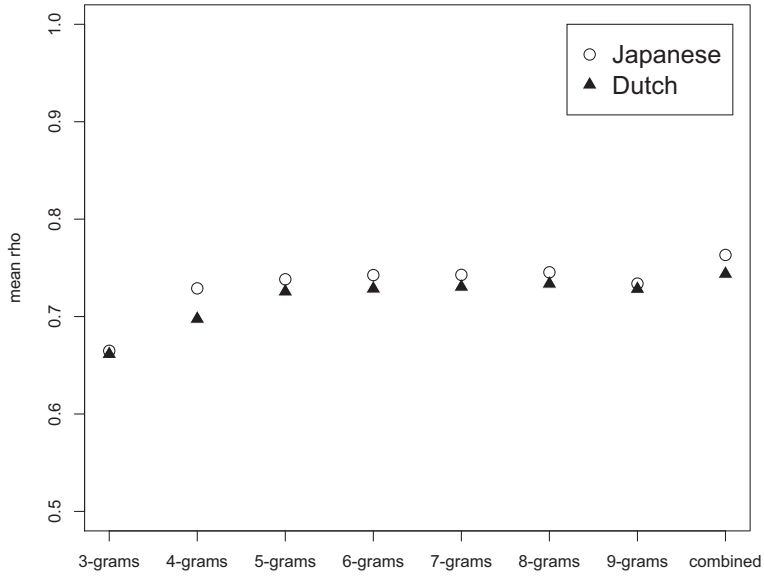

Fig. 3. Performance of allophone clustering (ρ-score) as a function of allophonic complexity measured by the number of following-context allophones in the corpus, for n-grams of lengths 3, 4, 5, 6, 7, 8, and 9 on Japanese input (circles) and Dutch input (triangles). The rightmost point indicates the performance of a combination of 4- through 8-grams. Each point indicates the mean performance over the entire range of corpora for that language.

as n-gram length increases, there is very little difference in the range from 4- to 9-grams. Third, the effects of n-gram length on performance are very similar for the two languages. The fact that n-grams behave similarly in languages as different as Japanese and Dutch offers hope that the 4- to 8-gram range will prove effective for a wide range of languages. Finally, Fig. 3 demonstrates that combining 4- through 8-grams results in better performance than any n-gram length by itself.

Because both the word and n-gram filters rely on minimally differing pairs of word forms, they are vulnerable to noise caused by the occurrence of pairs of words in the input that have

different meanings but happen to differ by a single segment. For example, in Japanese verbs whose non-past forms end in *-ku* have a corresponding imperative form ending in *-ke,* as in *aruku* ''walk'' and *aruke* ''walk-IMP.'' Despite the fact that the vowels /a/ and /e/ are different phonemes in Japanese, the existence of such verbal pairs may prevent the word filter from recognizing these vowels as different phonemes. The extent to which this is a problem for the algorithm, of course, depends on the number of such minimally differing word pairs, compared to the number of word pairs created by the phonological rules. Tables 3 (Japanese) and 4 (Dutch) give, for each corpus type, the numbers of *hits*—allophone pairs correctly passed by the word filter—as well as the number of *false alarms*—allophone pairs derived from different phonemes that are incorrectly passed by the word filter due to the presence of minimal word pairs.

Two trends may be observed in these results. First, unsurprisingly, the number of hits increases as the allophonic complexity increases. This is a straightforward consequence of the fact that the overall number of allophone pairs increases with complexity. More unusual is the relationship between corpus complexity and false alarms—as the number of allophones in the corpus increases, the number of false alarms triggered by minimally differing

Table 3
Numbers of hits versus false alarms passed by the word form filter in Japanese corpora.

Mean Allophonic Complexity	*Same* Pairs Passed by Filter (hits)	*Different* Pairs Passed by Filter (false alarms)
79.0	11.0	210.0
164.4	109.2	942.6
269.4	474.8	2,221.4
425.8	1,818.8	3,176.8
567.6	3,974.0	2,595.2
737.2	7,802.8	0.0

Note. All values are averaged over five corpora generated with the same parameters.

Table 4
Numbers of hits versus false alarms passed by the word form filter in Dutch corpora.

Mean Allophonic Complexity	*Same* Pairs Passed by Filter (hits)	*Different* Pairs Passed by Filter (false alarms)
92.8	29.2	297.8
220.8	247.6	707.0
409.6	918.4	969.4
741.8	2,724.6	1,198.6
1,037.8	5,215.4	960.2
1,334.2	8,099.2	276.4
1,633.0	11,557.0	0.0

Note. All values are averaged over five corpora generated with the same parameters.

words at first increases, and then decreases. This U-shaped trend is caused by two opposing forces: first, as with hits, the number of possible allophone pairs increases with the number of allophones. Second, however, as the number of allophones increases, the range of contexts assigned to each allophone shrinks. This means that the penultimate segments in the two words will be less likely to be grouped in the same phonemic category. To use the Japanese word for ''walk'' as an example, the only way that *aruku* and *aruke* will be mistakenly categorized as variants of the same word is if the allophones of /k/ that occur in each word are also treated as belonging to the same phonemic category, a mistake which is unlikely if /k/ is split into a high number of allophones, and impossible if /k/ is split into the maximum possible number of allophones (since ''/k/ before /u/'' and ''/k/ before /e/'' will always be treated as different allophones).

In short, the more attention the learner pays to the fine phonetic detail of each allophone, the odds of accidentally mistaking a pair of different words for a pair of word form variants of a single word decrease. This type of mistake only becomes dangerous once the infant has constructed fairly large and abstract categories, meaning that minimal pairs like *aruku* and *aruke* are unlikely to pose a serious problem in the early stages of category learning.

4. Experiment 3

Experiments 1 and 2 use phonological rules that are unilaterally conditioned, in particular, rules that are triggered by the phoneme's following context. Actual phonological processes in natural languages, however, are often conditioned by bilateral contexts. In Korean, for example, stop consonants become voiced when both preceded and followed by a voiced segment (Cho, 1990). In this Experiment we therefore test the algorithms used in Experiments 1 and 2 on data in which allophones are dependent on both the preceding and the following segment.

4.1. Method

4.1.1. Corpora

The corpora in Experiment 3 are based on the same Japanese and Dutch corpora used in Experiment 2. Allophonic rules were implemented as in the previous experiment, with the difference that each context consisted of both a preceding and a following segment.

4.1.2. Procedure

The implementation of the word form and *n*-gram filters was performed as in Experiment 2, with the exception of how relevant word form pairs were identified. Because the corpora in this experiment contain allophones that are conditioned by bilateral contexts, a pair of word forms (or *n*-grams) was considered relevant if either or both the initial and final segments differ. Thus, if the corpus contains two word forms *AXC* and *BXD*, where *X* is a string containing at least two segments, the pair of segments *A* and *B* are considered potential

allophones of the same phoneme, as is the pair *C* and *D*. This procedure is able to discover both unilaterally conditioned and bilaterally conditioned rules, and it would be effective in a language with both types of rule. In this experiment, however, we implement only bilateral rules in the training data, as the large number of contexts make this the most complex possible learning scenario.

4.2. Results and discussion

As in Experiment 2, we compare the results of KL divergence in combination with the word form filter and the *n*-gram filter to the results of Experiment 1 (KL alone). The results are shown in Fig. 4.

As in the previous experiments, the KL measure alone yields an exponential drop in performance as the allophonic complexity increases, while the algorithms incorporating a word form filter or an *n*-gram filter display a stronger resistance to allophonic complexity; for the latter algorithms, performance is around 0.8 and 0.7, respectively, for corpora with maximum allophonic complexity. A comparison of these results with the ones of Experiment 2 reveals that increasing the complexity of the rules themselves by making them sensitive to bilateral contexts does not substantially affect the performance of the two filters—in fact, performance is slightly better for bilateral contexts (this Experiment) than for unilateral contexts (Experiment 2).

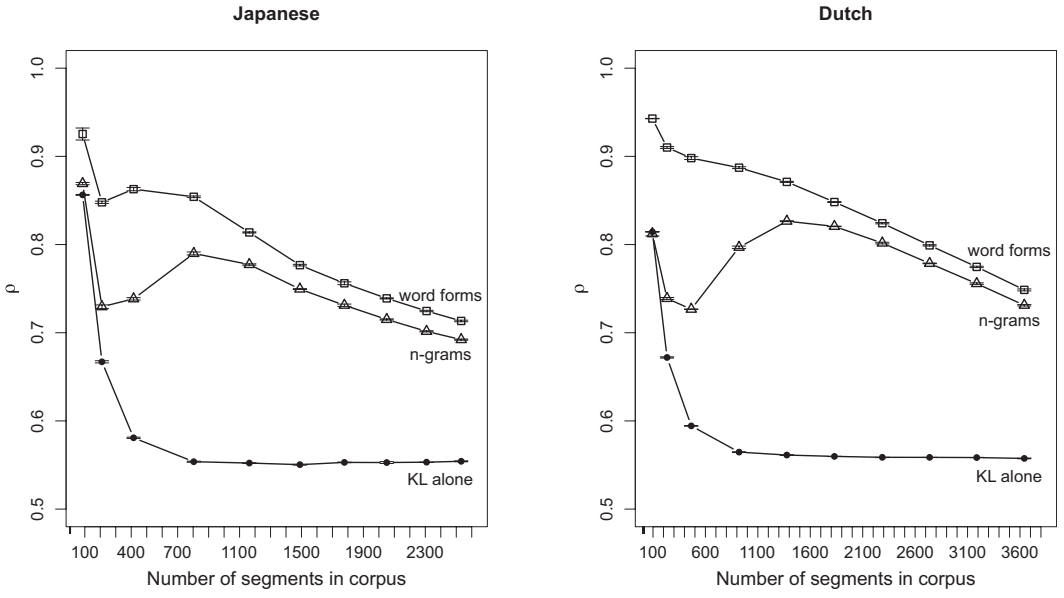

Fig. 4. Performance of allophone clustering (ρ-score) as a function of allophonic complexity measured by the number of bilateral allophones in the corpus, for three algorithms (KL alone, KL + word form filter, and KL + *n*-gram filter), on Japanese input (left panel) and Dutch input (right panel). Each point represents the mean performance of the algorithm on five corpora randomly generated using identical parameters. Error bars indicate the standard error across all five corpora.[5]

5. General discussion

The development of phonetic perception over the first year of life poses a conundrum. By the end of this period, despite having little access to semantic information, infants treat semantically meaningful (i.e., phonemic) and meaningless (i.e., allophonic) phonetic distinctions differently. We have argued that one way out of this conundrum for infants involves building phonemic categories, effectively classifying these distinctions as either phonemic or allophonic, using a procedure that exploits the lexical structure of the input.

We have shown, firstly, that when searching for phonemic categories, a bottom-up procedure which looks for sounds that are in complementary distribution becomes extremely inefficient when allophonic complexity (i.e., the number of allophones) increases. Secondly, we found that adding top-down lexical word form information allows for robust discrimination among segment pairs that belong to the same phoneme and those that belong to different phonemes, even in the presence of increased allophonic complexity. Finally, we have shown that lexical word forms can be crudely approximated with *n*-grams, which still yield results that are both good and resistant to allophonic complexity. These results are obtained with the same types of contextual variants as used in Peperkamp et al. (2006) and Le Calvez et al. (2007), that is, allophones that depend upon the following context (Experiments 1 and 2), as well as with bilaterally conditioned allophones (Experiment 3). Moreover, the results hold for both Dutch and Japanese, two languages that are very different from the viewpoint of syllable structure and phonotactics.

The reason for the lack of robustness of the bottom-up complementary distribution approach is fairly simple to grasp: as the number of allophones increases, the allophones become more and more tied to highly specific contexts. Ultimately, in a language where all possible bilateral allophones are instantiated, every segment is in complementary distribution with almost every other one, rendering distributional information nearly useless for phonemic categorization. Only when the number of allophonic variants is very small (not more than two per phoneme), and complementary distribution of segments thereby relatively rare, is this type of distributional information useful. Unfortunately for the learner, this means that looking for complementary distribution between segments in the input is only an efficient strategy when the problem has already been almost completely solved.

The top-down approach is successful because it relies on the fact that allophony changes not just individual sounds, but entire word forms, and that for sufficiently long words, the probability that two different words happen to be identical except for their final sounds is very low. Crucially, this fact is independent of the allophonic complexity of the input. Of course, this criterion alone is not sufficient, as there are true minimal pairs in many languages, but they are relatively rare (especially for longer words) and non-systematic, unlike the minimal pairs created by contextual allophony. Finally, the reason for the success of the *n*-gram strategy is that the low probability of true minimal pairs also applies to frequent *n*-grams, and this probability does not depend on allophonic complexity.

Our algorithm could be improved and extended in a number of ways. First, instead of using top-down information in an all-or-none fashion, we could implement a statistical

model of possible word forms (Swingley, 2005) and use it to compute the probability of obtaining the observed pattern of minimal pairs by accident. Such a procedure would allow us to probabilistically integrate the effect of word length instead of postulating a fixed window of word lengths as in the present study, and it would also be less sensitive to the occasional existence of true minimal pairs. Second, the crude *n*-gram segmentation strategy could be replaced by some less crude—although still sub-optimal—lexical segmentation procedure (e.g., Brent, 1999; Daland & Pierrehumbert, 2010; Goldwater et al., 2009; Monaghan & Christiansen, 2010; Venkataraman, 2001). Third, as noted in Peperkamp et al. (2006), performance can be improved by providing phonetic constraints as to the nature of allophonic processes. Peperkamp et al., 2006. proposed two such constraints, one to the effect that allophones of the same phoneme must be minimally different from a phonetic point of view, and another to the effect that allophones tend to result from contextual assimilation (feature spreading). How such phonetic constraints can be implemented in a language encoded with massive allophony remains to be assessed. Another example of a possibly helpful linguistic constraint is rudimentary semantic knowledge, which could serve as further evidence that two word forms are in fact realizations of a single word. Even if infants do not know many words, the few words they do know could improve the performance of the *n*-gram filter (Swingley, 2009). Fourth, our model could be extended to learn patterns of phonological variations that go beyond contextual allophony. Word forms can also vary through processes of insertion or deletion of segments, yielding a multiplication of closely related word forms (Greenberg, 1998). A proto-lexicon has the potential of capturing some of these variations, which would be impossible to do in a purely bottom-up fashion (Peperkamp & Dupoux, 2002).

Of course, the procedure we have described represents only the first step in the learning of phonemic categories. Our algorithm assigns each pair of allophones a rating which indicates how likely that pair is to belong to the same phoneme; a real infant would need to then use these ratings to group allophones into phoneme-sized clusters of allophones. This next step is not trivial. Given pairs of allophones and ratings, the infant must decide on an appropriate cutoff value above which allophones will be considered members of the same cluster. Choosing the optimal cutoff will depend on the relative costs of false alarms (allophones incorrectly grouped together) and misses (allophones incorrectly placed in different clusters). At this point, very little is known about what these costs might be or how easy it is for infants to recover from errors at this stage of learning. Although modeling this entire process is thus beyond the scope of the present article, we hope that our results provide a foundation for future progress on these questions.

We also emphasize that, although we have couched our proposal in terms of a specific algorithm, our findings allow us to draw more general conclusions that go beyond the question of whether infants use this exact procedure to learn phonemes. These results demonstrate, first, that the lexical structure of speech input contains information on phonemic categories that is missed by approaches which focus only on sublexical units, and second, that this lexical information can be extracted from data of realistic complexity, even using an extremely simple procedure. It is therefore likely that any approach to learning phonemes would benefit from making use of top-down lexical information.

We should, however, point out that our approach employs a number of simplifications that make it unable to address the entire complexity of the acquisition problem. First, the assumption that the learner starts by establishing a large set of discrete allophones may not adequately capture some of the phonetic effects of between-talker variation, nor the more continuous effects of speaking rate, or variability induced by noise in the transmission channel. Clearly, adequate signal preprocessing/normalization is needed if a running speech application is envisioned. Second, as mentioned earlier, the use of a minimal pair constraint may be problematic in languages with mono-segmental inflectional affixes that create systematic patterns of word-final or word-initial minimal pairs (as in the Japanese *aruku-aruke* example mentioned in section 3.2). To solve these cases, the proto-lexicon of word forms must be supplemented with semantic information which may only be acquired later during development (Regier, 2003). This is consistent with the view that the acquisition of phonemic categories is not closed by the end of the first year of life but continues to be refined thereafter (Sundara, Polka, & Genesee, 2006). Third, we should note that our procedure can only discover phonological rules that operate at word boundaries; a context-dependent rule that only ever applies within a word will not create word form variants in the way discussed here and will have to be learned in some other way. But precisely because such rules, if applied consistently, do not create multiple word forms, they do not impede word recognition or segmentation and so are not as crucial for a language-learning infant.

The traditional bottom-up scenario of early language acquisition holds that infants begin by learning the phonemes and constraints upon their sequencing during the first year of life (Jusczyk et al., 1993; Pegg & Werker, 1997; Werker & Tees, 1984) and then learn the lexicon on the basis of an established phonological representation (Clark, 1995). While infants have been shown to be capable of extracting phonological regularities in a non-lexical, bottom-up fashion (Chambers, Onishi, & Fisher, 2003; Maye et al., 2002; Saffran & Thiessen, 2003; White, Peperkamp, Kirk, & Morgan, 2008), our results cast serious doubt on the idea that such a mechanism is by itself sufficient to establish phonological categories. Indeed, we have shown that attempts to re-cluster allophones in a bottom-up fashion based on complementary distributions is inefficient in the face of massive allophony. However, we also showed that it is possible to replace the bottom-up scenario with one that is nearly the reverse, in which an approximate lexicon of word forms is used to acquire phonological regularities. In fact, an interactive scenario could be proposed, in which an approximate phonology is used to yield a better lexicon, which in turn is used to improve the phonology, and so on, until both phonology and the lexicon converge on the adult form (Swingley, 2009).

The present approach opens up novel empirical research questions in infant language acquisition. For instance, does a proto-lexicon exist in infants before 12 months of age? If so, what are its size and growth rate? Ngon et al. (in press) provide preliminary answers to these questions. They found that 11-month-old French-learning infants are able to discriminate highly frequent *n*-grams from low-frequency *n*-grams, even when neither set of stimuli contained any actual words. This suggests that at this age, infants have indeed constructed a proto-lexicon of high-frequency sequences which consists of a mixture of words and

nonwords. The Ngon et al. (in press). study raises the question of how to estimate the size and composition of the proto-lexicon as a function of age. Such an estimation should be linked to modeling studies like the present one, in order to determine the extent to which the protolexicon can help the acquisition of phonological categories.[4]

There are other questions raised by our results that will be more difficult to answer. In particular, does the growth of the proto-lexicon predict the acquisition of phonological categories and of phonological variation? And how does the acquisition of phonology help the acquisition of a proper lexicon of word forms? These questions have been neglected in the past, perhaps because of the belief that a proper lexicon cannot be learned before phonemic categories are acquired. The present results, however, suggest that an understanding of lexical acquisition will be a fundamental component of a complete theory of phonological acquisition. Clearly, more research is needed to understand the mechanisms that could make it possible to simultaneously learn lexical and phonological regularities, and whether infants can use these mechanisms during the first year of life.

Appendix: Kullback-Leibler measure of dissimilarity between two probability distributions

Let s be a segment, c a context, and $P(c|s)$ the probability of observing c given s. Then the Kullback-Leibler measure of dissimilarity between the distributions of two segments s_1 and s_2 is defined as:

$$m_{KL}(s_1, s_2) = \sum_{c \in C} \left(P(c|s_1) \quad \log\left(\frac{P(c|s_1)}{P(c|s_2)}\right) \; + \; P(c|s_2) \quad \log\left(\frac{P(c|s_2)}{P(c|s_1)}\right)\right)$$

with $P(c|s) = \frac{n(c,s)+1}{n(s)+N}$

with $n(c,s)$ the number of occurrences of the segment s in the context c (i.e., the number of occurrences of the sequence sc),

$n(s)$ the number of occurrences of the segment s,

and N the total number of contexts.

In order to smooth the probability estimates of the distributions in finite samples, $1/N$ occurrence of each segment is added in each context, where N is the total number of contexts.

Notes

1. Dale and Fenson (1996) found that English-learning 11-month-old infants comprehended an average of 113 words.
2. Note that contexts for our rules are limited to underlying phonemes. In actual languages, the outputs of some rules can serve as the inputs to other rules, further complicating the learning process.
3. A number of words present in the Dutch orthographic corpus (largely proper nouns) were not listed in the pronunciation lexicon. We eliminated any utterances containing

these words from our corpora, resulting in a roughly 20% reduction in corpus size.

4. For instance, the present algorithm uses the 10% most frequent *n*-grams as a protolexicon. Given the size of the corpus, this turns out to be a rather large set (over a million word forms). The use of a more realistic segmentation procedure would certainly cut down this number and bring it closer to the size of the protolexicon as it could be measured in infants.
5. With bilateral contexts, implementing our algorithm becomes computationally prohibitive on the most complex corpora. The rightmost points in Fig. 4 represent the most complex corpora we were able to process given our available resources.

Acknowledgments

This research was made possible by support from the Centre National de la Recherche Scientifique and the RIKEN Brain Science Institute, in addition to grants ANR-2010-BLAN-1901-1 from the Agence Nationale pour la Recherche and ERC-2011-AdG-295810 from the European Research Council.

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*, 387–415.

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30* (4), 591–627.

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134.

Boruta, L., Peperkamp, S., Crabbé, B., & Dupoux, E. (2011). Testing the robustness of online word segmentation: Effects of linguistic diversity and phonetic variation. *Proceedings of CMCL, ACL.*, *2*, 1–9.

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

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77.

Cho, Y. Y. (1990). Syntax and phrasing in Korean. In S. Inkelas & D. Zec (Eds.), *The phonology-syntax connection* (pp. 47–62). Chicago: University of Chicago Press.

Choi, J. D., & Keating, P. (1991). Vowel-to-vowel coarticulation in three Slavic languages. *UCLA Working Papers in Phonetics*, *78*, 78–86.

Clark, E. V. (1995). *The lexicon in acquisition*. Cambridge, England: Cambridge University Press.

Clements, G. N. (2009). The role of features in phonological inventories. In E. Raimy & C. E. Cairns (Eds.), *Contemporary views on architecture and representations in phonological theory* (pp. 19–68). Cambridge, MA: MIT Press.

Cutler, A., Eisner, F., McQueen, J., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. *Papers in Laboratory Phonology*, *10*, 91–111.

Daland, R., & Pierrehumbert, J. (2010). Learning diphone-based segmentation. *Cognitive Science*, *35*(1), 119–155.

Dale, P., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods*, *28*(1), 125–127.

Dehaene-Lambertz, G., & Baillet, S. (1998). A phonological representation in the infant brain. *NeuroReport*, *9*(8), 1885.

Dresher, B. E., & Kaye, J. D. (1990). A computational learning model for metrical phonology. *Cognition*, *34*(2), 137–195.

Feldman, N., Griffiths, T., & Morgan, J. (2009). Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2208–2213.

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research*, *24*, 127–139.

Fowler, C. A., & Smith, M. (1986). Speech perception as ''vector analysis'': An approach to the problems of segmentation and invariance. In J. S. Perkell & D. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 123–136). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gauthier, B., Shi, R., & Xu, Y. (2007a). Learning phonetic categories by tracking movements. *Cognition*, *103* (1), 80–106.

Gauthier, B., Shi, R., & Xu, Y. (2007b). Simulating the acquisition of lexical tones from continuous dynamic input. *Journal of the Acoustical Society of America*, *121*(5), EL190–EL195.

Goldsmith, J., & Xanthos, A. (2009). Learning phonological categories. *Language*, *85*(1), 4–38.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

Gow, D. W. Jr, & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(2), 344–359.

Greenberg, S. (1998). Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation. *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, 47–56.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, *100*(2), 1111–1121.

Jelinek, F. (1998). *Statistical Methods of Speech Recognition*. Cambridge, MA: MIT Press.

Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*, *21*, 3–28.

Jusczyk, P. (2000). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23.

Jusczyk, P. W., Friederici, A., Wessels, J., Svenkerud, V., & Jusczyk, A. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32* (3), 402–420.

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics*, *61*(8), 1465.

Kempton, T., & Moore, R. K. (2009). Finding allophones: An evaluation on consonants in the TIMIT corpus. *Interspeech*, *2009*, 1651–1654.

Kraljik, T., & Samuel, A. (2005). Perceptual learning in speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B*, *363*, 979–1000.

Le Calvez, R., Peperkamp, S., & Dupoux, E. (2007). Bottom-up learning of phonemes: a computational study. *Proceedings of the Second European Cognitive Science Conference*, *2*, 167–172.

Lee, K.-F. (1988). On the use of triphone models for continuous speech recognition. *JASA*, *84*(S1), S216–S216.

Lehiste, I., & Shockey, L. (1972). On the perception of coarticulation effects in English VCV syllables. *Journal of Speech, Language, and Hearing Research*, *15* (3), 500–506.

Maekawa, K., Koiso, H., Furui, S., & Isahara, H. (2000). Spontaneous speech corpus of japanese. *Proceedings of LREC*, *2*, 947–952.

Makhoul, J., & Schwartz, R. (1995). State of the art in continuous speech recognition. *Proceedings of the National Academy of Sciences*, *92*, 9956–9963.

Manuel, S. (1999). Cross-language studies: relating language-particular coarticulation patterns to other language-particular facts. In W. J. Hardcastle & N. Hewlett (Eds.), *Coarticulation: Theory, data and techniques* (pp. 179–198). Cambridge, UK: Cambridge University Press.

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

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Phonetic details in perception and production allow various patterns in phonological change. *Cognition*, *82*(3), B101–B111.

McMurray, B., & Aslin, R. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, *95*(2), B15–B26.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1609–1631.

Mcmurray, B., Aslin, R., & Toscano, J. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*(3), 369–378.

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *J. Child Lang.*, *37*(03), 545.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, *62*(3), 714–719.

Ngon, C., Martin, A. T., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (In press). Nonwords, nonwords, nonwords: Evidence for a proto-lexicon during the first year of life. Developmental Science.

Norris, D., McQueen, J., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, *39*, 151–168.

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first evaluation. *Proceedings of LREC-2000, Athens*, *2*, 887–894.

Pegg, J. E., & Werker, J. F. (1997). Adult and infant perception of two English phones. *Journal of the Acoustical Society of America*, *102*(6), 3742–3753.

Peperkamp, S., & Dupoux, E. (2002). Coping with phonological variation in early lexical acquisition. In I. Lasser (Ed.), *The Process of Language Acquisition: Proceedings of the 1999 GALA Conference* (pp. 359–385). Frankfurt: Peter Lang.

Peperkamp, S., Le Calvez, R., Nadal, J., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, *101*(3), B31–B41.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*(2-3), 115–154.

Ramus, F., Peperkamp, S., Christophe, A., Jacquemot, C., Kouider, S., & Dupoux, E. (2010). A psycholinguistic perspective on the acquisition of phonology. *Papers in Laboratory Phonology*, *10*, 311–340.

Regier, T. (2003). Emergent constraints on word-learning: A computational perspective. *Trends in Cognitive Sciences*, *7*(6), 263–268.

Rytting, C., Brew, C., & Fosler-Lussier, E. (2010). Segmenting words from natural speech: subsegmental variation in segmental cues. *Journal of Child Language*, *37*(3), 513.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*(3), 484–494.

Seidl, A., Cristia, A., Bernard, A., & Onishi, K. H. (2009). Allophonic and phonemic constrasts in infants' learning of sound patterns. *Language Learning and Development*, *5*(3), 191–202.

Sundara, M., Polka, L., & Genesee, F. (2006). Language-experience facilitates discrimination of/d-/in monolingual and bilingual acquisition of English. *Cognition*, *100*(2), 369–388.

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

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

Tesar, B., & Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, *29* (2), 229–268.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *PNAS*, *104*(33), 13273–13278.

Varadarajan, B., Khudanpur, S., & Dupoux, E. (2008). Unsupervised learning of acoustic sub-word units. *Proceedings of ACL-08: HLT, Short Papers (Companion Volume)*, *46*, 165–168.

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

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

Werker, J. F., & Tees, R. C. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology*, *50*, 509–535.

White, K. S., Peperkamp, S., Kirk, C., & Morgan, J. L. (2008). Rapid acquisition of phonological alternations by infants. *Cognition*, *107*(1), 238–265.