Available online at www.sciencedirect.com
ScienceDirect
COGNITION
Cognition 101 (2006) B31–B41
www.elsevier.com/locate/COGNIT

Brief article

# The acquisition of allophonic rules: Statistical learning with linguistic constraints

Sharon Peperkamp [a,b,*], Rozenn Le Calvez [a,c], Jean-Pierre Nadal [d], Emmanuel Dupoux [a]

[a] *Laboratoire de Sciences Cognitives et Psycholinguistique, EHESS–ENS–CNRS, 46 Rue d'Ulm, 75005 Paris, France*
[b] *Département des Sciences du Langage, Université de Paris 8, Saint-Denis, France*
[c] *Université de Paris 6, Paris, France*
[d] *Laboratoire de Physique Statistique, ENS–Université de Paris 6–Université de Paris 7–CNRS, Paris, France*

Received 20 May 2005; revised 5 October 2005; accepted 17 October 2005

## Abstract

Phonological rules relate surface phonetic word forms to abstract underlying forms that are stored in the lexicon. Infants must thus acquire these rules in order to infer the abstract representation of words. We implement a statistical learning algorithm for the acquisition of one type of rule, namely allophony, which introduces context-sensitive phonetic variants of phonemes. This algorithm is based on the observation that different realizations of a single phoneme typically do not appear in the same contexts (ideally, they have complementary distributions). In particular, it measures the discrepancies in context probabilities for each pair of phonetic segments. In Experiment 1, we test the algorithm's performances on a pseudo-language and show that it is robust to statistical noise due to sampling and coding errors, and to non-systematic rule application. In Experiment 2, we show that a natural corpus of semiphonetically transcribed child-directed speech in French presents a very large number of near-complementary distributions that do not correspond to existing allophonic rules. These spurious allophonic rules can be eliminated by a linguistically motivated filtering

[*] Corresponding author. Tel. +33 1 44322358.
*E-mail address:* Sharon.Peperkamp@ens.fr (S. Peperkamp).

doi:10.1016/j.cognition.2005.10.006

mechanism based on a phonetic representation of segments. We discuss the role of a priori linguistic knowledge in the statistical learning of phonology.

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

## 1. Introduction

The speed and accuracy with which infants bootstrap into their native language(s) is still a matter of puzzlement. While the very fact that infants can learn language at all has prompted the hypothesis that there exists an innately specified Language Acquisition Device (LAD), it was only recently that research started to investigate what learning algorithm or constraints the putative LAD might contain. Much research has shown that infants have great abilities to process fine-grained phonetic information (see Jusczyk, 1997, for a review) and to extract statistical regularities from the signal. These latter studies provide some hypotheses about possible algorithms for the acquisition of a language's segment inventory (Maye, Werker, & Gerken, 2002) and phonotactic structure (Chambers, Onishi, & Fisher, 2003; Saffran & Thiessen, 2003), as well as the extraction of words from the continuous speech stream (Saffran, Aslin, & Newport, 1996). Yet, there are other aspects of language acquisition that have barely been looked at, such as the acquisition of phonological rules. Phonological rules relate surface phonetic word forms to abstract underlying forms that are stored in the lexicon. A given word, like *kitten*, can be realized differently (e.g. [kʰɪʔən], [kʰɪtn̩], [kʰɪtən]), depending on the phonological context, speech rate, dialect, etc. The distinction between a concrete surface level of representation and an abstract underlying representation is at the very heart of phonological theory, and there is experimental evidence to suggest that this distinction is psychologically real (Lahiri & Marslen-Wilson, 1991). This raises the question how infants acquire the phonological rules of their language and hence infer the abstract representation of words.

We focus on the acquisition of so-called allophonic rules, which introduce phonetic variants of phonemes. For instance, English has an allophonic rule that nasalizes vowels before nasal consonants: the phoneme /æ/ is realized as oral in *mad* [mæd] but as nasal in *man* [mæ̃n]. The oral vowel is called the default segment and the nasal vowel an allophone. Adults have difficulties perceiving contrasts between default segments and allophones (Pegg & Werker, 1997; Peperkamp, Pettinato, & Dupoux, 2003; Whalen, Best, & Irwin, 1997). These difficulties arise in infants at 10–12 months (Pegg & Werker, 1997), suggesting that allophonic rules are acquired around this age. Peperkamp and Dupoux (2002) argued that infants might base their acquisition on the fact that allophones and default segments that are realizations of a single phoneme have non-overlapping distributions. For instance, in English, the allophone [æ̃] only occurs before nasals, whereas the default segment [æ] occurs everywhere else. Hence, computing contextual statistics would allow infants to detect allophonic rules. In this paper, we explore the feasibility of this strategy.

An algorithm that looks for complementary distributions of pairs of segments faces two challenges. First, allophonic rules might be overlooked whenever there is some noise in the signal, for instance because a segment is either misproduced or misperceived. This is why we propose a statistical algorithm that looks for near-complementary distributions. Second, because natural languages have constraints on the sequencing of segments within syllables and words, spurious allophonic rules might emerge (Peperkamp, 2003). For instance, in French, the semivowel [ɥ] only occurs as the last element in the syllable onset (as in *pluie* [plɥi] ‘rain’), hence before vowels, whereas the vowel [œ] only occurs in closed syllables (as in *peur* [pœʁ] ‘fear’), hence before consonants. These two segments, then, have complementary distributions, but in no phonological theory are they considered realizations of a single phoneme. Spurious allophonic rules are especially likely to emerge with a statistical algorithm. Indeed, whereas absolute complementary distributions like those between French [ɥ] and [œ] might be rare, near-complementary distributions can be quite frequent. We can thus foresee that our statistical algorithm needs to be complemented with other sources of information to constrain the set of candidate rules. In particular, we explore the addition of a linguistic filtering mechanism to the statistical algorithm.

In two experiments, we test the feasibility of acquiring allophonic rules on the basis of distributional information. In Experiment 1, we implement a bare statistical algorithm and evaluate its performance using a pseudo-language. In Experiment 2, we turn to natural language data and compare the performances of our statistical algorithm to ones that are equipped with one or two linguistic filters.

## 2. Experiment 1

We implement a statistical algorithm that looks for near-complementary distributions. Using a pseudo-language, we test its performances on corpora of various sizes. We also evaluate its resistance to noise (defined as the presence of segments being misproduced or misperceived) and its capacity to detect optional allophonic rules.

### 2.1. Method

The statistical algorithm calculates which pairs of segments in a given corpus have near-complementary distributions. We use a classical tool from information theory, the Kullback–Leibler measure of dissimilarity between probability distributions (Kullback & Leibler, 1951), defined in Appendix A.[1] For pairs of segments that have different distributions, the Kullback–Leibler number is high, whereas for pairs of segments that have identical distributions it equals 0. A relative threshold in terms of *Z*-score is set, which separates the pairs of segments with a high Kullback–Leibler

[1] Using a different measure, the Bhattacharyya coefficient (Bhattacharyya, 1943), we obtained very similar results in both this experiment and the next one.

number from those with a low Kullback–Leibler number; the former are considered to have allophonic distributions, hence to be realizations of a single phoneme. Within a pair of segments with an allophonic distribution, the allophone is the one that is least frequent and appears in the smallest number of context. A statistically adequate method to single out the allophone is to use a criterion of relative entropy: the allophone is the segment with the highest relative entropy (Appendix B).

### 2.2. Corpora

We created a pseudo-language composed of 46 phonemes with equal relative frequency. One of the phonemes is the target of an allophonic rule that applies in eight contexts. Utterances are composed of random strings of 8–28 segments (mean: 13).[2] Using this pseudo-language, we generated three sets of corpora. The first set consisted of 12 corpora of varying lengths (ranging from 10 to 20,000 utterances), in which no noise is present and in which the allophonic rule applies obligatorily. The second set consisted of 11 corpora based on a sample of 5000 utterances in which the allophonic rule applies obligatorily; these corpora varied in the amount of noise present (ranging from 0% to 100%), where noise is defined as one segment token being replaced randomly by another one. The third set consisted of 11 corpora based on the same sample of 5000 utterances as before, varying this time in the amount of application of the allophonic rule in the appropriate context (ranging from 100% to 0%), in the absence of noise.

### 2.3. Results and discussion

The performances of the statistical algorithm are presented in Fig. 1, showing the $Z$-scores of the Kullback–Leibler measure of all pairs of segments as a function of corpus size, amount of noise, and amount of application of the allophonic rule.

As can be seen, in all corpora containing at least 400 utterances, the pair of segments involved in the allophonic rule is separated from all other pairs of segments in that it has the highest $Z$-score; in these corpora, a threshold can thus be set at which the real allophonic rule is detected without detecting any spurious allophonic rules (Fig. 1A). In the corpora containing 5000 utterances, the pair of segments involved in the real allophonic rule is separated from all other pairs of segments in the presence of up to 40% of noise (Fig. 1B) and starting from 20% of rule application (Fig. 1C).

The present results show that given our pseudo-language, the statistical algorithm distinguishes between real and spurious allophonic rules, except in very small corpora, it is resistant to noise, and it can detect optional allophonic rules.

[2] The mean utterance length corresponds to that in the French child-directed corpus used in Experiment 2.

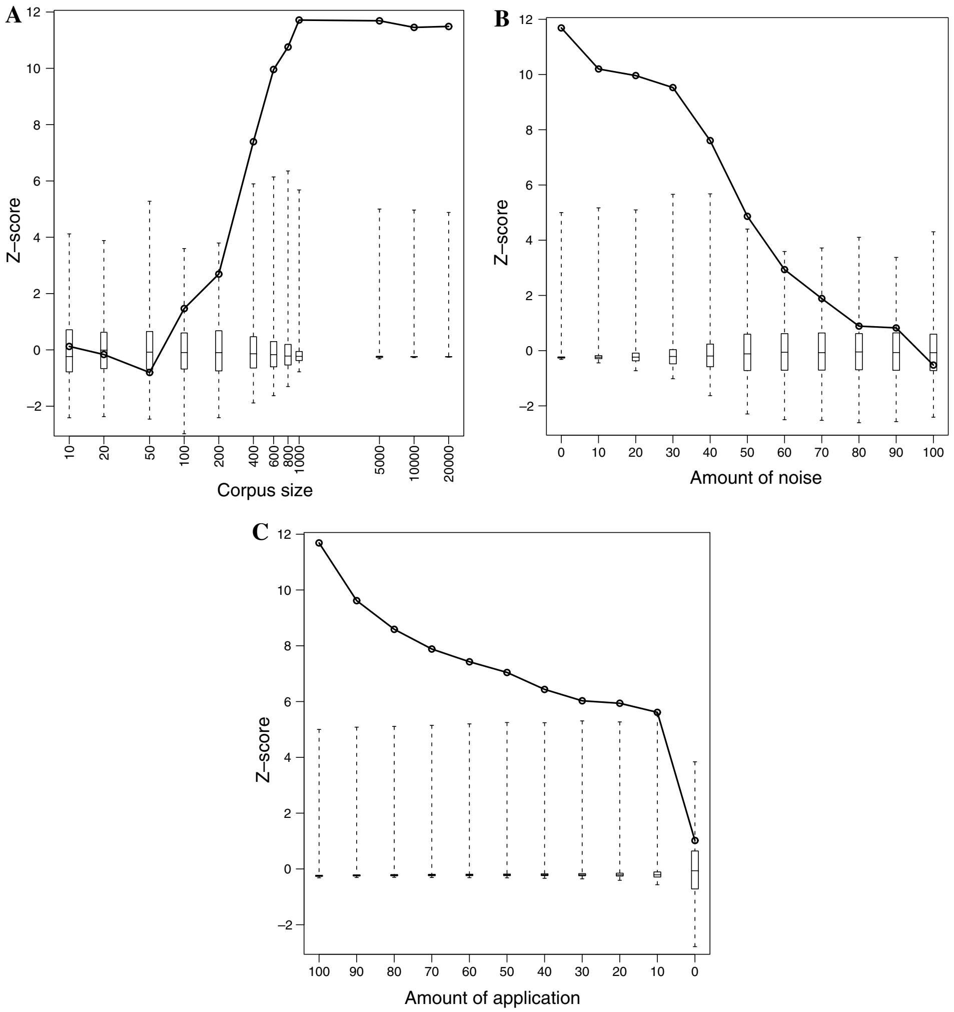

Fig. 1. $Z$-scores of the Kullback–Leibler measure of all pairs of segments as a function of: (A) corpus size, (B) amount of noise, and (C) amount of application of the rule. The scores are represented in box plots. The score of the pair of segments involved in the real allophonic rule is traced by a solid line. The scale of the $x$-axis in (A) is logarithmic; the values of the $x$-axis in (C) are in decreasing order.

## 3. Experiment 2

In this experiment, we test our algorithm on real language data from French. We compare the algorithm of Experiment 1 to ones that are identical to it but have in addition one or two linguistic filters. Cross-linguistically, allophonic rules share two formal properties, which are exploited by these filters. First, the allophone and the default segment are minimally distant from a phonological point of view. Second, the conditioning context spreads some of its features to the target of the rule; hence, the allophone is phonologically closer to the context than the default segment.[3]

### 3.1. Method

We represent each segment as a unique numerical vector encoding its articulatory properties. Phoneticians generally use different sets of properties for the description of vowels and consonants, respectively. In particular, vowels are described by front/backness, height, lip rounding, and nasality, whereas consonants are described by place, manner, and voicing. For the present purposes, we use a single set of properties that provides a unique representation for vowels and consonants alike, i.e., place of articulation, sonority, voicing, nasality, and rounding. Nine different places of articulation are recognized, ranging from bilabial to uvular.[4] Sonority likewise is multi-valued and coded on a scale from 1 (for voiceless stops) to 13 (for low vowels).[5] The remaining properties are binary.

We define two linguistic filters (Appendix C). The first one considers allophonic distributions of two segments as spurious if there is a third segment that is intermediate between them (that is, for each of the components of the vector representation, the third segment lies within the closed interval defined by the other two). The second filter considers allophonic distributions of two segments as spurious if for each of the five articulatory components, the allophone is more distant from its contexts than the default segment.

### 3.2. Corpus

We used a French corpus of child-directed speech, containing some 42,000 short utterances, drawn from the CHILDES database (MacWhinney, 2000). All utterances were first transcribed phonemically, using VoCoLex (Dufour, Peereman, Pallier, & Radeau, 2002). Commas, colons, and semicolons were transcribed as intonational phrase boundaries,[6] full stops as utterance boundaries. Two allophonic rules of French were then implemented: palatalization, which palatalizes /k/ and /g/ before front vowels and semivowels, and devoicing, which devoices /r/, /l/, /m/, /n/, /ɲ/, /ŋ/, and /j/ before voiceless consonants.[7] We took the domains of both rules to be the intonational phrase.

[3] This holds both for rules that are traditionally described as assimilation and for rules that are described as lenition. Dissimilation and fortition, the opposite processes, do not conform to the second principle but they are normally not allophonic. Note also that we do not consider prosodically conditioned rules.

[4] Front vowels are considered to be labial, central vowels dental, and back vowels velar.

[5] Sonority is usually defined acoustically rather than articulatorily. This property has robust articulatory correlates, though: the sonority scale indeed groups together vowels of the same height as well as consonants that share their manner of articulation.

[6] Intonational phrase boundaries can be part of a segment's context. For instance, the right-context of an intonational phrase-final segment is composed of the boundary plus the next segment.

[7] Sonorants also devoice *after* voiceless consonants; we do not consider this rule.

The number of French phonemes being 35, the number of segments in the final corpus was 44 (35 default phones + 2 palatalized allophones + 7 devoiced allophones); yielding 946 (= 44 × 43 × 1/2) segment pairs.

### 3.3. Results and discussion

Let a *hit* be the detection of a real allophonic distribution, a *false alarm* the detection of a spurious allophonic distribution, and a *miss* the non-detection of a real allophonic distribution. Fig. 2 shows the numbers of hits and false alarms as a function of the detection threshold for the statistical algorithm equipped with: (1) no filter; (2) the first filter only; (3) the second filter only; (4) both filters.

Let us first look at the number of hits. Our corpus contained nine pairs of segments with allophonic distributions, due to palatalization ($N = 2$) and devoicing ($N = 7$). With the detection threshold set at any $Z$-score between 0 and 1, there are seven hits. The two misses concern the application of devoicing to /ɲ/ and /ŋ/; these phonemes are the most infrequent ones in the corpus (combined relative frequencies of allophone and default segment: .02% and .01%). With higher thresholds, the number of hits gradually decreases.

Next, we consider the presence of false alarms. In the absence of a linguistic filter, the false alarms always outnumber the hits. With the threshold set at $Z = 1$, the

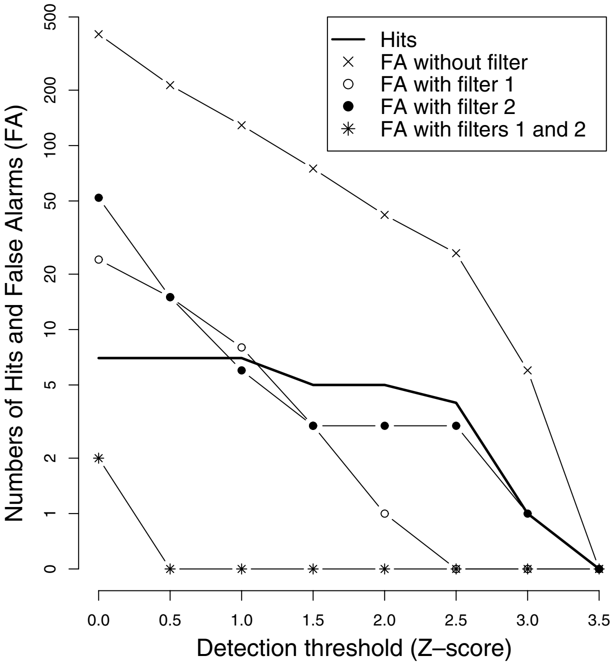

Fig. 2. Numbers of hits and false alarms as a function of the detection threshold for the statistical algorithm with: (1) no filter; (2) the first filter only; (3) the second filter only; (4) both filters. The scale of the $y$-axis is logarithmic.

absolute number of false alarms is 129, corresponding to a false alarm rate of 13.6%.[8] In the presence of one or the other linguistic filter, the number of false alarms is much smaller. In particular, at $Z = 1$ there are 8 and 6 false alarms left after the application of the first and the second filters, respectively. With higher thresholds, the first filter fares best, in that it eliminates all false alarms at $Z = 2.5$. However, it is only in the presence of both filters that all false alarms can be eliminated without losing any hits. This optimal result is indeed reached with a threshold set at $0.5 \leqslant Z \leqslant 1$.

These results show that the statistical algorithm can detect real allophonic distributions in French, except when both the allophone and the default segment are very rare. In order to obtain a 100% hit rate, the algorithm should have a much larger input. Furthermore, in the absence of a linguistic filter, the algorithm also detects a great many spurious allophonic distributions. The two linguistic filters each discard a certain number of these, but only the presence of both linguistic filters allows one to discard spurious allophonic distributions while keeping the real ones.

## 4. Conclusion

Our results show that statistics based on the distribution of segments provide a robust indicator of allophonic distributions in artificial corpora, but are not sufficient to distinguish between real and spurious allophonic distributions in a natural corpus. Hence, natural languages contain a lot of near-complementary distributions over and above those due to allophonic rules. This is probably due to the fact that segments do not occur with equal frequencies in all positions within words and syllables.[9] The addition of linguistic constraints on the form of possible allophonic rules, though, allows one to discard the spurious allophonic distributions. Specifically, we used the conjunction of two constraints, one stating that allophone and default segment should be neighbors in phonetic space, the other that the allophone should be more similar to the context in which it appears than the default segment. Thus, local statistics plus linguistic constraints are sufficient to learn allophonic rules. Preliminary evidence regarding the psychological plausibility of the statistical part of our algorithm is provided by White, Peperkamp, and Morgan (submitted), who showed that prelexical infants can detect allophonic distributions on the basis of distributional information only.

From the point of view of statistical inference it is not surprising that efficient learning may require some prior knowledge. The particular type of linguistic knowledge that we implemented, though, is quite substantial: it consists of an articulatory representation of segments and an assumption about the nature of possible

[8] It should be noted that the large number of false alarms is not due to the presence of real allophonic rules in the corpus. Indeed, if we do not implement palatalization and devoicing we still obtain a false alarm rate of 12.8% at $Z = 1$.

[9] It would be interesting to study languages with different kinds of phonotactic constraints. So far, we obtained similar results using a Japanese corpus.

allophonic rules. Future research could examine the effect of weakening our linguistic filters, as well as the effect of using a representation in terms of acoustic properties rather than articulatory ones. The specific linguistic assumptions regarding allophonic rules on which our filters are based could also be questioned. In fact, our algorithm will fail to discover allophonic rules in which the default segment and the allophone are not neighbors in phonetic space, as well as dissimilatory rules. We are unaware of the existence of such phonetically unnatural allophonic rules. But if they exist, one would have to study the consequences of modifications in the linguistic filters that would be needed to account for their acquisition. In parallel, it should be examined if infants acquire phonetically unnatural rules in the same way as natural ones.

Finally, our approach can be extended to more complicated cases such as non-allophonic rules, which show complementary distributions between word forms rather than between individual segments, and rule interaction (Peperkamp, 2003). Ultimately, the implementation of statistical algorithms might shed light on the acquisition of detailed phonological representations, including phonological features and their hierarchical organization within phonemes.

## Acknowledgements

Research for this article was supported by a master student research grant and a graduate student research grant from the French Ministry of Research to R. Le Calvez, as well as by a grant from the CNRS ('Cognition et Traitement de 1'Information').

## Appendix A. 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:

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

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

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.[10]

[10] In order to smooth the probability estimates of the distributions in finite samples, we add one occurrence of each segment in each context.

## Appendix B. Relative entropy criterion

In a pair of segments $s_{\mathrm{a}}$ and $s_{\mathrm{d}}$ with an allophonic distribution, the allophone $s_{\mathrm{a}}$ is defined as:

$$s_{\mathrm{a}} = \max_{s_{\mathrm{a}},s_{\mathrm{d}}} \left[ \sum_{c} P(c|s) \log \frac{P(c|s)}{P(c)} \right].$$

The other segment, $s_{\mathrm{d}}$, is defined the default segment.

## Appendix C. Linguistic filters

*Filter 1*: Allophonic distributions of segments $s_{\mathrm{a}}$ and $s_{\mathrm{d}}$ are spurious if:

$$\exists s \; [\forall i \in \{1,\ldots,5\}, v_i(s_{\mathrm{a}}) \leqslant v_i(s) \leqslant v_i(s_{\mathrm{d}}) \text{ or } v_i(s_{\mathrm{d}}) \leqslant v_i(s) \leqslant v_i(s_{\mathrm{a}})]$$

with $v_i(s)$ the $i$th component of the vector representation of $s$.

*Filter 2*: Allophonic distributions of segments $s_{\mathrm{a}}$ and $s_{\mathrm{d}}$ are spurious if:

$$\exists i \in \{1,\ldots,5\}, \; \left| \sum_{s \in C[s_{\mathrm{a}}]} (v_i(s_{\mathrm{a}}) - v_i(s)) \right| > \left| \sum_{s \in C[s_{\mathrm{a}}]} (v_i(s_{\mathrm{d}}) - v_i(s)) \right|$$

with $C[s_{\mathrm{a}}]$ the set of contexts of $s_{\mathrm{a}}$, and $v_i(s)$ the $i$th component of the vector representation of $s$.

## References

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society, 35*, 99–109.

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

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.

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

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 76–86.

Lahiri, A., & Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition, 38*, 245–294.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Peperkamp, S. (2003). Phonological acquisition: Recent attainments and new challenges. *Language and Speech, 46*, 87–113.

Peperkamp, S., & Dupoux, E. (2002). Coping with phonological variation in early lexical acquisition. In I. Lasser (Ed.), *The process of language acquisition* (pp. 359–385). Frankfurt: Peter Lang.

Peperkamp, S., Pettinato, M., & Dupoux, E. (2003). Allophonic variation and the acquisition of phoneme categories. In B. Beachley, A. Brown, & F. Conlin (Eds.). *Proceedings of the 27th annual Boston University conference on language development* (Vol. 2, pp. 650–661). Sommerville, MA: Cascadilla Press.

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

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

Whalen, D., Best, C., & Irwin, J. (1997). Lexical effects in the perception and production of American English /p/ allophones. *Journal of Phonetics, 25*, 501–528.

White, K., Peperkamp, S., & Morgan, J. (submitted). Rapid acquisition of phonological alterations by infants.