Bottom-Up Learning of Phonemes: A Computational Study

Rozenn Le Calvez (rozenn.le.calvez@ens.fr)
Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS, DEC-ENS, CNRS) & Université de Paris 6
46 rue d'Ulm, 75005 Paris, France

Sharon Peperkamp (sharon.peperkamp@ens.fr)
Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS, DEC-ENS, CNRS) & Université de Paris 8
46 rue d'Ulm, 75005 Paris, France

Emmanuel Dupoux (dupoux@lscp.ehess.fr)
Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS, DEC-ENS, CNRS)
46 rue d'Ulm, 75005 Paris, France

Abstract

We present a computational evaluation of a hypothesis according to which distributional information is sufficient to acquire allophonic rules (and hence phonemes) in a bottom-up fashion. The hypothesis was tested using a measure based on information theory that compares distributions. The test was conducted on several artificial language corpora and on two natural corpora containing transcriptions of speech directed to infants from two typologically distant languages (French and Japanese). The measure was complemented with three filters, one concerning the statistical reliability due to sample size and two concerning the following universal properties of allophonic rules: constituents of an allophonic rule should be phonetically similar, and allophonic rules should be assimilatory in nature.

Acquisition of Allophonic Rules

During their first year of life, infants learn many aspects of the phonology of their native language. At birth, they all discriminate speech segments (atomic units corresponding to consonants and vowels) in a language-universal way. Their perception then becomes attuned to their native language: at 6-8 months (Kuhl et al. 1992), infants learn the vowel categories of their native language and at 10-12 months the consonant categories (Werker & Tees 1984). These remarkable steps are reached before infants have a lexicon and before they can talk. Infants' learning mechanisms include extracting statistical regularities present in the speech signal such as frequency distributions of segments and transitional probabilities between segments (Jusczyk 1997; Maye, Werker & Gerken 2002). One aspect of early phonological acquisition that remains to be studied is how children go beyond the segmental representation and acquire the phonemes of their language.

Phonemes and Allophonic Rules

Languages represent speech sounds at two levels. At the abstract (underlying) level, word forms are made up of a combination of a finite set of phonemes. At the surface level, the pronunciation of a word is specified in terms of a larger set of context-dependent segments. For instance in Mexican Spanish, the word /felis/ ("feliz", happy) is pronounced as [feliz] when it is followed by a voiced consonant and [felis] otherwise.

Allophonic rules express the phonetic realizations of a given phoneme according to its context. For instance, the allophonic rule of voicing in Mexican Spanish is written as follows:

$$/\mathrm{s}/ \rightarrow \begin{cases} [\mathrm{z}] & \text{before voiced consonants} \\ [\mathrm{s}] & \text{elsewhere} \end{cases} \tag{1}$$

[z] is called the allophone and [s] the default segment. These two segments never occur in the same contexts: they are in "complementary distributions".

Whether a given pair of segments is in an allophonic relationship or not is language-specific: unlike in Mexican Spanish, /s/ and /z/ are two distinct phonemes in French so that /felis/ and /feliz/ could be two different French words. Therefore, phonemes (and allophonic rules) must be learned at some point in the course of language acquisition.

A Bottom-Up Hypothesis

When and how phonemes are learned is controversial. Phonemes might be learned in a top-down fashion with the help of the lexicon or the orthography: knowing the abstract form of a word, children would learn to match phonemes with their phonetic realizations (Kazanina, Phillips & Idsardi 2006). Alternatively, phonemes might be learned very early in life *before* infants have a lexicon, based on complementary distributions of segments (Peperkamp & Dupoux 2002). This is the hypothesis we endorse here.

A Computational Evaluation

We present a computational study of the bottom-up learning of phonemes. Our approach is that a significant number of allophonic pairs can be acquired without the help of lexical information. We suppose that, prior to this acquisition, infants can extract phonetic segments from the speech they hear, use statistical learning mechanisms, and that they have a similarity metrics allowing them to compare the segments of their native language (Liberman & Mattingly 1985). We investigate to what extent statistical information is sufficient or to what extent other information (in the form of linguistic biases) might be necessary.

Related Work

Phonological rule induction models have been studied within various frameworks. Structural linguists have formulated procedures to discover phonemes from a set of language data by hand (Harris 1951 and references therein). Johnson (1984) presented a formal procedure in a generative linguistics approach for the learning of ordered phonological rules. Neither of these approaches, though, were robust to noise.

Gildea & Jurafsky (1996) introduced a stochastic algorithm using finite-state transducers. They included three learning biases often implicit in linguistic theories: faithfulness (surface forms are close to underlying forms), community (similar segments tend to behave similarly) and context (phonological rules are accessed in their context). While in this machine learning approach the algorithm was robust to noise, it had the disadvantage of being supervised by a virtual teacher.

In order to understand how children might learn their language, we develop a statistical algorithm that is unsupervised. It tracks complementary distributions of segments using a measure from information theory. This algorithm was shown to efficiently detect allophonic pairs in a French corpus provided two linguistic filters restricting learning to universally possible allophonic rules were added (Peperkamp et al. 2006). This first study had a number of limitations: the measure was not shown to scale up according to the number of rules and rule complexity; it did not take into account spurious complementary distributions due to small sample size ; it was only tested in one language and the artificial corpora only used very unrealistic languages with equiprobable “phonemes”. In the present study, we add a reliability filter to remove statistically unreliable rules due to small sample size. Tests were performed on a wider range of languages: more realistic artificial languages were used to study the influence of corpus size and number of rules; finally two natural languages (French and Japanese) were studied.

The paper is structured as follows: in the next section, we present the algorithm. We then evaluate it on several artificial language corpora. Finally, the algorithm is tested on natural language corpora of speech transcriptions from two typologically distant languages, namely French and Japanese.

Algorithm

Looking for Complementary Distributions

The algorithm looks for near-complementary distributions of segments using the symmetric Kullback-Leibler divergence (henceforth *KL-measure*) that compares two probability distributions (Kullback & Leibler 1951). Specifically we compared the probability distributions of two segments s_1 and s_2 as follows:

$$m_{KL}(s_1, s_2) = \sum_c \left(P_1 \ \log\left(\frac{P_1}{P_2}\right) + P_2 \ \log\left(\frac{P_2}{P_1}\right)\right) \quad (2)$$

where s_1 and s_2 are two segments,
c are the contexts (right segment, left segment or both) occuring in a corpus,
$P_1 = P(c|s_1)$, $P_2 = P(c|s_2)$,
$P(c|s) = \frac{n(c,s)+1}{n(s)+N}$ with $n(c,s)$ the number of occurences of segment s in context c, $n(s)$ the number of occurences of segment s and N the number of different contexts[1].

The measure is high for segment pairs that have complementary distributions. All segment pairs above a certain threshold (Z-score > 1, corresponding to the mean of measures plus one standard deviation) are selected as candidate allophonic pairs.

Default Phone or Allophone?

A relative entropy criterion then determines the roles of the two segments of the pair, either default segment (that globally appears more often and in more contexts) or allophone. The default segment s_d has the smallest relative entropy:

$$s_d = \arg\min_s \left[\sum_c P(c|s) \log \frac{P(c|s)}{P(c)} \right] \quad (3)$$

where s are the two segments s_1 and s_2 of the phoneme,
and c are the contexts of the segments.

Reliability Filter

The statistical reliability of probability estimations depends on the sample size of corpora. We use a reliability filter to discard unreliable pairs that were selected as candidate allophonic pairs by the KL-measure. The Ψ-criterion (Jaynes 2003) compares observed frequency counts to a theoretical probability distribution. It is similar to a χ^2-test but it is also valid for small samples. It is defined as follows:

$$\Psi(s_1, s_2) = n(s_1) \sum_c f_c(s_1) \ \log\left(\frac{f_c(s_1)}{P(c|s_2)}\right) \quad (4)$$

where $n(s_1)$ is the number of occurences of segment s_1,
$f_c(s) = \frac{n(c,s)}{n(s)}$ with $n(c,s)$ the number of occurences of segment s in context c,
$P(c|s_2)$ is the conditional probability estimation of c given s_2 defined as in Equation 2.

The criterion thus evaluates whether the frequency counts of a segment s_1 are different from the theoretical probability distribution of a segment s_2. If they are

[1] We add one occurence of each segment to the corpus to avoid null probabilities that may arise in limited size corpora.

not considered sufficiently different, the pair of segments is discarded. We use a conservative level of confidence of 10^{-3} (one learner in 1,000 fails): pairs with $\Psi < 3$ are discarded as unreliable.

When we calculate this criterion, we compare the distributions of segments pairwise. To correct for the number of comparisons, we divide the Ψ criterion by the number of comparisons (Bonferroni correction)[2].

Linguistic Filters In Peperkamp et al. (2006), we found that the KL-measure selected spurious allophonic pairs due to phonotactic (i.e. distributional) constraints in natural languages. Two linguistic filters were added to discard them based on linguistic properties of possible allophonic rules. First, allophonic pairs consist of phonetically close segments. In particular, there should not be any intermediate segment between them:

$$\begin{array}{ll} \nexists\ s, & \forall i \in \{1 \ldots 6\}, d_i(s_a) \le d_i(s) \le d_i(s_d) \\ & \text{or } \forall i \in \{1 \ldots 6\}, d_i(s_d) \le d_i(s) \le d_i(s_a) \end{array} \qquad (5)$$

where s is a segment appearing in at least one context of the allophone,
s_d the default segment and s_a the allophone,
$d_i(s)$ is the i^{th} component of the distance representation.

Second, allophonic rules are assimilatory in nature. That is, the allophone should be closer to its contexts than the default segment:

$$\forall i,\ \Big|\sum_{C_{s_a}} \Big(d_i(s_a) - d_i(C_{s_a})\Big)\Big| \le \Big|\sum_{C_{s_a}} \Big(d_i(s_d) - d_i(C_{s_a})\Big)\Big| \qquad (6)$$

where s, s_d, s_a, d_i are defined as above,
C_{s_a} are the contexts of the allophone.

To apply the filters, segments are defined along a numerical articulatory-phonetic distance. The six dimensions used were: *place of articulation* from 1 (bilabial) to 13 (uvular), *sonority* from 1 (voiceless stops) to 12 (low vowels), *voicing* (0 or 1), *nasality* (0 or 1), *rounding* (0 or 1) and *length* (0 for simple segments, 1 for geminates and long vowels).

Linguistic filters were not used for tests on artificial language corpora in which segments are arbitrary symbols rather than segments with phonetic properties. The complete algorithm is summarized in Figure 1.

Simulations with Artificial Languages

Two series of simulations were performed to evaluate the performance of the algorithm. We used artificial languages in order to examine the efficiency of the reliability filter and precisely characterise the sensitivity of the algorithm. We studied the influence of two parameters that are important for our problem: corpus size and the number of allophonic rules.

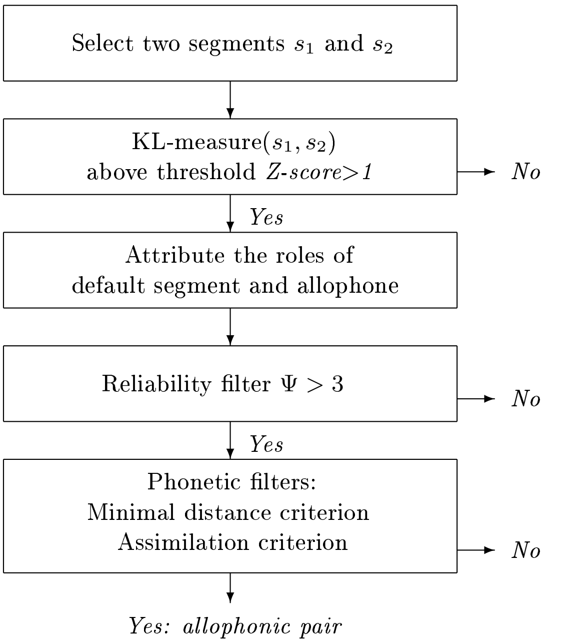
Select two segments s_1 and s_2
KL-measure(s_1, s_2) above threshold *Z-score>1*
No
Yes
Attribute the roles of default segment and allophone
Reliability filter $\Psi > 3$
No
Yes
Phonetic filters: Minimal distance criterion Assimilation criterion
No
Yes: allophonic pair

Figure 1: Summary of the algorithm. Is the current segment pair an allophonic pair?

Corpus Size

Methods We generated artificial language corpora having the following characteristics: the language has 60 segments (similar to the natural corpora used in the next section) with a frequency ratio of 1.000 between the most frequent and least frequent segments and a logarithmic distribution of the frequencies. In this language, ten allophonic rules triggered by randomly determined right contexts were implemented. We chose six corpus sizes varying from 100 to 10^7 segments and drew 20 random corpora of each.

Performance was measured as follows: segment pairs were ranked according to their KL-measure with the pair with the highest KL-measure having rank one. The optimal performance would result in the ten allophonic pairs being ranked from 1 to 10. Hence, the higher the median rank, the worse the performance.

Results Results are shown in Figure 2: box-and-whiskers plots include minimum rank, quartiles and maximum rank.

Median ranks and quartiles decrease with the application of the filter: the reliability filter considerably improves the performance although some outliers remain. The filter is especially efficient on small and middle size corpora. Curves are bell-shaped whether the reliability filter is applied or not (with a maximum around 10^5 segments).

Further analyses (not shown here) revealed that the bell shape is due to two factors: segment frequency and

[2] For instance, for 100 (= 10 × 10) comparisons (about 10 segments in the language), the corrected criterion will be $\Psi = -\log(\frac{10^{-3}}{10^2}) = 5$.

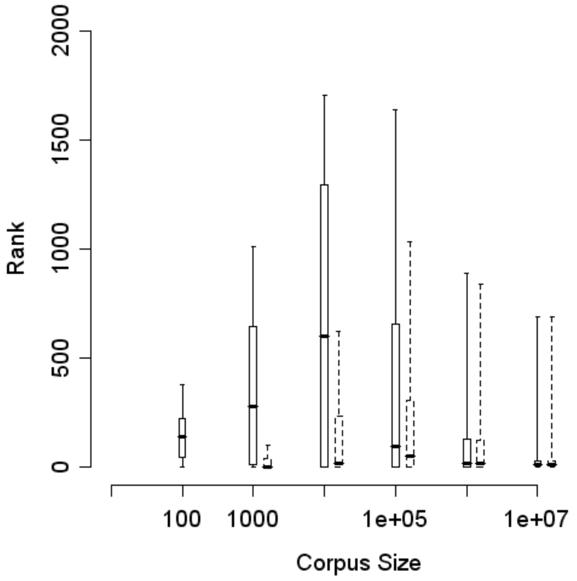

Figure 2: Influence of corpus size with corpora ranging from 100 segments to 10^7 segments. Box-and-whiskers plots show the results of 20 random corpora of each corpus size. Plain (left): Ranks of allophonic pairs before the application of the reliability filter. Dashed (right): Ranks of allophonic pairs after the application of the reliability filter.

interactions among allophonic pairs. Concerning segment frequency, allophonic pairs comprising a rare segment are not present in small corpora; in middle size corpora they are present but mostly unreliably so, leading to a rank increase; in large corpora, rare phonemes are reliably found and ranked well. Concerning interactions, allophonic pairs can for instance share the same allophone. The effect of these interactions is to increase the ranks of allophonic rules in middle size corpora (10^5 segments). Both effects are reduced with the application of the filter.

Number of Rules

Methods The artificial language has 60 segments with a frequency ratio of 1.000 between the most frequent and least frequent segments and a logarithmic distribution of the frequencies. Corpus size was set at a reliable sample size of 10^7 segments, the number of rules varied within a reasonable range for the natural language corpora we use: 1 to 35 rules were triggered by randomly determined right contexts. Twenty corpora were drawn randomly for each number of rules. Simulations were performed with the application of the reliability filter.

Results Results are shown in Figure 3: box-and-whiskers plots include minimum rank, quartiles and maximum rank.

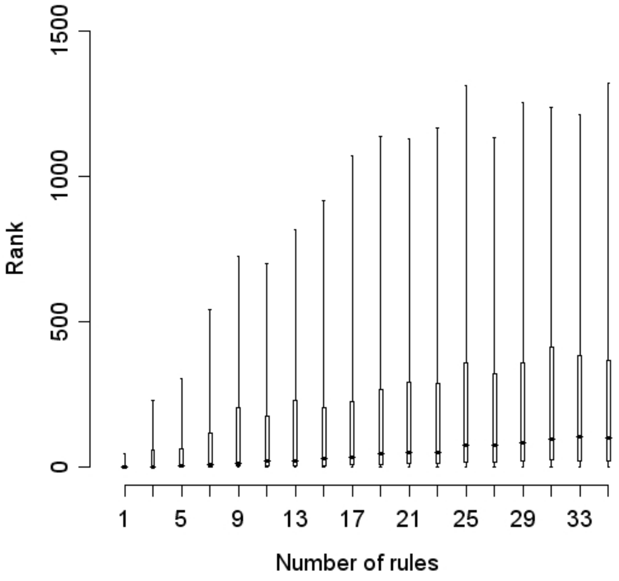

Figure 3: Influence of the number of rules (1 to 35) implemented in a random corpus. Box-and-whiskers plots show the results of 20 random corpora for each number of rules.

Quartiles increase with the number of rules. A few outliers are always badly ranked. The worst rank augments gradually as the number of rules gets bigger, until there are around 25 of them. The median rank is always worse than its optimal value, indicating that some non-allophonic pairs are ranked better than allophonic pairs. These pairs mainly consist of one allophone and another segment (allophone or not) and are thus the result of "allophonic confusion".

Overall, simulations on artificial languages suggest that the algorithm is quite robust to corpus size variation and to variation in the number of rules. The algorithm is particularly sensitive to three characteristics: segment frequency (frequent segments tend to be better ranked), interactions among allophonic pairs in middle size corpora, and allophonic confusion. In natural language corpora, linguistic filters reduce (or remove) the negative effects of these latter two characteristics.

Simulations with Natural Languages

Finally, the algorithm was evaluated on transcribed corpora of child-directed speech in order to examine the performance of the algorithm on two phonologically diverse languages: French and Japanese.

French

The corpus consists of child-directed speech from the CHILDES corpus (MacWhinney 2000). This corpus contains dialogs between parents and children that were orthographically transcribed. Only utterances from adults were kept. We used the VoCoLex dictionary (Dufour et al. submitted) to get a phonemic transcription of the corpus and implemented 11 allophonic rules of French (Dell 1973):

- Sonorants /ʁ,l,m,n,ɲ,ŋ,ɥ,j,w/ are devoiced when followed or preceded by a voiceless consonant /p,t,k,f,s,ʃ/.
- Velars /k,g/ are palatalized when followed by front vowels and semi-vowels /i,y,e,ɛ,ø,œ,j,ɥ,ɛ̃/.

The resulting semi-phonetic corpus included 45 distinct segments of which 11 were allophones. It was 43.000 utterances long (for a total of about 200.000 segments). We ran the algorithm on the corpus. 432 pairs were selected by the KL-measure as candidate allophonic pairs, none of which was discarded by the reliability filter. Among the 432, 8 were correct allophonic pairs. The rest were spurious pairs due to phonotactics (distributional restrictions of phonemes in the language) and allophonic confusion.

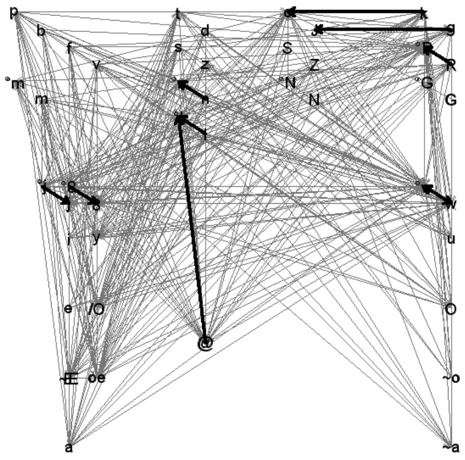

Figure 4: Representation of the results obtained with the CHILDES French corpus. Black lines: pairs kept after the application of the linguistic filters. Gray lines: spurious pairs removed by the filters.

The application of the linguistic filters removed 422 of the 424 spurious pairs. The remaining spurious pairs were [w̥]-[ɥ] (two segments belonging to two different allophonic pairs) and [ə]-[l̥] (due to phonotactic constraints). The action of the linguistic filters is shown in Figure 4 on a 2-dimensional representation of our 6-dimensional distance, roughly showing place of articulation on the horizontal axis and sonority on the vertical axis. Without the filters, all the segment pairs on the figure were selected as candidate allophonic pairs. The filters removed all the gray pairs and kept only the black ones. Notice that most of the spurious pairs (in gray) are distant on the figure.

The three allophonic pairs that were not found by the algorithm were [m]-[m̥], [ŋ]-[ŋ̊] and [ɲ]-[ɲ̊]. Allophones of these pairs were rare in the corpus, hence had a small KL-measure and were not selected.

Japanese

The corpus consists of child-directed speech from the CHILDES corpus of Japanese (MacWhinney 2000). We introduced a number of well-known phonological rules of Japanese (palatalization, affrication, nasal assimilation). The resulting corpus contained 15 allophonic pairs, due to the following allophonic rules:

- /t,d,z/ and their geminates turn into affricates before [u].
- /h/ turns into [f] before [u].
- the moraic nasal /N/ is velarized when followed by velar consonants /k,g/.
- /a,i,u,e,o,aː,iː,uː,eː,oː/ are nasalized when followed by the moraic nasal /N/.

The corpus included 53 distinct segments and was 81.000 utterances long (for a total of about 800.000 segments).

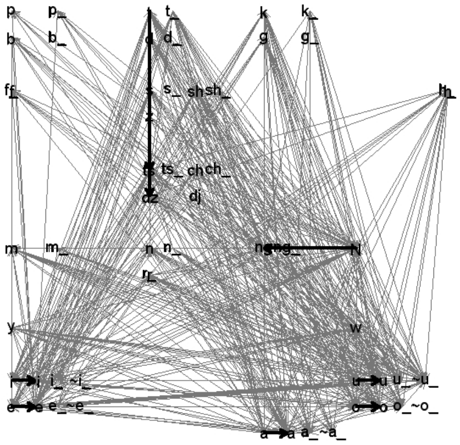

Figure 5: Representation of the results obtained with the CHILDES Japanese corpus. Black lines: pairs kept after the application of the linguistic filters. Gray lines: spurious pairs removed by the filters.

The KL-measure selected 725 candidate allophonic pairs, five of which were removed by the reliability filter. Of the resulting 720 pairs, 8 were allophonic pairs and the remaining were spurious. After the application of the two linguistic filters, only 9 pairs were left: 8 allophonic pairs and 1 spurious pair involving [h] and [ɴ] (due to phonotactic constraints). The action of the filters is represented in Figure 5. As for French, all candidate allophonic pairs selected by the KL-measure are pictured. Pairs discarded by the linguistic filters are pictured in gray, pairs passing the filters in black. The 7 allophonic pairs that were not found (nasalisation of the 5 long vowels, affrication of geminate /t/, [h]-[f]) contained rare allophones.

Discussion

The algorithm with linguistic filters performed very well: it discovered 8/11 and 7/15 of the allophonic pairs in French and Japanese respectively. It should be noted that in both languages there were interactions between rules. For instance in French, several rules applied in the same contexts, leading to complementary distributions between default segments and allophones of different allophonic pairs (such as [m] and [j̥]). These interactions didn't impede the performance of the algorithm, due to the fact that the linguistic filters removed most of these spurious complementary distributions.

Very few spurious pairs were kept: two for French and one for Japanese. They were due to phonotactic constraints and confusion between elements of different allophonic pairs. Adding constraints on the participation of a segment to several allophonic pairs might help to discard them. For instance, we may not allow to keep two allophonic pairs and a third pair consisting of one segment of each of the other two pairs. Such constraints would act on the set of allophonic pairs instead of on individual allophonic pairs. They would thus constrain the phonological system as a whole.

In Japanese, the reliability filter removes several allophonic pairs, thus indicating that the corpus is too small to get reliable information on all pairs. A bigger corpus would be needed to improve the performance.

Conclusion

We presented an algorithm for the bottom-up learning of phoneme categories. Simulations on artificial languages studied the influence of corpus size and number of allophonic rules. The algorithm was applied on data from two languages: French and Japanese. We obtained a good performance provided the algorithm is complemented with three filters: a reliability filter removing unreliable pairs due to insufficient sample size, and two linguistic filters constraining the nature of allophonic rules. The statistical part of our algorithm yielded a very large number of false alarms, most of which are due to phonotactic constraints in these languages. As in Peperkamp et al. (2006), we showed that these false alarms can be pruned down using linguistic filters based on a phonetic representation of the segments that introduce constraints on the universal properties of allophonic rules. Yet, it is unclear as to whether such a phonetic representation is available or not to infants during their first year of life. Further work using an acoustic representation instead of a handmade phonetic one is needed. Another line of research would be to independently acquire phonotactic constraints and use this knowledge to prune spurious allophonic pairs. In brief, in agreement with current language acquisition theories, this study suggests that infants may gain considerable knowledge without a lexicon using a few computational principles and appropriate learning biases.

Acknowledgments

We would like to thank Jean-Pierre Nadal for his help in conceiving the algorithm for the detection of complementary distributions. Research for this work was supported by a graduate student research grant from the French Ministery of Research to R. Le Calvez, as well as by a grant from the Agence Nationale de la Recherche n° ANR-05-BLAN-0065-01.

References

Dell, F. (1973). *Les règles et les sons.* Paris: Hermann.

Dufour, S., Peereman, R., Pallier, C., & Radeau, M. (2002). VoCoLex: Une base de données lexicales sur les similarités phonologiques entre les mots français. *L'Année Psychologique*, *102*, 725–746.

Gildea, D. & Jurafsky, D. (1996). Learning bias and phonological rule induction. *Computational Linguistics*, *22*, 497–530.

Harris, Z. (1951). *Methods in Structural Linguistics.* Chicago: University of Chicago Press.

Jaynes, E.T. (2003). *Probability theory: The logic of science.* Cambridge University Press.

Johnson, M. (1984). A discovery procedure for certain phonological rules. *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics.*

Jusczyk, P. (1997). *The discovery of spoken language.* Cambridge, MA: MIT Press.

Kazanina, N., Phillips C. and W. Idsardi (2006). The influence of meaning on the perception of speech sounds. *PNAS*, *103(30)*, 11381–11386.

Kuhl, P., Williams, K., Lacerda, F., Stevens, K. & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by six months of age. *Science*, *255*, 606–608.

Liberman, A. M. & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82 (3)*, B101–B111.

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

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

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