A Rudimentary Lexicon and Semantics Help Bootstrap Phoneme Acquisition

Abdellah Fourtassi
abdellah.fourtassi@gmail.com

Emmanuel Dupoux
emmanuel.dupoux@gmail.com

Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS, Paris

Abstract

Infants spontaneously discover the relevant phonemes of their language without any direct supervision. This acquisition is puzzling because it seems to require the availability of high levels of linguistic structures (lexicon, semantics), that logically suppose the infants having a set of phonemes already. We show how this circularity can be broken by testing, in real-size language corpora, a scenario whereby infants would learn approximate representations at all levels, and then refine them in a mutually constraining way. We start with corpora of spontaneous speech that have been encoded in a varying number of detailed context-dependent allophones. We derive, in an unsupervised way, an approximate lexicon and a rudimentary semantic representation. Despite the fact that all these representations are poor approximations of the ground truth, they help reorganize the fine grained categories into phoneme-like categories with a high degree of accuracy.

One of the most fascinating facts about human infants is the speed at which they acquire their native language. During the first year alone, i.e., before they are able to speak, infants achieve impressive landmarks regarding three key language components. First, they tune in on the phonemic categories of their language (Werker and Tees, 1984). Second, they learn to segment the continuous speech stream into discrete units (Jusczyk and Aslin, 1995). Third, they start to recognize frequent words (Ngon et al., 2013), as well as the semantics of many of them (Bergelson and Swingley, 2012).

Even though these landmarks have been documented in detail over the past 40 years of research, little is still known about the mechanisms that are operative in infant's brain to achieve such a result. Current work in early language acquisition has proposed two competing but incomplete hypotheses that purports to account for this stunning development path. The bottom-up hypothesis holds that infants converge onto the linguistic units of their language through a statistical analysis over of their input. In contrast, the top-down hypothesis emphasizes the role of higher levels of linguistic structure in learning the lower level units.

1 A chicken-and-egg problem

1.1 Bottom-up is not enough

Several studies have documented the fact that infants become attuned to the native sounds of their language, starting at 6 months of age (see Gervain & Mehler, 2010 for a review). Some researchers have claimed that such an early attunement is due to a statistical learning mechanism that only takes into account the distributional properties of the sounds present in the native input (Maye et al., 2002). Unsupervised clustering algorithms running on simplified input have, indeed, provided a proof of principle for bottom-up learning of phonemic categories from speech (see for instance Vallabha et al., 2007).

It is clear, however, that distributional learning cannot account for the entire developmental pattern. In fact, phoneme tokens in real speech exhibit high acoustic variability and result in phonemic categories with a high degree of overlap (Hillenbrand et al., 1995). When purely bottom up clustering algorithms are tested on realistic input, they ended up in either a too large number of subphonemic units (Varadarajan et al., 2008) or a too small number of coarse grained categories (Feldman et al., 2013a).

1.2 The top-down hypothesis

Inspection of the developmental data shows that infants do not wait to have completed the acquisition of their native phonemes to start to learn words. In fact, lexical and phonological acquisition largely overlap. Infant can recognize highly frequent word forms like their own names, by as early as 4 months of age (Mandel et al., 1995). Vice versa, the refinement of phonemic categories does not stop at 12 months. The sensitivity to phonetic contrasts has been reported to continue at 3 years of age (Nittrouer, 1996) and beyond (Hazan and Barrett, 2000), on par with the development of the lexicon.

Some researchers have therefore suggested that there might be a learning synergy which allows infants to base some of their acquisition not only on bottom up information, but also on statistics over lexical items or even on the basis of word meaning (Feldman et al., 2013a; Feldman et al., 2013b; Yeung and Werker, 2009)

These experiments and computational models, however, have focused on simplified input or/and used already segmented words. It remains to be shown whether the said top-down strategies scale up when real size corpora and more realistic representations are used. There are indeed indications that, in the absence of a proper phonological representation, lexical learning becomes very difficult. For example, word segmentation algorithms that work on the basis of phoneme-like units tend to degrade quickly if phonemes are replaced by contextual allophones (Boruta et al., 2011) or with the output of phone recognizers (Jansen et al., 2013; Ludusan et al., 2014).

In brief, we are facing a chicken-and-egg problem: lexical and semantic information could help to learn the phonemes, but phonemes are needed to acquire lexical information.

1.3 Breaking the circularity: An incremental discovery procedure

Here, we explore the idea that instead of learning adult-like hierarchically organized representations in a sequential fashion (phonemes, words, semantics), infants learn approximate, provisional linguistic representations in parallel. These approximate representations are subsequently used to improve each other.

More precisely, we make four assumptions. First, we assume that infants start by paying attention to fine grained variation in the acoustic input, thus constructing perceptual phonetic categories that are not phonemes, but segments encoding fine grained phonetic details (Werker and Curtin, 2005; Pierrehumbert, 2003). Second, we assume that these units enable infants to segment proto-words from continuous speech and store them in this detailed format. Importantly, this proto-lexicon will not be adult-like: it will contain badly segmented word forms, and store several alternant forms for the same word. Ngon et al. (2013) have shown that 11 month old infants recognize frequent sound sequences that do not necessarily map to adult words. Third, we assume that infants can use this imperfect lexicon to acquire some semantic representation. As shown in Shukla et al. (2011), infants can simultaneously segment words and associate them with a visual referent. Fourth, we assume that as their exposure to language develops, infants reorganize these initial categories along the relevant dimensions of their native language based on cues from all these representations.

The aim of this work is to provide a proof of principle for this general scenario, using real size corpora in two typologically different languages, and state-of-the-art learning algorithms.

The paper is organized as follows. We begin by describing how we generated the input and how we modeled different levels of representation. Then, we explain how information from the higher levels (word forms and semantics) can be used to refine the learning of the lower level (phonetic categories). Next, we present the results of our simulations and discuss the potential implications for the language learning process.

2 Modeling the representations

Here, we describe how we model different levels of representation (phonetic categories, lexicon and semantics) starting from raw speech in English and Japanese.

2.1 Corpus

We use two speech corpora: the Buckeye Speech corpus (Pitt et al., 2007), which contains 40 hours of spontaneous conversations in American English, and the 40 hours core of the Corpus of Spontaneous Japanese (Maekawa et al., 2000), which contains spontaneous conversations and public speeches in different fields, ranging from engineering to humanities. Following Boruta (2012),

we use an inventory of 25 phonemes for transcribing Japanese, and for English, we use the set of 45 phonemes in the phonemic transcription of Pitt et al. (2007).

2.2 Phonetic categories

Here, we describe how we model the perceptual phonetic categories infants learn in a first step before converging on the functional categories (phonemes). We make the assumption that these initial categories correspond to fine grained *allophones*, i.e., different systematic realizations of phonemes, depending on context. Allophonic variation can range from categorical effects due to phonological rules to gradient effects due to coarticulation, i.e, the phenomenon whereby adjacent sounds affect the physical realization of a given phoneme. An example of a rather categorical allophonic rule is given by /r/ devoicing in French:

$$/\text{ʁ}/ \rightarrow \begin{cases} [\chi] \,/ & \text{before a voiceless obstruent} \\ [\text{ʁ}] & \text{elsewhere} \end{cases}$$

Figure 1: Allophonic variation of French /r/

The phoneme /r/ surfaces as voiced ([ʁ]) before a voiced obstruent like in [kanaʁ ʒon] ("canard jaune", yellow duck) and as voiceless ([χ]) before a voiceless obstruent as in [kanaχ puχpʁ] ("canard pourpre", purple duck). The challenge facing the leaner is, therefore, to distinguish pairs of segments that are in an allophonic relationship ([ʁ], [χ]) from pairs that are two distinct phonemes and can carry a meaning difference ([ʁ],[l]).

Previous work has generated allophonic variation artificially (Martin et al., 2013). Here, we follow Fourtassi et al. (2014b) in using a linguistically and statistically controlled method, starting from audio recordings and using a standard Hidden Markov Models (HMM) phone recognizer to generate them, as follows.

We convert the raw speech waveform into successive 10ms frames containing a vector of Mel Frequency Cepstrum Coefficients (MFCC). We use 12 MFC coefficients (plus the energy) computed over a 25ms window, to which we add the first and second order derivatives, yielding 39 dimensions per frame.

The HMM training starts with one three-state model per phoneme. Each state is modeled by a mixture of 17 diagonal Gaussians. After training, each phoneme model is cloned into context-dependent triphone models, for each context in which the phoneme actually occurs (for example, the phoneme /ɑ/ occurs in the context [d–ɑ–g] as in the word /dɑg/ ("dog"). The triphone models cloned from the phonemes are then retrained, but, this time, only on the relevant subset of the data, corresponding to the given triphone context. Finally, these detailed models are clustered back into inventories of various sizes (from 2 to 20 times the size of the phonemic inventory) and retrained. Clustering is done state by state using a phonetic feature-based decision tree, and results in tying together the HMM states of linguistically similar triphones so as to maximize the likelihood of the data. The HMM were built using the HMM Toolkit (HTK: Young et al., 2006).

2.3 The proto-lexicon

Finding word boundaries in the continuous sequence of phones is part of the problem infants have to solve without direct supervision. We model this segmentation using a state-of-the-art unsupervised word segmentation model based on the Adaptor Grammar framework (Johnson et al., 2007). The input consists of a phonetic transcription of the corpus, with boundaries between words eliminated (we vary this transcription to correspond to different inventories with different granularity in the allophonic representation as explained above). The model tries to reconstruct the boundaries based on a Pitman-Yor process (Pitman and Yor, 1997), which uses a language-general statistical learning process to find a compact representation of the input. The algorithm stores high frequency chunks and re-uses them to parse novel utterances. We use a grammar which learns a hierarchy of three levels of chunking and use the intermediate level to correspond to the lexical level. This grammar was shown by Fourtassi et al. (2013) to avoid both over-segmentation and under-segmentation.

2.4 The proto-semantics

It has been shown that infants can keep track of co-occurrence statistics (see Lany and Saffran (2013) for a review). This ability can be used to develop a sense of semantic similarity as suggested by Harris (1954). The intuition behind the *distributional hypothesis* is that words that are similar in meaning occur in similar contexts. In order to model the acquisition of this semantic similarity from a

transcribed and segmented corpus, we use one of the simplest and most commonly used distributional semantic models, Latent Semantic Analysis (LSA: Landauer & Dumais, 1997). The LSA algorithm takes as input a matrix consisting of rows representing word types and columns representing contexts in which tokens of the word type occur. A context is defined as a fixed number of utterances. Singular value decomposition (a kind of matrix factorization) is used to extract a more compact representation. The cosine of the angle between vectors in the resulting space is used to measure the semantic similarity between words. Two words have a high semantic similarity if they have similar distributions, i.e., if they co-occur in most contexts. The model parameters, namely the dimension of the semantic space and the number of utterances to be taken as defining the context of a given word form, are set in an unsupervised way to optimize the latent structure of the semantic model (Fourtassi and Dupoux, 2013). Thus, we use 20 utterances as a semantic window and set the semantic space to 100 dimensions.

3 Method

Here we explore whether the approximate high level representations, built bottom-up and without supervision, still contain useful information one can use to refine the phonetic categories into phoneme-like units. To this end, we extract potential cues from the lexical and the semantic information, and test their performance in discriminating allophonic contrasts from non-allophonic (phonemic) contrasts.

3.1 Top down cues

3.1.1 Lexical cue

The top down information from the lexicon is based on the insight of Martin et al. (2013). It rests on the idea that true lexical minimal pairs are not very frequent in human languages, as compared to minimal pairs due to mere phonological processes (figure 1). The latter creates alternants of the same lexical item since adjacent sounds condition the realization of the first and final phoneme. Therefore, finding a minimal pair of words differing in the first or last segment (as in [kanaχ] and [kanaʁ]) is good evidence that these two phones ([ʁ], [χ]) are allophones of one another. Conversely, if a pair of phones is not forming any minimal pair, it is classified as non-allophonic (phonemic).

However, this binary strategy clearly gives rise to false alarms in the (albeit relatively rare) case of true minimal pairs like [kanaχ] ("duck") and [kanal] ("canal"), where ([χ], [l]) will be mistakenly labeled as allophonic. In order to mitigate the problem of false alarms, we use Boruta's continuous version (Boruta, 2011) and we define the lexical cue of a pair of phones $Lex(x, y)$ as the number of lexical minimal pairs that vary on the first segment (xA, yA) or the last segment (Ax, Ay). The higher this number, the more the pair of phones is likely to be considered as allophonic.

The lexical cue is consistent with experimental findings. For example Feldman et al. (2013b) showed that 8 month-old infants pay attention to word level information, and demonstrated that they do not discriminate between sound contrasts that occur in minimal pairs (as suggested by our cue), and, conversely, discriminate contrasts that occur in non-minimal pairs.

3.1.2 Semantic cue

The semantic cue is based on the intuition that true minimal pairs ([kanaχ] and [kanal]) are associated with different events, whereas alternants of the same word ([kanaχ] and [kanal]) are expected to co-occur with similar events.

We operationalize the semantic cue associated with a pair of phones $Sem(x, y)$ as the average semantic similarity between all the lexical minimal pairs generated by this pair of phones. The higher the average semantic similarity, the more the learner is prone to classify them as allophonic. We take as a measure of the semantic similarity, the cosine of the angle between word vectors of the pairs that vary on the final segment $cos(\widehat{Ax, Ay})$ or the first segment $cos(\widehat{xA, yA})$.

This strategy is similar in principle to the phenomenon of *acquired distinctiveness*, according to which, pairing two target stimuli with distinct events enhances their perceptual differentiation, and *acquired equivalence*, whereby pairing two target stimuli with the same event, impairs their subsequent differentiation (Lawrence, 1949). In the same vein, Yeung and Werker (2009) tested 9 month-olds english learning infants in a task that consists in discriminating two non-native phonetic categories. They found that infants succeeded only when the categories co-occurred with two distinct visual cues.

	Segmentation						Lexicon					
	English			Japanese			English			Japanese		
Allo./phon.	F	P	R	F	P	R	F	P	R	F	P	R
2	0.61	0.57	0.65	0.45	0.44	0.47	0.29	0.42	0.22	0.23	0.54	0.15
4	0.52	0.46	0.59	0.38	0.34	0.43	0.22	0.37	0.15	0.16	0.50	0.10
10	0.51	0.45	0.59	0.34	0.30	0.38	0.21	0.34	0.16	0.16	0.41	0.10
20	0.42	0.38	0.47	0.28	0.26	0.32	0.21	0.29	0.17	0.16	0.32	0.10

Table 1 : Scores of the segmentation and the resulting lexicon, as a function of the average number of allophones per phoneme. P=Precison, R=Recall and F=F-score.

3.1.3 Combined cue

Finally, we consider the combination of both cues in one single cue where the contextual information (semantics) is used as a weighing scheme of the lexical information, as follows:

$$Comb(x,y) = \sum_{(Ax,Ay)\in L^2} cos(\widehat{Ax,Ay}) + \sum_{(xA,yA)\in L^2} cos(\widehat{xA,yA}) \quad (1)$$

where $\{Ax \in L\}$ is the set of words in the lexicon L that end in the phone x, and $\{(Ax, Ay) \in L^2\}$ is the set of phonological minimal pairs in $L \times L$ that vary on the final segment.

The lexical cue is incremented by one, for every minimal pair. The combined cue is, instead, incremented by one, times the cosine of the angle between the word vectors of this pair. When the words have similar distributions, the angle goes to zero and the cosine goes to 1, and when the words have orthogonal distributions, the angle goes to 90° and the cosine goes to 0.

The semantic information here would basically enable us to avoid false alarms generated by potential true minimal pairs like the above-mentioned example of ([kanaχ] and [kanal]). Such a pair will probably score high as far as the lexical cue is concerned, but it will score low on the semantic level. Thus, by taking the combination, the model will be less prone to mistakenly classify ([χ], [l]) as allophones.

3.2 Task

For each corpus we list all possible pairs of allophones. Some of these pairs are allophones of the same phoneme (allophonic pair) and others are allophones of different phonemes (non-allophonic pairs). The task is a same-different classification, whereby each of these pairs is given a score from the cue that is being tested. A good cue gives higher scores to allophonic pairs.

Only pairs of phones that generate at least one lexical minimal pair are considered. Phonetic variation that does not cause lexical variation is "invisible" to top down strategies, and is, therefore, more probably clustered through purely bottom up strategies (Fourtassi et al., 2014b)

3.3 Evaluation

We use the same evaluation procedure as Martin et al. (2013). This is carried out by computing the associated ROC curve (varying the z-score threshold and computing the resulting proportions of misses and false alarms). We then derive the Area Under the Curve (AUC), which also corresponds to the probability that given two pairs of phones, one allophonic, one not, they are correctly classified on the basis of the score. A value of 0.5 represents chance and a value of 1 represents perfect performance.

In order to lessen the potential influence of the structure of the corpus (mainly the order of the utterances) on the results, we use a statistical resampling scheme. The corpus is divided into small blocks of 20 utterances each (the semantic window). In each run, we draw randomly with replacement from this set of blocks a sample of the same size as the original corpus. This sample is then used to retrain the acoustic models and generate a phonetic inventory that we used to retranscribe the corpus and re-compute the cues. We report scores averaged over 5 such runs.

4 Results and discussion

4.1 Segmentation

We first explore how phonetic variation influences the quality of the segmentation and the resulting lexicon. For the evaluation, we use the same measures as Brent (1999) and Goldwater et al. (2009), namely Segmentation Precision (P), Recall (R) and F-score (F). Segmentation precision is defined

as the number of correct word tokens found, out of all tokens posited. Recall is the number of correct word tokens found, out of all tokens in the ideal segmentation. The F-score is defined as the harmonic mean of Precision and Recall:

$$F = \frac{2 * P * R}{P + R}$$

We define similar measures for word types (lexicon). Table 1 shows the scores as a function of the number of allophones per phonemes. For both corpora, the segmentation performance decreases as we increase the number of allophones. As for the lexicon, the recall scores show that only 15 to 22% of the 'words' found by the algorithm in the English corpus are real words; in Japanese, this number is even lower (between 10 and 15%). This pattern can be attributed in part to the fact that increasing the number of allophones increases the number of word forms, which occur therefore with less frequency, making the statistical learning harder. Table 2 shows the average number of word forms per word as a function of the average number of allophones per phoneme, in the case of ideal segmentation.

Allo./Phon.	W. forms/Word	
	English	Japanese
2	1.56	1.20
4	2.03	1.64
10	2.69	2.11
20	3.47	2.83

Table 2 : Average number of word-forms per word as a function of the average number of allophones per phoneme.

Another effect seen in Table 1 is the lower overall performance of Japanese compared to English. This difference was shown by Fourtassi et al. (2013) to be linked to the intrinsic segmentation ambiguity of Japanese, caused by the fact that Japanese words contain more syllables compared to English.

4.2 Allophonic vs phonemic status of sound contrasts

Here we test the performance of the cues described above, in discriminating between allophonic contrasts from phonemic ones. We vary the number of allophones per phoneme, on the one hand (Figure 2a), and the amount of data available to the learner, on the other hand, in the case of two allophones per phonemes (Figure 2b). In both situations, we compare the case wherein the lexical and semantic cues are computed on the output of the unsupervised segmentation (right), to the control case where these cues are computed on the ideally segmented speech (left).

We see that the overall accuracy of the cues is quite high, even in the case of bad word segmentation and very small amount of data.

The lexical cue is robust to extreme variation and to the scarcity of data. Indeed, it does not seem to vary monotonically neither with the number of allophones, nor with the size of the corpus. The associated f-score generally remains above the value of 0.7 (chance level is 0.5). The semantics, on the other hand, gets better as the variability decreases and as the amount of data increases. This is a natural consequence of the fact that the semantic structure is more accurate with more data and with word forms consistent enough to sustain a reasonable co-occurrence statistics.

The comparison with the ideal segmentation, shows, interestingly, that the semantics is more robust to segmentation errors than the lexical cue. In fact, while the lexical strategy performs, overall, better than the semantics under the ideal segmentation, the patterns reverses as we move to a a more realistic (unsupervised) segmentation.

These results suggest that both lexical and semantic strategies can be crucial to learning the phonemic status of phonetic categories since they provide non-redundant information. This finding is summarized by the combined cue which resists to both variation and segmentation errors, overall, better than each of the cues taken alone.

From a developmental point of view, this shows that infants can, in principle, benefit from higher level linguistic structures to refine their phonetic categories, even if these structures are rudimentary. Previous studies about top down strategies have mainly emphasized the role of word forms; the results of this work show that the semantics can be at least as useful. Note that the notion of semantics used here is weaker than the classic notion of referential semantics as in a word-concept matching. The latter might, indeed, not be fully operative at the early stages of the child development, since it requires some advanced conceptual abilities (like forming symbolic representations and understanding a speaker's referential

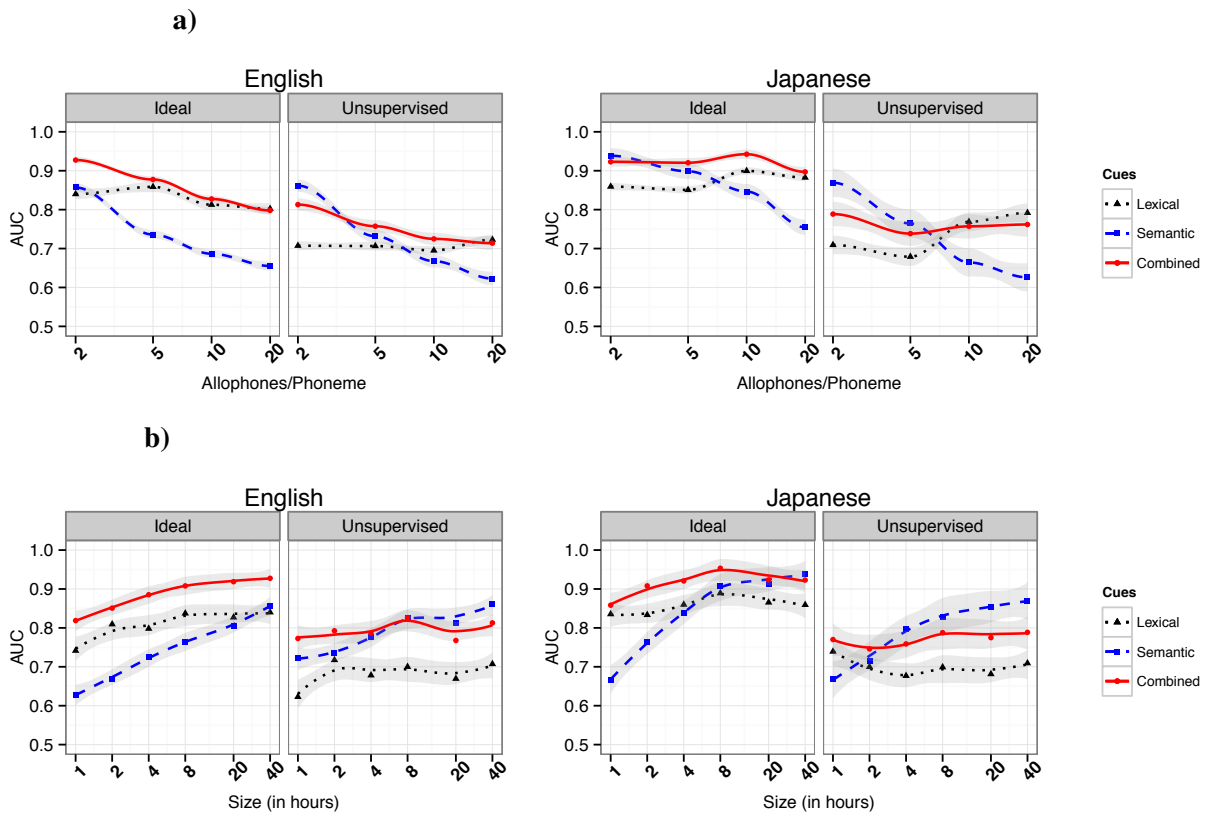

Figure 2: Same-different scores (AUC) for different cues as a function of the average number of allophones per phoneme (a), and as a function of the size of the corpus, in the case of two allophones per phonemes (b). The scores are shown for both ideal and unsupervised word segmentation in English and Japanese. The points show the mean scores over 5 runs. The lines are smoothed interpolations (local regressions) through the means. The grey band shows a 95% confidence interval.

intentions) (Waxman and Gelman, 2009). What we call the "semantics" of a word in this study, is the general context provided by the co-occurrence with other words. Infants have been shown to have a powerful mechanism for tracking co-occurrence relationships both in the speech and the visual domain (Lany and Saffran, 2013) . Our experiments demonstrate that a similar mechanism could be enough to develop a sense of semantic similarity that can successfully be used to refine phonetic categories.

5 General discussion and future work

Phonemes are abstract categories that form the basis for words in the lexicon. There is a traditional view that they should be defined by their ability to contrast word meanings (Trubetzkoy, 1939). Their full acquisition, therefore, requires lexical and semantic top-down information. However, since the quality of the semantic representations depends on the quality of the phonemic representations that are used to build the lexicon, we face a chicken-and-egg problem. In this paper, we proposed a way to break the circularity by building approximate representation at all the levels.

The infants' initial attunement to language-specific categories was represented in a way that mirrors the linguistic and statistical properties of the speech closely. We showed that this detailed (proto-phonemic) inventory enabled word segmentation from continuous transcribed speech, but, as expected, resulted in a low quality lexicon. The poorly segmented corpus was then used to derive a semantic similarity matrix between pairs of words, based on their co-occurrence statistics. The results showed that information from the derived lexicon and semantics, albeit very rudimentary, help discriminate between allophonic and phonemic contrasts, with a high degree of accuracy. Thus, this works strongly support the claim that the lexicon and semantics play a role in the refinement of the phonemic inventory (Feldman et

al., 2013a; Frank et al., 2014), and, interestingly, that this role remains functional under more realistic assumptions (unsupervised word segmentation, and bottom-up inferred semantics). We also found that lexical and semantic information were not redundant and could be usefully combined, the former being more resistant to the scarcity of data and variation, and the latter being more resistant to segmentation errors.

That being said, this work relies on the assumption that infants start with initial perceptual categories (allophones), but we did not show how such categories could be constructed from raw speech. More work is needed to explore the robustness of the model when these units are learned in an unsupervised fashion (Lee and Glass, 2012; Huijbregts et al., 2011; Jansen and Church, 2011; Varadarajan et al., 2008).

This work could be seen as a proof of principle for an iterative learning algorithm, whereby phonemes emerge from the interaction of low level perceptual categories, word forms, and the semantics (see Werker and Curtin (2005) for a similar theoretical proposition). The algorithm has yet to be implemented, but it has to address at least two major issues: First, the fact that some sound pairs are not captured by top down cues because they do not surface as minimal word forms. For instance, in English, /h/ and /ŋ/ occur in different syllable positions and therefore, cannot appear in any minimal pair. Second, even if we have enough information about how phonetic categories are organized in the perceptual space, we still need to know how many categories are relevant in a particular language (i.e., where to stop the categorization process).

For the first problem, Fourtassi et al. (2014b) showed that the gap could, in principle, be filled by bottom-up information (like acoustic similarity). As for the second problem, a possible direction could be found in the notion of *Self-Consistency*. In fact, (Fourtassi et al., 2014a) proposed that an optimal level of clustering is also a level that globally optimizes the predictive power of the lexicon. Too detailed allophones result in too many synonyms. Too broad classes result in too many homophones. Somewhere in the middle, the optimal number of phonemes optimizes how lexical items predict each other. Future work will address these issues in more detail in order to propose a complete phoneme learning algorithm.

Acknowledgments

This work was supported in part by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the Ecole de Neurosciences de Paris, and the Région Ile de France (DIM cerveau et pensée).

References

Elika Bergelson and Daniel Swingley. 2012. At 6 to 9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9).

Luc Boruta, Sharon Peperkamp, Benoît Crabbé, and Emmanuel Dupoux. 2011. Testing the robustness of online word segmentation: Effects of linguistic diversity and phonetic variation. In *Proceedings of CMCL*, pages 1–9. Association for Computational Linguistics.

Luc Boruta. 2011. Combining Indicators of Allophony. In *Proceedings ACL-SRW*, pages 88–93.

Luc Boruta. 2012. *Indicateurs d'allophonie et de phonémicité*. Doctoral dissertation, Université Paris-Diderot - Paris VII.

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

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

N. Feldman, B. Myers, K. White, T. Griffiths, and J. Morgan. 2013b. Word-level information influences phonetic learning in adults and infants. *Cognition*, 127:427–438.

Abdellah Fourtassi and Emmanuel Dupoux. 2013. A corpus-based evaluation method for distributional semantic models. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 165–171, Sofia, Bulgaria. Association for Computational Linguistics.

Abdellah Fourtassi, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. 2013. WhyisEnglishsoeasytosegment? In *Proceedings of CMCL*, pages 1–10. Association for Computational Linguistics.

Abdellah Fourtassi, Ewan Dunbar, and Emmanuel Dupoux. 2014a. Self-consistency as an inductive bias in early language acquisition. In *Proceedings of the 36th annual meeting of the Cognitive Science Society*.

Abdellah Fourtassi, Thomas Schatz, Balakrishnan Varadarajan, and Emmanuel Dupoux. 2014b. Exploring the Relative Role of Bottom-up and Top-down Information in Phoneme Learning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Stella Frank, Naomi Feldman, and Sharon Goldwater. 2014. Weak semantic context helps phonetic learning in a model of infant language acquisition. In *Proceedings of the 52nd Annual Meeting of the Association of Computational Linguistics*.

Judit Gervain and Jacques Mehler. 2010. Speech perception and language acquisition in the first year of life. *Annual Review of Psychology*, 61:191–218.

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

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Valerie Hazan and Sarah Barrett. 2000. The development of phonemic categorization in children aged 6 to12. *Journal of Phonetics*, 28:377–396.

James Hillenbrand, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler. 1995. Acoustic characteristics of american english vowels. *Journal of the Acoustical Society of America*, 97:3099–3109.

M. Huijbregts, M. McLaren, and D. van Leeuwen. 2011. Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection. In *Proceedings of ICASSP*, pages 4436–4439.

A. Jansen and K. Church. 2011. Towards unsupervised training of speaker independent acoustic models. In *Proceedings of INTERSPEECH*, pages 1693–1696.

Aren Jansen, Emmanuel Dupoux, Sharon Goldwater, Mark Johnson, Sanjeev Khudanpur, Kenneth Church, Naomi Feldman, Hynek Hermansky, Florian Metze, Richard Rose, Mike Seltzer, Pascal Clark, Ian McGraw, Balakrishnan Varadarajan, Erin Bennett, Benjamin Borschinger, Justin Chiu, Ewan Dunbar, Abdallah Fourtassi, David Harwath, Chia ying Lee, Keith Levin, Atta Norouzian, Vijay Peddinti, Rachel Richardson, Thomas Schatz, and Samuel Thomas. 2013. A summary of the 2012 jhu clsp workshop on zero resource speech technologies and models of early language acquisition. In *Proceedings of ICASSP*.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 641–648. MIT Press, Cambridge, MA.

Peter W Jusczyk and Richard N Aslin. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology*, 29(1):1–23.

Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

J. Lany and J. Saffran. 2013. Statistical learning mechanisms in infancy. In J. Rubenstein and P. Rakic, editors, *Comprehensive Developmental Neuroscience: Neural Circuit Development and Function in the Brain*, volume 3, pages 231–248. Elsevier, Amsterdam.

D.H. Lawrence. 1949. Acquired distinctiveness of cues: I. transfer between discriminations on the basis of familiarity with the stimulus. *Journal of Experimental Psychology*, 39(6):770–784.

C. Lee and J. Glass. 2012. A nonparametric bayesian approach to acoustic model discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 40–49.

Bogdan Ludusan, Maarten Versteegh, Aren Jansen, Guillaume Gravier, Xuan-Nga Cao, Mark Johnson, and Emmanuel Dupoux. 2014. Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems. In *Proceedings of LREC*.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of japanese. In *LREC*, pages 947–952, Athens, Greece.

D.R. Mandel, P.W. Jusczyk, and D.B. Pisoni. 1995. Infants' recognition of the sound patterns of their own names. *Psychological Science*, 6(5):314–317.

Andrew Martin, Sharon Peperkamp, and Emmanuel Dupoux. 2013. Learning phonemes with a proto-lexicon. *Cognitive Science*, 37(1):103–124.

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

C. Ngon, A. Martin, E. Dupoux, D. Cabrol, M. Duthat, and S. Peperkamp. 2013. (non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science*, 16(1):24–34.

S. Nittrouer. 1996. Discriminability and perceptual weighting of some acoustic cues to speech perception by 3-year-olds. *Journal of Speech and Hearing Research*, 39:278–297.

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

J. Pitman and M. Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.

M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and Fosler-Lussier. 2007. Buckeye corpus of conversational speech.

M Shukla, K White, and R Aslin. 2011. Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences*, 108(15):6038–6043.

N. S. Trubetzkoy. 1939. *Grundzüge der Phonologie (Principles of phonology)*. Vandenhoeck & Ruprecht, Göttingen, Germany.

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

Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux. 2008. Unsupervised learning of acoustic sub-word units. In *Proceedings of ACL-08: HLT, Short Papers*, pages 165–168. Association for Computational Linguistics.

Sandra R. Waxman and Susan A. Gelman. 2009. Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, 13(6):258–263.

J. F. Werker and S. Curtin. 2005. PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, 1(2):197–234.

Janet F. Werker and Richard C. Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49 – 63.

H Yeung and J Werker. 2009. Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, 113:234–243.

Steve J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. 2006. *The HTK Book Version 3.4*. Cambridge University Press.