Developmental Science

Developmental Science 16:1 (2013), pp 24–34 DOI: 10.1111/j.1467-7687.2012.01189.x

PAPER

# (Non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life

Céline Ngon,[1] Andrew Martin,[2] Emmanuel Dupoux,[1] Dominique Cabrol,[3] Michel Dutat[1] and Sharon Peperkamp[1]

1. *Laboratoire de Sciences Cognitives et Psycholinguistique, (Département d'Etudes Cognitives – Ecole Normale Supérieure, Ecole des Hautes Etudes en Sciences Sociales, Centre National de la Recherche Scientifique), Paris, France*
2. *Laboratory for Language Development, RIKEN Brain Science Institute, Saitama, Japan*
3. *Maternité Port-Royal, Université Paris Descartes, France*

## Abstract

*Previous research with artificial language learning paradigms has shown that infants are sensitive to statistical cues to word boundaries (Saffran, Aslin & Newport, 1996) and that they can use these cues to extract word-like units (Saffran, 2001). However, it is unknown whether infants use statistical information to construct a receptive lexicon when acquiring their native language. In order to investigate this issue, we rely on the fact that besides real words a statistical algorithm extracts sound sequences that are highly frequent in infant-directed speech but constitute nonwords. In three experiments, we use a preferential listening paradigm to test French-learning 11-month-old infants' recognition of highly frequent disyllabic sequences from their native language. In Experiments 1 and 2, we use nonword stimuli and find that infants listen longer to high-frequency than to low-frequency sequences. In Experiment 3, we compare high-frequency nonwords to real words in the same frequency range, and find that infants show no preference. Thus, at 11 months, French-learning infants recognize highly frequent sound sequences from their native language and fail to differentiate between words and nonwords among these sequences. These results are evidence that they have used statistical information to extract word candidates from their input and stored them in a 'protolexicon', containing both words and nonwords.*

## Introduction

Infants hear speech as a continuous stream of sounds in which words are strung together with no silences to delimit them. Thus, one of the early tasks infants face on their route towards language acquisition is to find effective ways to extract word-forms from the speech flow. Experimental studies have revealed that they succeed at this task at an early age. For example, Hallé and de Boysson-Bardies (1994) provide evidence that a receptive lexicon emerges in French-learning infants at 11 months: they found that at this age, infants listen longer to lists of words that are frequent in infant-directed speech, such as *ballon* 'ball' and *canard* 'duck', than to lists of rare words, such as *busard* 'harrier'. The same pattern of results has been obtained in English learners of the same age (Vihman, dePaolis, Nakai & Hallé, 2004). Thus, infants of this age have minimally stored the sound forms of some familiar words, while they might not yet know their meaning.[1]

To account for these observations, psycholinguists have focused their attention on the sources of information infants might attend to in order to detect words in the speech stream. Several word-finding procedures have been proposed as being available by the end of the first year of life. For instance, lexical stress constitutes an efficient indicator of word beginnings in some languages, including English, in which most words begin with a stressed syllable (Cutler & Carter, 1987). English-learning infants become sensitive to the predominant stress pattern of their language during the first year of life: At 9 but not at 6 months of age, they prefer listening to disyllabic words with initial stress over ones with final stress (Jusczyk, Cutler & Redanz, 1993). Furthermore, at 7.5 months, they can use this stress cue to detect disyllabic stress-initial words in fluent speech (Jusczyk, Houston & Newsome, 1999; see also Curtin, Mintz & Christiansen, 2005). Other cues that infants can exploit include phonotactic regularities (Mattys & Jusczyk, 2000) and allophonic variation (Jusczyk, Hohne & Bauman, 1999).

Although these cues are often reliable indicators of word boundaries, they are also language-specific, and therefore must be learned before they can be used to extract words. However, in order to learn which cues

[1] Recent work on English-learning infants provides evidence that at an even younger age they can simultaneously segment a word-form and associate it to a visual referent (Shukla, White & Aslin, 2011), and indeed know the meaning of some words of their language (Bergelson & Swingley, 2012).

Address for correspondence: Céline Ngon, Laboratoire de Sciences Cognitives et Psycholinguistique, Pavillon Jardin, 29, rue d'Ulm, 75005 Paris, France; e-mail: celine.ngon@gmail.com

correspond to word boundaries, infants must first be able to identify at least some word boundaries in the input speech. For example, in order to infer that a stressed syllable often signals the beginning of a word, English-learning infants must have learned some word-forms in the first place, allowing them to notice that they tend to have initial stress. One possibility would be that they infer this on the basis of words spoken in isolation. However, while it has been shown that isolated words occur reliably in infant-directed speech (Brent & Siskind, 2001), infants have no way of knowing whether an isolated short sequence of, say, two syllables constitutes a disyllabic word or a sequence of two monosyllabic words. Another possibility would be that they infer language-specific stress cues by focusing on utterance edges, which are marked in the speech signal. It has been shown that infants extract words more readily from utterance edges than from utterance-medial positions (Seidl & Johnson, 2006). However, using this strategy would cause English infants to incorrectly infer that words tend to begin with weak syllables and end with strong syllables; indeed, utterances often begin with weak function words and end with strong monosyllabic content words, respectively. Finally, speech contains a universal, language-general cue to word boundaries in the form of statistical sequential regularities across speech units. Thus, infants could rely on this cue to extract an initial set of word-forms, which could then be used to learn more accurate language-specific cues (Thiessen & Saffran, 2003). This strategy exploits a universal property of human languages, namely the fact that they place restrictions on the sounds that may co-occur within words. Thus, simply computing the probabilities with which phonological units co-occur can be a useful strategy to locate word boundaries.

Evidence for the plausibility of statistical learning as a mechanism to extract words has come from both modelling and experimental studies. As to computational models of word segmentation, several techniques have been implemented, including minimum description length compression (Brent & Cartwright, 1996), mutual information over syllables (Swingley, 2005), and syllable chunking (Perruchet & Vinter, 1998; Batchelder, 2002). These studies demonstrate that exploiting bare statistical language learning mechanisms would lead the infant to be familiar with a fair number of word-forms. For example, Batchelder (2002) implemented a distributional algorithm that could recognize 65% of the words in a corpus of English infant-directed speech, and 56% of the words in a corpus of Japanese speech. Likewise, Swingley (2005) implemented an algorithm that selects syllables or syllable sequences having high frequency and high mutual information as candidate words; in corpora of English and Dutch infant-directed speech, some 80% of these candidate words are real words. Concerning experimental evidence, Goodsitt, Morgan and Kuhl (1993) found that 7-month-old infants are more likely to treat two syllables that co-occur frequently in speech as a potential word than two syllables that rarely co-occur. Using an artificial language learning paradigm, Saffran, Aslin and Newport (1996) further showed that 8-month-old infants can distinguish syllable sequences on the basis of transitional probabilities across syllables only, and Saffran (2001) demonstrated that they group sequences with high transitional probabilities into word-like units.

This previous research has established the plausibility of statistical algorithms by showing that they are effective in computational simulations on natural language data, and by demonstrating that infants possess the ability to utilize sequential regularities in order to extract words from completely novel data. What is still missing, however, is empirical evidence that infants in fact use these abilities to extract words from the speech stream during the course of acquiring their native language, before they have mastered the relevant language-specific cues. Here, we exploit the fact that a statistical learning strategy is necessarily crude and incomplete: Although sequential statistics correlate with word boundaries, the correlation is not perfect, meaning that infants using a purely statistical strategy would be expected to store not only words, but also highly frequent chunks of speech that are not themselves words. We test this prediction of the statistical learning hypothesis, and present evidence that French-learning infants construct a ‘protolexicon’ containing a mixture of real words and nonwords.

Our research enterprise requires that we choose a specific version of the statistical learning hypothesis to test. The segmentation procedures proposed in the studies cited earlier may be broadly divided into two types: those which focus on identifying word boundaries using measures of predictability between small units such as phonemes or syllables, and those which attempt to recognize entire words by storing chunks of speech that recur frequently (see Brent, 1999, for an overview).

We have elected to test a frequency-based measure, in part because this is a simpler statistic to compute than transitional probability or mutual information, and therefore seems an appropriate starting point for an experimental study of statistical learning.

We thus implement an algorithm to extract high-frequency disyllabic sequences from a corpus of French infant-directed speech and test 11-month-old French-learning infants on their recognition of these sequences. At this age, French infants recognize familiar disyllabic word-forms (Hallé & de Boysson-Bardies, 1994), while their word segmentation capacities are still rudimentary (Nazzi, Lakimova, Bertoncini, Frédonie & Alcantara, 2006). If infants use a crude statistical algorithm, they should likewise recognize nonwords that correspond to equally frequent stretches of speech. Using the central visual fixation paradigm (Cooper & Aslin, 1990), we conduct three preferential looking experiments. In the first two experiments we test infants’ sensitivity to the frequency of disyllabic nonword sequences in their speech input, by presenting them with lists of high-frequency disyllabic sequences vs. lists of low-frequency ones. In the last experiment we test whether among high-frequency disyllables infants can distinguish between words and nonwords.

## Experiment 1

In this experiment, we test infants' capacity to recognize high-frequency disyllabic sequences over low-frequency ones.

### *Methods*

### Stimuli

High- and low-frequency disyllabic sequences were extracted using a corpus of French infant-directed speech containing over 285,000 word tokens. The corpus was a subset of several French corpora from the CHILDES database (MacWhinney, 2000), consisting of orthographically transcribed parent–infant speech dialogues recorded in French-speaking families (Suppes, Smith & Leveillé, 1973; MacWhinney, 1995; Bassano & Maillochon, 1994; De Cat & Plunkett, 2002; Hunkeler, 2005; Hamann, Ohayon, Dubé, Frauenfelder, Rizzi, Starke & Zesiger, 2003; Morgenstern, 2006; Demuth & Tremblay, 2008). We only included the parental speech addressed to infants who were at most 24 months old, transcribed word-by-word using the 'Lexique 3' electronic dictionary (New, 2006) and ignoring co-articulatory and prosodic information. We then manually applied the following phonological rules of French: obligatory liaison (the insertion of a word-final consonant before a vowel-initial word in certain words that in other contexts end in a vowel, e.g. *les amis*, 'the friends' [le.za.mi] from [le] and [a.mi]), enchaînement (across-word resyllabification – the final consonant of a word is resyllabified with the initial vowel of the following word, e.g. *balle orange*, 'orange ball' [ba.lo.ʁɑ̃ʒ]), liquid deletion (the deletion of the liquid in word-final obstruent-liquid clusters if followed by a consonant-initial word, e.g. *mètre carré*, 'square meter' [mɛt.ka.ʁe] from [mɛtʁ] and [ka.ʁe]) and schwa insertion (the insertion of schwa after word-final consonant clusters if followed by a word that begins with a consonant cluster, e.g. *porte-clé*, 'key ring' [pɔʁ.tə.kle] from [pɔʁt] and [kle]).

Using this processed corpus, we extracted all disyllabic sequences and ordered them on the basis of their frequency in the corpus. Given infants' sensitivity to utterance boundaries (Hirsh-Pasek, Kemler Nelson, Jusczyk, Wright Cassidy, Druss & Kennedy, 1987), syllable pairs were extracted within but not across utterances. By focusing on disyllabic sequences, we matched the syllable structure of the familiar words used in Hallé and de Boysson-Bardies (1994), to which we will compare nonwords in Experiment 3.

Twelve high-frequency nonwords were selected within the top 1.8% of disyllables ranked by frequency, their number of occurrences in the corpus ranging from 61 to 547 (mean: 174) (see Appendix A). They were phonotactically legal sequences, but none of them were real words or phrases. All but one were consonant-initial.[2] Their internal syllabification was identical to that shown in more than 99% of the utterances in the corpus from which they had been extracted. For instance, the nonword [na.ply] occurred in the corpus almost exclusively in sequences with a syllable boundary before [p], e.g. *tu n'as plus faim* [ty.na.ply.fɛ̃], 'you're not hungry anymore', rather than in sequences with a syllable boundary before [l], e.g. *la nappe lumineuse* [la.nap.ly.mi.nəz], 'the luminous tablecloth'. In addition, no nonword began with a high-frequency function word (specifically, a pronoun, e.g. *tu*, 'you$_{\mathrm{SING}}$', or an article, e.g. *les*, 'the$_{\mathrm{PLU}}$'). Such sequences were excluded because from 6 months of age, French-learning infants have stored high-frequency function words and can use them to segment the following word (e.g. Shi, Cutler, Werker & Cruickshank, 2006; Hallé, Durand & de Boysson-Bardies, 2008; Shi & Lepage, 2008). Finally, in the corpus from which they were extracted, the majority of the selected nonwords appeared in the middle of an utterance most of the time (see Appendix C), and without the final stress pattern that is typical of French words and with which they were recorded for the test; hence, infants would not be able to rely on either distributional cues (nonwords that occur in isolation or at utterance boundaries should be easier to recognize) or a potential stress cue to recognize the selected high-frequency nonwords during the test.

For each of the high-frequency nonwords, a matched low-frequency nonword was constructed by interchanging either the onset consonants of the two syllables, the vowels, or both (see Appendix A). In this fashion, the high- and low-frequency sets contained exactly the same phonemes. All low-frequency nonwords respected the phonotactic constraints of French. Their number of occurrences in the corpus ranged from 0 to 2 (mean < 1; that is, most of them did not occur in the corpus).

High- and low-frequency nonwords were matched pair-wise with regard to syllable structure and did not differ in their mean segment-to-segment co-occurrence frequency ($M_{\mathrm{high}} = 0.007$; $M_{\mathrm{low}} = 0.005$; $t(22) = 1.6$; *ns*), nor in the mean number of embedded words contained within the sequences ($M_{\mathrm{high}} = 7.5$; $M_{\mathrm{low}} = 7.2$; $t < .1$).

All items were recorded in a soundproof room by an adult female native speaker of French, digitized, and stored in computer files. Each nonword was produced in isolation, with a neutral voice and moderate speaking rate. For each of the two stimulus sets, six lists were constructed, each containing a randomization of the 12 nonwords with an ISI of one second. All lists for the same set started with a different nonword. The lists lasted around 20 seconds each.

[2] One vowel-initial item was included in order to perfectly match the syllable structure of the items in the high-frequency list with the familiar words of Hallé and de Boysson-Bardies (1994) that will be used in Experiment 3; one of these familiar words is indeed vowel-initial. Given the absence of vowel-initial nonwords in the corpus that satisfy all the selection criteria, we created it by deleting the initial segment of a consonant-initial nonword.

## Participants

Infants from French-speaking homes in the Paris region were recruited based on parental interest in research participation. French accounted for at least 90% of their language exposure. Parental consent was obtained before the experiment, in accordance with ethical standards for the treatment of human participants. Sixteen 11-month-old infants were tested (eight females and eight males, mean age = 11;3, range = 10;29–11;14).

## Procedure

We used a central visual fixation paradigm (Cooper & Aslin, 1990). Infants were presented with a central auditory test stimulus played binaurally from two speakers on each side of a monitor. The presentation of the audio stimuli was made contingent on the looking time to a central pattern displayed on the monitor. The onset of a trial was triggered when infants had oriented towards an attention-getter. During each trial, infants heard 12 stimuli separated by a 1000 ms interval. The trial ended when the infant looked away from the central pattern for more than 2 seconds, or when the entire list had been played. If the infant looked away but reoriented towards the centre of the monitor within 2 seconds, the list continued, but the time spent looking away was not included in the total looking duration. There were 12 trials, six using lists of high-frequency nonwords and six using lists of low-frequency ones, presented in a pseudo-random order, with the constraint that no more than three trials of the same type could occur in a row.

The experiment was conducted in a dark and quiet room where infants were seated on the caregiver's lap, facing a computer monitor (display size: 47 cm × 30 cm) at a distance of about 75 cm. The monitor and speakers were connected to a computer hidden from the infant behind curtains, and the presentation of the auditory and visual stimuli was controlled by the program Lincoln Infant Lab (Meints & Woodford, 2008). A video camera positioned above the computer monitor recorded the infant's gaze. During the test, the observer pressed a button to start a trial whenever the infant's attention was drawn towards an animated visual stimulus – the attention getter – at the centre of the monitor. On each trial, the auditory stimulus was presented together with an unrelated static visual display at the centre of the monitor, a black-and-red checkerboard pattern on a white background. Throughout the experiment, both the caregiver and the experimenter listened to masking sound through headphones. The experiment lasted 3 minutes on average.

### *Results and discussion*

Infants' looking times were re-coded offline frame by frame (40 ms interval). Mean looking times per trial type are shown in Figure 1.

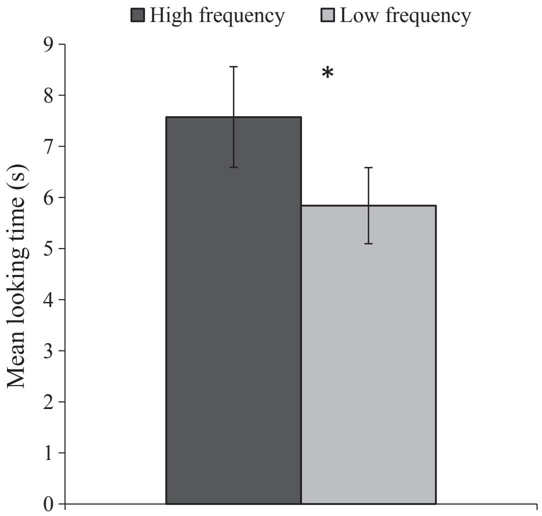

**Figure 1** Mean looking times per trial to high-frequency and low-frequency nonwords. * $p < .01$.

A paired $t$-test revealed that infants listened longer to high-frequency than to low-frequency nonwords ($t(15) = 3.0$; $p < .01$). Thirteen out of 16 infants showed this pattern.[3]

These results show that 11-month-old French-learning infants have extracted high-frequency disyllabic sequences from their speech input and recognize them when they are presented in isolation. Note, however, that even though we had matched the diphone frequency of the items in the two lists, we had not likewise matched the frequency of the syllables that composed the items. The mean number of occurrences of these component syllables turns out to be significantly higher in the high-frequency than in the low-frequency list ($M_{\text{high}} = 4798$; $M_{\text{low}} = 1140$; $t(22) = 3.27$; $p < .003$). Could it be that infants are not really storing disyllables as units, but rather, are reacting to the frequency of the component syllables in our stimuli? In the next experiment, we test this by using stimuli that are matched in component syllable frequency. Thus, if infants in this experiment still distinguish between high- and low-frequency disyllables, their performance must be based on their sensitivity to the frequency of the entire sequence.

[3] These results are similar to those obtained in a pilot experiment in which the materials were not matched as well as in the present experiment: Among the frequent nonwords, one formed a common isolated utterance, and some others started with very frequent adverbs such as *oui*, 'yes' and *non*, 'no'. Moreover, infrequent nonwords were selected among existing low-frequency sequences generated from the corpus, rather than being constructed by swapping a couple of segments in each of the selected high-frequency ones; thus, high- and low-frequency nonwords were not matched in terms of their segmental make-up. As in the present experiment, 11-month-old French infants listened longer to high-frequency nonwords as opposed to low-frequency ones ($M_{\text{high}} = 8.9$s; $M_{\text{low}} = 6.8$s; $t(15) = 3.59$; $p < .003$).

## Experiment 2

### *Methods*

#### Stimuli

The high-frequency disyllabic nonwords were the same as those used in Experiment 1. The low-frequency nonwords were constructed by rearranging syllables from the high-frequency list, while keeping them in the same position (initial or final). Thus, initial and final syllables were recombined with one another so as to get low-frequency sequences (see Appendix A). They were all phonotactically legal, and occurred between 0 and 17 times (mean: 6) in the corpus.

As in Experiment 1, high- and low-frequency nonwords were matched pair-wise with regard to syllable structure and did not differ in their mean segment-to-segment co-occurrence frequency ($M_{\text{high}} = 0.007$; $M_{\text{low}} = 0.007$; $t(22) = 0.24$; *ns*), nor in the mean number of embedded words contained within the sequences ($M_{\text{high}} = 7.5$; $M_{\text{low}} = 7.5$; $t(22) = 0.11$; *ns*). Crucially, the frequency of the syllables contained within the sequences did not differ either, since the two lists were composed of the same syllables. A female native speaker of French, different from the one who recorded the stimuli for Experiment 1, recorded both the high- and the low-frequency sequences.

#### Participants

Sixteen infants (seven females, mean age: 11;2, range 10;14–11;21) participated. None of them had participated in Experiment 1. Four additional infants were tested but excluded from analyses due to crying (1), fussiness (1), parental interference (1) and experimenter error (1).

#### Procedure

The experiment followed the same procedure as described in Experiment 1.

### *Results and discussion*

As in Experiment 1, infants' looking times were re-coded offline. Mean looking times per trial are shown in Figure 2.

A paired *t*-test revealed that infants listened longer to high-frequency than to low-frequency nonwords ($t(15) = 2.63$; $p < .02$). Twelve out of 16 infants showed this pattern. A mixed ANOVA with the factors Experiment (1 vs. 2) and Frequency (High vs. Low) yielded an effect of Frequency only ($F(1, 30) = 15.9$; $p < .0004$). Thus, infants recognized high-frequency nonwords over low-frequency nonwords that are matched in component syllable frequency, and their behaviour was not different from that of infants in Experiment 1. This, then, is evidence that at 11 months of age French-learning infants are sensitive to the frequency of disyllables in their input: they have extracted the high-frequency ones and recognize them when presented in isolation.

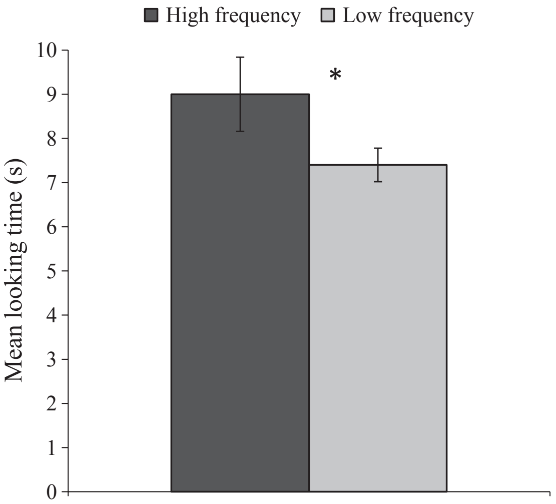

**Figure 2** Mean looking times per trial to high-frequency nonwords and low-frequency nonwords that are matched in single syllable frequency. * $p < .02$.

By themselves, these results do not show that infants treat high-frequency nonwords on a par with words. If their recognition of these disyllables is indeed a consequence of their use of a statistical word-finding procedure, we predict that infants should recognize real words of the same frequency range and, crucially, fail to distinguish between high-frequency words and nonwords. We test this prediction in the next experiment.

## Experiment 3

### *Methods*

#### Stimuli

The high-frequency disyllabic nonwords were the same as those used in Experiment 1. The high-frequency words consisted of the 12 disyllabic words that 11-month-olds were shown to recognize in Hallé & de Boysson-Bardies (1994), except that we replaced *bonjour*, 'hello', and *encore*, 'again', because they often appear as isolated utterances (see Appendix B).[4] They were recorded by the same speaker who recorded the stimuli for Experiment 1.

High-frequency nonwords and words were matched pair-wise with regard to syllable structure. They did not

[4] We overlooked the presence of the word *lapin* in the list, which starts with the frequent function word *la*, 'the$_{\text{FEM}}$'. Note that if this were to influence infants' behaviour, it should favour longer listening times to the words as opposed to the nonwords.

differ in their mean frequency in our infant-directed speech corpus ($M_{\text{high}} = 174$; $M_{\text{words}} = 145$, $t < 1$), their mean segment-to-segment co-occurrence frequency ($M_{\text{high}} = 0.007$; $M_{\text{words}} = 0.004$; $t(22) = 1.38$; *ns*), or the mean number of the embedded words contained within them ($M_{\text{high}} = 7.5$; $M_{\text{words}} = 8.9$; $t(22) = 1.8$; *ns*).

### Participants

Sixteen infants (seven females, mean age: 11;1, range 10;28–11;15) participated. None of them had participated in Experiments 1 or 2. One additional infant was tested but excluded from analyses due to experimenter error.

### Procedure

The experiment followed the same procedure as in the two previous experiments.

### *Results and discussion*

As in Experiment 1 and Experiment 2, infants' looking times were re-coded offline. Mean looking times per trial are shown in Figure 3.

A paired *t*-test revealed that the difference between the mean total looking times for the two lists is not significant ($t(15) < 1$), even though 11 infants out of 16 preferred the real words. These results show that 11-month-old French infants do not discriminate between high-frequency words and high-frequency nonwords.

Together with the results from Experiments 1 and 2, this suggests that early in development, infants extract highly recurring sound sequences and store them in a receptive protolexicon, without distinguishing between words and nonwords. In other terms, the familiar words that 11-month-old infants have been observed to recognize (Hallé & de Boysson-Bardies, 1994) are in fact blended among these extracted high-frequency sound sequences.

## General discussion

Artificial language learning studies have demonstrated that infants can rely on statistical information to segment speech into word-like units (e.g. Saffran *et al*., 1996). The present study examined whether infants in fact use statistical information to extract candidate words from running speech towards the end of their first year of life. We focused on 11-month-old French-learning infants and tested the prediction that they recognize co-occurring syllable pairs that are highly frequent in the input, regardless of whether they are words or nonwords. A simple algorithm that stores high-frequency disyllabic sequences was used to simulate infants' extraction of candidate words from infant-directed speech. In Experiments 1 and 2, 11-month-old French-learning infants recognized the most highly frequent disyllabic nonwords over low-frequency ones. In Experiment 3, they failed to distinguish between those high-frequency nonwords and real words from the same frequency range. Together, these results reveal that at 11 months – when they cannot yet segment words accurately from fluent speech (Nazzi *et al*., 2006) – French-learning infants indeed use statistical information to extract word candidates from their input. This guides them to construct a protolexicon in which real words are not yet distinguishable from nonwords.

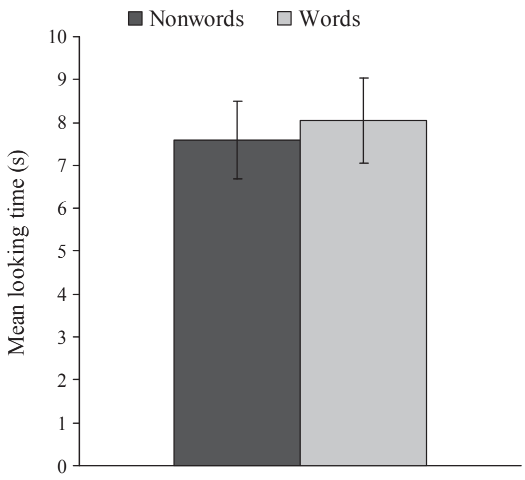

**Figure 3** Mean looking times per trial to high-frequency nonwords and high-frequency words.

The present results provide further insight into the nature of the sound sequences that infants recognize in their first year of life. While earlier work found that French 11-month-olds could recognize frequent words (Hallé & de Boysson-Bardies, 1994), our results show that this is only part of the story. In addition to these actual words, the word-finding procedure used by infants at this age appears to also generate a potentially large number of 'false alarms', i.e. high-frequency sequences that straddle word boundaries but are nonetheless treated as words. This raises the question of the size, composition, and development of such a protolexicon.

Answering this question would ideally involve testing individual items on individual infants, and sampling the stimuli from a large number of frequency ranges. In our experiments, we only have group results for a list of 12 items sampled from a rather large frequency range. This makes it difficult to know whether infants' preference for high-frequency items was driven by the entire set of stimuli, or restricted to the most frequent ones. It is unlikely that the results are due to one or two items. To see why, note that while each list of 12 items lasted about 20 seconds, the average duration of a trial was roughly 7 seconds; in any given trial infants were thus only presented with on average four items in the list. However, this still leaves a substantial range, from a few items to all 12 items, within which the crucial frequency cut-off could fall. In order to calculate the size of the protolexicon, we consider three possibilities, according to which infants' reactions must have been based on the 25, 50, or 75% most frequent items in each list, respectively.

We calculated the mean frequency of the items within the first quartile, the median, and the third quartile of all the high-frequency items from which the 12 nonword sequences used in Experiments 1 and 2 were extracted. The mean frequency of the items within the first quartile is 407.9 (corresponding to the top 0.46% of the distribution of disyllables in the corpus), that of the items within the median is 270 (top 0.92%) and that of the items within the third quartile is 209.3 (top 1.38%). A list of disyllables was compiled for each frequency range, excluding sequences beginning with a high-frequency function word.[5] We have found that the frequency ranges within the first, second and third quartile contain 104, 213 and 323 sequences, of which 55%, 45% and 43% correspond to real words, respectively.[6] These estimates of the size of the infants' protolexicon and its word/nonword ratio is of course rather crude. In addition to the fact that a more sensitive test paradigm is necessary to establish the actual frequency cut-off above which infants recognize frequent sequences, two more factors should be taken into account for a more accurate size estimate. First, we did not consider monosyllabic or trisyllabic word-forms, suggesting that our computation *underestimates* the size of the protolexicon. Second, we did include vowel-initial disyllables in our count; given that our experimental stimuli included only one such sequence, we do not know to what extent infants extract vowel-initial sequences, and hence in this respect our computation might *overestimate* the size of the protolexicon.

The exploration of the size and composition of the protolexicon is also dependent on the postulated segmentation algorithm that infants use. We compared the fit of our frequency-based algorithm to the data with that of three other algorithms proposed in previous research: two based on the transitional probabilities (TP) between syllables (e.g. Saffran *et al.*, 1996; Aslin, Saffran & Newport, 1998), and another based on the mutual information (MI) of syllables (e.g. Swingley, 2005). We implemented the forward TP algorithm of Saffran *et al.* (1996), the backward TP algorithm of Pelucchi, Hay and Saffran (2009), and a simplified version of the MI algorithm of Swingley (2005)[7] on the same corpus that we used for the selection of our stimuli, classifying disyllabic sequences from most to least probable word candidates. Each of our experimental stimuli was then assigned four values: disyllable probability, forward TP, backward TP, and MI.[8] For each of these measures, we calculated the difference in the mean values between the two sets of stimuli used in each experiment. Larger differences predict better discrimination on that pair of stimuli types. These differences are displayed in Figure 4, together with the corresponding differences in looking times from Experiments 1, 2, and 3 for comparison.

The figure shows that both disyllable probability[9] and forward transitional probabilities are consistent with the experimental results. In fact, the latter appears to be more consistent in that it correctly predicts no difference between Experiments 1 and 2. Our results do not, however, rule out a simple frequency-based model, for two reasons. First, a more sensitive experimental paradigm might uncover a difference between the two experiments that our paradigm was unable to detect. Second, the stimuli in Experiments 1 and 2 differ only in the probabilities of the low-frequency items. The disyllable probability model therefore predicts a difference between these two experiments only if infants are sensitive to differences between very low-frequency items and slightly higher (but still very low) frequency items. This would require them to store, and track the frequency of, virtually every disyllable they hear. For these reasons, we take the conservative position that our experimental results provide support for both forward transitional probability and disyllable probability models, and leave it to future research to determine which one is a better model of infants' actual statistical learning abilities.

As for mutual information and backward transitional probability, these measures also predict a difference between Experiments 1 and 2, which, as noted above, does not by itself rule them out as models of the data. More problematic, however, is the fact that MI predicts that discrimination in Experiment 3 should be better than that in Experiment 1 and backward TP predicts that discrimination should be the same in both experiments, whereas the infants in our experiment demonstrated significantly better discrimination in Experiment 1 than Experiment 3. These incorrect predictions cannot be explained away as the result of an insufficiently sensitive experimental paradigm.

[5] As mentioned in Experiment 1, French-learning infants are known to recognize function words and be able to use them to aid segmentation from 6 months of age (e.g. Hallé *et al.*, 2008).

[6] In a non-reported experiment, we tested another group of infants using high-frequency nonwords of a somewhat lower frequency range than those used in Experiment 1 and matched low-frequency nonwords. The high-frequency nonwords were between the top 1.9% and 3.9% of sequences ranked by frequency (31 to 58 occurrences, mean: 43). For infants in this experiment, the difference in looking times for the two lists was not significant ($t(15) < 1$). Thus, infants' recognition of high-frequency nonwords seems to be limited to the highest frequency range. Based on this result, a similar computation leads us to infer the maximum size of the protolexicon as well as its minimum word/nonword ratio: considering the frequent items in this non-reported experiment, 227 disyllables are within the first quartile, 483 within the second quartile and 745 within the third quartile, with a word/nonword ratio of 45%, 41% and 37%, respectively.

[7] In particular, we considered only disyllabic sequences (in French, almost all legal syllables are real words), and included word candidates that are embedded within a real word.

[8] For stimuli consisting of successive syllables $A$ and $B$, disyllable probability is defined as $p(AB)$, forward transitional probability as $p(AB) / p(A)$, backward transitional probability as $p(AB) / p(B)$, and mutual information as $\log_2(p(AB) / p(A)p(B))$.

[9] Although we assume that infants track frequencies over sequences of syllables, the results reported here are largely unchanged if frequencies are instead computed over strings of phonemes.

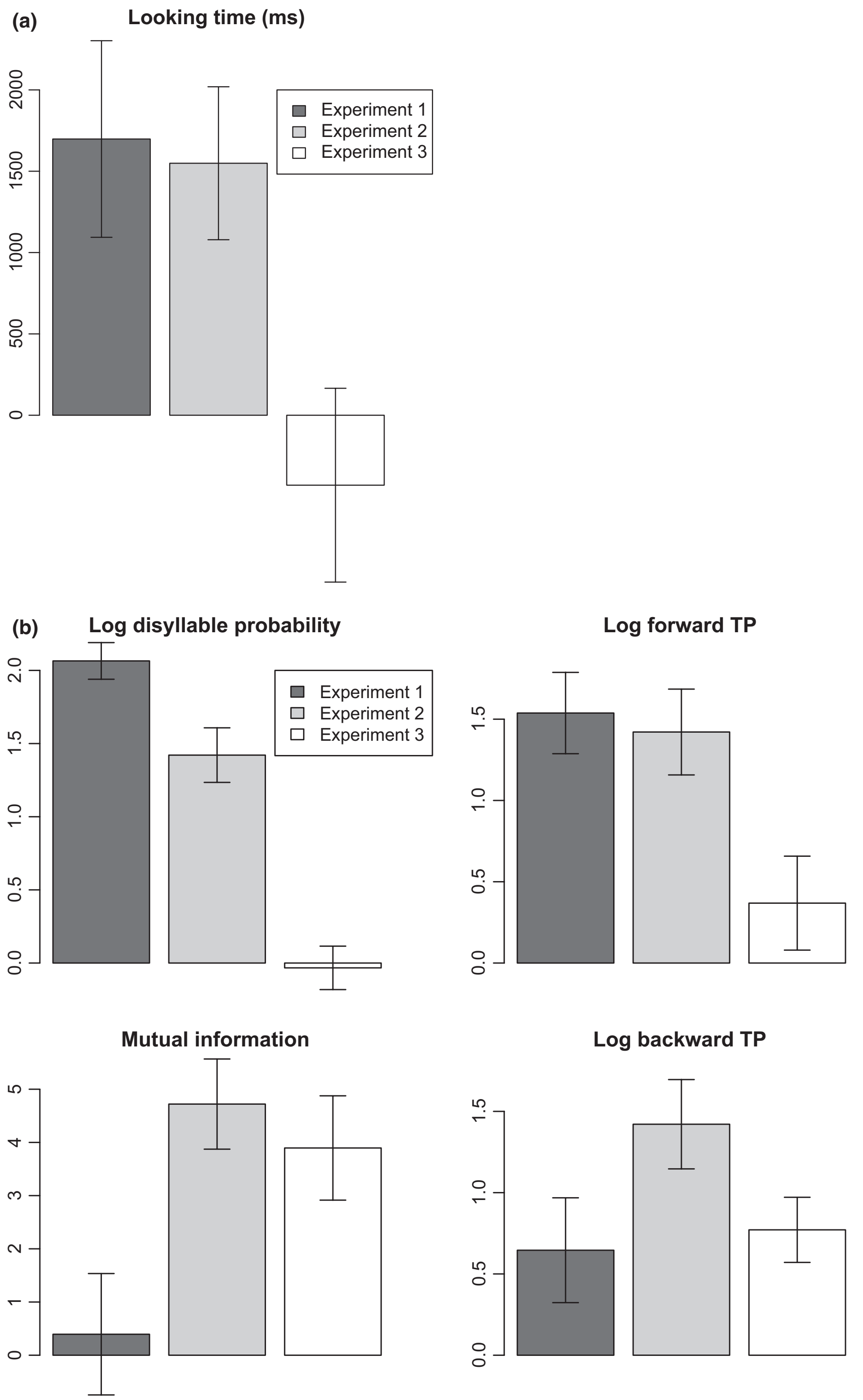

**Figure 4** Comparison of experimental results (a) to four potential models (b). The height of each bar represents the difference between the mean values of high- and low-frequency disyllables (Experiments 1 and 2) or of words and high-frequency disyllables (Experiment 3). Error bars represent the standard error of each difference in means.

Mutual information and backward TP are thus effective measures for distinguishing words from high-frequency nonwords, but are not as good at distinguishing high- from low-frequency nonwords, while actual infants display the opposite trend. The reasons for this can be found in the differing statistical structure of the two sets of stimuli: although disyllable probability ($p(AB)$) and the probability of the first syllable ($p(A)$) are nearly identical in words and nonwords, the probabilities of the second syllable ($p(B)$) differ greatly, with mean $p(B)$ for words (0.001) being a full order of magnitude smaller than that of high-frequency nonwords (0.014). Because MI and backward TP are both a function of $p(B)$, these metrics are particularly good at discriminating between words and high-frequency nonwords.

As for *why* second-syllable probabilities differ so greatly in words and nonwords, this is likely a consequence of a fundamental difference between the two types of sequence. That is, high-frequency nonwords recur often in the input precisely because they consist of independently high-frequency syllables, whereas word frequency is driven by other factors. Despite the potential usefulness of mutual information or backward TP as cues to word discovery, however, our experimental results provide no evidence that infants rely on either of these statistics.

Forward transitional probability and disyllable probability, on the other hand, based as they are on the values of $p(AB)$ and $p(A)$, make roughly the same predictions. Future experiments could manipulate these variables independently, allowing us to determine which of these two statistics provides a better model of infants' word-finding procedure. Although transitional probabilities constitute the algorithm that is the most clearly identified with statistical learning, there are models of word segmentation that do not rely on transitional probabilities (see Introduction; note also Endress & Mehler, 2009, for evidence that adults do not use transitional probabilities in word segmentation).

To conclude, the present results provide evidence for the presence of a protolexicon in French-learning infants of 11 months, an age at which they do not yet master language-specific word segmentation. This protolexicon contains a large number of nonwords, suggesting that infants – at least when it comes to words – go through a stage in which their linguistic knowledge is not just less precise than that of adults but is by and large contradictory to it. That is, rather than gradually building up a lexicon of word-forms in which they have high confidence, they begin by assembling a large set of mostly meaningless phonological forms, which will later be pruned as more information becomes available.

In addition to providing a list of potential word candidates, a protolexicon could reap benefits in the acquisition of phonological categories. The modelling study of Swingley (2005) shows that a protolexicon can help to infer the prototypical stress pattern of English and Dutch, which can in turn help improve word segmentation. In a similar vein, Martin, Peperkamp and Dupoux (in press) show that a protolexicon can provide powerful top-down information that helps to construct phoneme categories out of allophonic variants, giving infants an advantage over a purely bottom-up procedure that tries to learn these categories from attending only to distributional cues. And vice versa, knowledge of word-forms that contain a given phonemic contrast could help infants distinguish that contrast, as shown in experimental work by Thiessen (2011). More generally, learning several linguistic levels simultaneously is not harder but actually easier than learning them separately (Johnson, Demuth, Frank & Jones, 2010). The reason is that even poorly specified linguistic information from one level can help learning another level, thereby creating positive synergies in learning. The existence of a protolexicon of approximate word-forms is thus expected as an intermediate stage of such synergistic learning.

As mentioned above, more research is necessary to test the potential use of other types of statistical learning algorithm that take into account more than just frequency, in particular transitional probabilities. In addition, future experiments could test for the existence of a protolexicon in other languages, for instance in English. Indeed, since English-learning infants are known to be ahead of French learners in their word segmentation capacities, which appear between 6 and 7.5 months (Jusczyk & Aslin, 1995), they might also be ahead in the construction of a protolexicon. Finally, the existence of a protolexicon might have consequences for word learning. Specifically, Graf-Estes, Evans, Alibali and Saffran (2007) showed that infants consider the sequences that they segment out of an artificial language based on statistical information to be potential words: in an object-label task that follows the segmentation task, they succeed when the labels are words in the artificial language but not when they are sequences crossing a word boundary. We expect that until infants develop a real lexicon, they likewise associate meanings more readily to high-frequency sequences in their native language than to low-frequency ones, the former being stored in their protolexicon of candidate words.

## Acknowledgements

This study was funded by grants from the Agence Nationale de la Recherche (ANR-10-BAN-1901-01) and the Fondation de France. We thank the parents and infants who participated in this study. We are grateful to Isabelle Brunet and Sylvie Margules for help in recruiting participants, and to Stéphanie Iannuzzi and Mathilde Fort for stimuli recording.

## References

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

Bassano, D., & Maillochon, I. (1994). Early grammatical and prosodic marking of utterance modality in French: a longitudinal case study. *Journal of Child Language*, **21**, 649–675.

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

Bergelson, E., & Swingley, D. (2012). At 6 to 9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, USA*, **109**, 3253–3258.

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

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

Brent, M.R., & Siskind, J.M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, **81**, B33–B44.

Cooper, R., & Aslin, R. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, **61**, 1584–1595.

Curtin, S., Mintz, T.H., & Christiansen, M.H. (2005). Stress changes the representational landscape: evidence from word segmentation. *Cognition*, **96**, 233–262.

Cutler, A., & Carter, D.M. (1987). The predominance of stressed initial syllables in the English vocabulary. *Computer Speech and Language*, **2**, 133–142.

De Cat, C., & Plunkett, B. (2002). Qu'est ce qu'i (l) dit, celui-là? Notes méthodologiques sur la transcription d'un corpus francophone. In C.D. Pusch & W. Raible (Eds.), *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache (=ScriptOralia; 126)*. Tübingen: Narr, CD-rom.

Demuth, K., & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners. *Journal of Child Language*, **35**, 99–127.

Endress, A.D., & Mehler, J. (2009). The surprising power of statistical learning: when fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, **60**, 351–367.

Goodsitt, J.V., Morgan, J.L., & Kuhl, P.K. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language*, **20**, 229–252.

Graf Estes, K., Evans, J.L., Alibali, M., & Saffran, J.R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, **18**, 254–260.

Hamann, C., Ohayon, S., Dubé, S., Frauenfelder, U., Rizzi, L., Starke, M., & Zesiger, P. (2003). Aspects of grammatical development in young French children with SLI. *Developmental Science*, **6**, 151–158.

Hallé, P.A., & de Boysson-Bardies, B. (1994). Emergence of an early receptive lexicon: infants' recognition of words. *Infant Behavior and Development*, **17**, 119–129.

Hallé, P.A., Durand, C., & de Boysson-Bardies, B. (2008). Do 11-month-old French infants process articles? *Language and Speech*, **51**, 23–44.

Hirsh-Pasek, K., Kemler Nelson, D.G., Jusczyk, P.W., Wright Cassidy, K., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, **26**, 269–286.

Hunkeler, H. (2005). Aspects of the evolution of the early lexicon in mother–child interactions: case study of two dizygotic twin children between 15 and 26 months. Unpublished ms. University of Rouen.

Johnson, M., Demuth, K., Frank, M., & Jones, B. (2010). Synergies in learning words and their referents. *Advances in Neural Information Processing Systems*, **23**, 1018–1026.

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

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

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

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

MacWhinney, B. (1995). *The CHILDES project: Tools for analyzing talk* (2nd edn.). Hillsdale, NJ: L. Erlbaum.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd edn.). Mahwah, NJ: Lawrence Erlbaum Associates.

Martin, A., Peperkamp, S., & Dupoux, E. (in press). Learning phonemes with a pseudo-lexicon. *Cognitive Science.*

Mattys, S.L., & Jusczyk, P.W. (2000). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, **78**, 91–121.

Meints, K., & Woodford, A. (2008). *Lincoln Infant Lab Package 1.0: A new programme package for IPL, Preferential Listening, Habituation and Eyetracking*. [www document: Computer software & manual]. URL: http://www.lincoln.ac.uk/psychology/babylab.htm.

Morgenstern, A. (2006). *Un JE en construction. Ontogenèse de l'auto-désignation chez l'enfant*. Ophrys: Bibliothèque de Faits de langues.

Nazzi, T., Lakimova, G., Bertoncini, J., Frédonie, S., & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: emerging evidence for crosslinguistic differences. *Journal of Memory and Language*, **54**, 283–299.

New, B. (2006). *Lexique 3: Une nouvelle base de données lexicales*. Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006), April, Louvain.

Pelucchi, B., Hay, J.F., & Saffran, J.R. (2009). Learning in reverse: eight-month-old infants track backward transitional probabilities. *Cognition*, **113**, 244–247.

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

Saffran, J.R. (2001). Words in a sea of sounds: the output of statistical learning. *Cognition*, **81**, 149–169.

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

Seidl, A., & Johnson, E. (2006). Infant word segmentation revisited: edge alignment facilitates target extraction. *Developmental Science*, **9**, 565–573.

Shi, R., Cutler, A., Werker, J., & Cruickshank, M. (2006). Frequency and form as determinants of functor sensitivity in English-acquiring infants. *Journal of the Acoustical Society of America*, **119**, EL61–EL67.

Shi, R., & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science*, **11**, 407–413.

Shukla, M., White, K.S., & Aslin, R.N. (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, USA*, **108**, 6038–6043.

Suppes, P., Smith, R., & Leveillé, M. (1973). The French syntax of a child's noun phrases. *Archives de Psychologie*, **42**, 207–269.

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

Thiessen, E.D. (2011). When variability matters more than meaning: the effect of lexical forms on use of phonemic contrasts. *Developmental Psychology*, **47**, 1448–1458.

Thiessen, E.D., & Saffran, J.R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, **39**, 706–716.

Vihman, M., dePaolis, R., Nakai, S., & Hallé, P.A. (2004). The role of accentual pattern in early lexical representation. *Journal of Memory and Language*, **50**, 336–353.

Received: 6 October 2011
Accepted: 21 June 2012

## Appendix A
## *Items used in Experiments 1 and 2 and their number of occurrences in the CHILDES corpus*

| High-frequency | | Low-frequency (Exp. 1) | | Low-frequency (Exp. 2) | |
|---|---|---|---|---|---|
| [dɑ̃la] | 547 | [dalɑ̃] | 0 | [dɑ̃ty] | 8 |
| [sepuʁ] | 331 | [pesuʁ] | 0 | [seʁɛ̃] | 0 |
| [kwasa] | 242 | [sakwa] | 2 | [kwale] | 0 |
| [vafɛʁ] | 180 | [vɛfaʁ] | 0 | [vakɛl] | 0 |
| [kɔʁɛ̃] | 135 | [kɛ̃ʁo] | 0 | [kola] | 0 |
| [mɛty] | 118 | [tɛmy] | 0 | [mɛply] | 10 |
| [tule] | 117 | [telu] | 0 | [tupa] | 12 |
| [akɛl] | 100[a] | [ɛkal] | 1 | [asyʁ] | 16 |
| [vɛpa] | 92 | [vapɛ] | 0 | [vɛsa] | 17 |
| [naply] | 82 | [nypla] | 0 | [nafɛʁ] | 4 |
| [pasyʁ] | 81 | [sypaʁ] | 0 | [padiʁ] | 2 |
| [vødiʁ] | 61 | [divœʁ] | 0 | [vøpuʁ] | 2 |

[a] Note that the number of occurrences for this vowel-initial item corresponds to that of the sequences of *phonemes*, not syllables. This was because, as mentioned in the stimuli section of Experiment 1, this item was created by deleting the initial segment of a consonant-initial nonword generated by the disyllable-extracting algorithm.

## Appendix B
## *Items used in Experiment 3 and their number of occurrences in the CHILDES corpus*

| High-frequency nonwords: as in Experiment 1 and 2 (Appendix A) | |
|---|---|
| Words: | |
| [lapɛ̃] 'rabbit' | 273 |
| [apɛl] 'call' | 261[b] |
| [pupe] 'doll' | 181 |
| [bonɔm] 'little man' | 168 |
| [vwatyʁ] 'car' | 165 |
| [balɔ̃] 'ball' | 156 |
| [ʃosyʁ] 'shoe' | 116 |
| [kanaʁ] 'duck' | 115 |
| [ʃapo] 'hat' | 88 |
| [gato] 'cake' | 74 |
| [bibʁɔ̃] 'feeding bottle' | 70 |
| [wazo] 'bird' | 70 |

[b] See note in Appendix A.

## Appendix C
## *Corpus distribution of nonword (Experiments 1 and 2) and real word items (Experiment 3) as a function of their position within utterances*

| | High-frequency nonwords | High-frequency words |
|---|---|---|
| Complete utterance | 0.9% | 3.8% |
| Part of utterance | | |
| initial | 15.3% | 1.3% |
| medial | 63.3% | 29.4% |
| final | 20.5% | 65.5% |