Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Beyond transitional probability computations: Extracting word-like units when only statistical information is available

Pierre Perruchet [*], Bénédicte Poulin-Charronnat

CNRS-UMR 5022, University of Bourgogne, Dijon, France

ARTICLE INFO

Article history:
Received 15 April 2011
revision received 17 February 2012
Available online 21 March 2012

Keywords:
Statistical learning
Artificial language
Word segmentation
Chunking
Modeling

ABSTRACT

Endress and Mehler (2009) reported that when adult subjects are exposed to an unsegmented artificial language composed from trisyllabic words such as AB*X*, *Y*BC, and A*Z*C, they are unable to distinguish between these words and what they coined as the "phantom-word" ABC in a subsequent test. This suggests that statistical learning generates knowledge about the transitional probabilities (TPs) within each pair of syllables (AB, BC, and A···C), which are common to words and phantom-words, but, crucially, does not lead to the extraction of genuine word-like units. This conclusion is definitely inconsistent with chunk-based models of word segmentation, as confirmed by simulations run with the MDLChunker (Robinet, Lemaire, & Gordon, 2011) and PARSER (Perruchet & Vinter, 1998), which successfully discover the words without computing TPs. Null results, however, can be due to multiple causes, and notably, in the case of Endress and Mehler, to the reduced level of intelligibility of their synthesized speech flow. In three experiments, we observed positive results in conditions similar to Endress and Mehler after only 5 min of exposure to the language, hence providing strong evidence that statistical information is sufficient to extract word-like units.

Introduction

Seminal studies have shown that, after hearing an artificial language in which invented words have been concatenated without any phonological or prosodic markers, infants (Saffran, Aslin, & Newport, 1996), children (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), and adults (Saffran, Newport, & Aslin, 1996) become more familiar with the invented words of the language than with the part-words straddling word boundaries. The statistical structure being the only cue made available to the learners, this achievement attests to the fact that listeners are able to exploit the statistical information available in the input, and more precisely, the prevalent view is that this form of learning proceeds through the computation of transitional probabilities (hereafter: TPs; Aslin, Saffran, & Newport, 1998). Participants would exploit the fact that TPs between word internal syllables are stronger than TPs between syllables spanning word boundaries.

The role of statistical information in word segmentation

Since this earlier demonstration, the role of statistics in word extraction has been keenly challenged. A part of the debate stems from the a priori argument that statistical information would be too impoverished to be useful in word learning. For instance, Yang (2004) reported that using TPs leads to a far from optimal segmentation of a child-directed corpus of language: Precision was 41.6%, meaning that more than half of the extracted units were not words, and completeness was 23.3%, meaning that almost 80% of the actual words were not extracted.

Other authors have argued that many other sources of information are available to infants. The role of phonological and prosodic features, such as lexical stress placement,

[*] Corresponding author. Address: Université de Bourgogne, LEAD/CNRS, Pole AAFE, Esplanade Erasme, 21000 Dijon, France.
E-mail address: pierre.perruchet@u-bourgogne.fr (P. Perruchet).

0749-596X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2012.02.010

on word discovery has been well documented (e.g., Creel, Tanenhaus, & Aslin, 2006; Curtin, Mintz, & Christiansen, 2005; Thiessen & Saffran, 2007). The question of how statistical and phonological or prosodic cues combine has been investigated in experimental studies in which these cues are either consistent or inconsistent with the word-like units of a continuous speech flow (e.g., Creel et al., 2006; Onnis, Monaghan, Chater, & Richmond, 2005; Shukla, Nespor, & Mehler, 2007; Tyler, Perruchet, & Cutler, 2006). These studies have shown that performance in a word-segmentation test improved with consistent cues and strongly decreased (and potentially dropped at chance level) with inconsistent cues. Other studies have explored how these cues compete as function of age. Although Thiessen and Saffran (2003) reported a prevalence of statistics over prosody in 6-month-old infants, Johnson and Jusczyk (2001) reported that prosodic factors override statistics in 8-month-old infants.

Natural language acquisition also relies on the exploitation of known words to discover new words. To borrow an example given by Dahan and Brent (1999): "If *look* is recognized as a familiar unit in the utterance *Lookhere!* then *look* will tend to be segmented out and the remaining contiguous stretch, *here*, will be inferred as a new unit" (p. 165). Dahan and Brent (1999) and Perruchet and Tillmann (2010) provided experimental evidence of this phenomenon in adults, and Bortfeld, Morgan, Golinkoff, and Rathbun (2005) demonstrated the same capacity in 6-month-old infants. It has been suggested that such lexically-driven segmentation could progressively supersede cues during language development (e.g., Mattys, White, & Melhorn, 2005, p. 493).

Although these studies admittedly reduce the relative importance of statistical learning, they did not challenge the ability of statistical learning processes to trigger word-unit extraction when only statistical information is available. In a recent paper, Endress and Mehler (2009) went far beyond this earlier literature. They acknowledged the capacity of learners to compute TPs, but, "surprisingly", they wrote, "there is no evidence that TP-based computations lead to the extraction of word-candidates." The available experimental evidence, they claimed, "does not imply that the items with stronger TPs are represented as actual word-like units, or even that they have been extracted." (p. 352).

Endress and Mehler's (2009) results

Endress and Mehler (2009) based their conclusion on a set of ingeniously designed experiments in which participants were familiarized with a continuous language containing trisyllabic words, as in the studies cited above, but the words were generated from what the authors coined as a "phantom-word", which was never presented in the language. If the phantom-word is designated as ABC (with each letter standing for a syllable), the heard words were AB*X*, *Y*BC, and A*Z*C (with *X*, *Y*, and *Z* standing for invariant syllables). For instance, participants heard *tazepi*, *mizeRu*, and *tanoRu*, which were all derived from the (unheard) phantom-word *tazeRu*. In this way, the phantom-words had exactly the same TPs between their constituent syllables (i.e., AB, BC, and A···C) than the trisyllabic words composing the language. The reasoning was straightforward: If subjects have learned a word-like unit, that is some acoustical word candidates that could be later associated as a whole to a meaning, they should select words over phantom-words when both are played in a subsequent forced-choice test. However, if they only learned pairwise relations, they should be unable to distinguish between the actual words and the phantom-words.

The results indicate that participants failed to distinguish between words and phantom-words. Chance performance was observed in several experiments in which the number of words and the length of the familiarization phase (from 5 to 40 min) were varied. To quote the authors: "Even when collapsing all 161 participants who took part in the different experiments, no preference for words to phantom-words emerged ($M = 51.2\%$, $SD = 19.4\%$), $t(160) = 0.8$, $p = .438$)" (p. 358). In subsequent experiments, the authors made the word structure perceptually salient, either by introducing 25-ms silent pauses between words or by lengthening the last syllable of each word during the familiarization phase. Subjects now chose words over phantom-words in subsequent forced-choice tests. According to the authors, these findings demonstrate that "despite the general agreement that TP-based computations are crucial for word-learning, other cues seem to be required for actually extracting word-like units." (p. 359). In their view, extracting word-like units requires the construction of positional memories, which would be possible only when prosodic markers of word boundaries are provided in the input.

Theoretical implications of Endress and Mehler's (2009) conclusion

Insofar as Endress and Mehler's conclusion is taken for granted, it should elicit major changes in the current conceptions about the role of statistical learning in word segmentation and language acquisition. We focus below on their theoretical implications with regard to learning models. In the brief outline of the statistical approach above, we have assumed that chunks are inferred from the discovery of the boundaries, which are defined as the points where the predictability of the next element is the lowest. Because the primary aim of computations is to insert word boundaries within a continuous sequence, this view, which is currently prevalent in the literature, is sometimes coined as the *bracketing* approach[1] (Goodsitt, Morgan, & Kuhl, 1993; Swingley, 2005). The consequences of Endress and Mehler's results for a bracketing approach are relatively limited. Indeed, Endress and Mehler do not put into question the fact that learners compute TPs, which are at the core of the bracketing approach. They only challenge the additional postulate according to which TP computations directly trigger word-like unit extraction.

[1] The term "statistical learning" sometimes refers to this specific approach, and is taken as equivalent to "computations of transitional probabilities". Hereafter, "statistical learning" is used as a theoretically neutral label designating any form of exploitation of the statistical structure of the input. The chunk-based models described in this paper are construed as models of statistical learning, as those relying on the computation of TPs.

However, there is another view that, by opposition to the *bracketing* approach, has been called the *clustering* or chunking approach (Swingley, 2005). In this approach, the sensitivity to TPs is nothing else that a by-product of other processes. Instead of looking for units' boundaries, the general strategy shared by all the chunk-based models (e.g., Brent & Cartwright, 1996; Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Perruchet & Vinter, 1998; Robinet et al., 2011; Servan-Schreiber & Anderson, 1990) is that a large number of potential chunks are built online, then selected as a function of their relevance. The Endress and Mehler's conclusion, if valid, is in principle devastating for chunk-based models, because, in a nutshell, Endress and Mehler suggest that learners compute TPs without extracting the words, whereas chunk-based models posit that learners extract the words without computing TPs. We now turn to a brief presentation of chunk-based models and their predictions for the Endress and Mehler's task.

The chunk-based models

Although they share the same general objectives, current chunk-based models follow one of two very different strategies. For the sake of simplicity, we will consider below the MDLChunker of Robinet et al. (2011) as representing the first strategy, and PARSER (Perruchet & Vinter, 1998) as representing the second strategy, although the predictions for the current situation could certainly be generalized beyond these specific models. For a more formal presentation of the models, the reader is referred to the original papers.

In the MDLChunker, the model creates exhaustively all possible new chunks beginning with the shortest ones, then examines whether the consequences of creating a given chunk are positive or negative for the system. Let us consider the sequence CABCRTPABCGAFABCLB. There is a pattern of reoccurrence, namely ABC. If ABC is coded by X, the representation of the sequence becomes shorter: CXRTPXGAFXLB, but this length reduction is obtained at the expense of storing somewhere a new code, ABC = X. The question is to know whether creating the chunk ABC leads to simplify the overall representation of the system, including both the data (i.e., the sequence) and the codes. The length of the data and the length of the codes typically evolve in opposite directions, and therefore, assessing the effect of creating a new chunk on the overall length of the representation is not a trivial matter whenever the data increases in length and complexity. As several other models (Brent & Cartwright, 1996; Frank et al., 2010), the MDLChunker uses a powerful mathematical algorithm, known as the Minimum Description Length principle (MDL, Rissanen, 1978), to solve this trade-off.

A major advantage of the MDLChunker over past MDL-based models (e.g., Brent & Cartwright, 1996) is that it works online. Let us assume that the model has just processed CABCRTPA and now encounters B. The consequences of creating AB as a new chunk are examined. If creating a code for AB leads to a shorter representation of the stored data, which is not overcompensated by the cost of coding AB as a new unit, the code is definitely created, otherwise the code is withdrawn. If AB has been chunked during earlier steps when C is shown, the model may consider creating ABC, hence ensuring the step-by-step formation of longer chunks.

Concerning PARSER (Perruchet & Vinter, 1998), the primary motivation is to account for human behavior in terms of psychologically plausible processes. Based on the observation that, in humans, attentional coding of the ongoing information naturally segments the material into disjunctive parts, PARSER postulates that a sequence such as CABCRTPABCGAFABCLB will be perceived as, say, CAB/C/RT/P/AB/CGA/F/AB/CLB. Each of these randomly determined fragments are created as provisional chunks as they appear in the language. Clearly, some of them are relevant to the structure of the language (here: AB) and all the others are irrelevant. How does the model operate a selection without calling to a sophisticated algorithm? In PARSER, the fate of a new chunk does not depend on the consequences of a retrospective recoding of stored information as in the MDLChunker, but on the probability for the new chunk to be encountered later. The relevant units emerge through a selection process based on forgetting. Due to both decay and interference, forgetting leads to the selection of the most cohesive parts among all parts generated by the initial, presumably mostly irrelevant, chunking of the material. For instance, CAB, once created as a provisional unit, is doomed to quick forgetting, because it does not reoccur later in the sequence. By contrast, AB has more chance of surviving because it will be strengthened on its subsequent occurrences. Once a new chunk has been created, it plays the role of a new primitive, and hence it can become the component of a longer chunk. For instance, if AB is a new primitive, ABC can be created in a subsequent stage of learning (note that ABC could also have been created by chance from the outset instead of AB). This allows the system to build chunks whose components could hardly be perceived in one attentional focus if perception were driven only by the initial primitives in the corpus.

Chunk-based models' predictions in Endress and Mehler's task

Although Endress and Mehler results are, in principle, incompatible with the main tenets of chunk-based models, one needs to examine the actual predictions of these models to eliminate a possible drawback. Indeed, the languages used by Endress and Mehler are somewhat atypical, because given that words were derived from phantom-words, they were closer to each other than in earlier word segmentation studies (e.g., each syllable occurred in two different words). A possibility would be that chunk-based models also fail to extract such words, as participants did in Endress and Mehler's study, or at least need very extensive training to do so.

We performed a large set of simulations using the Endress and Mehler's task with the MDLChunker (Robinet et al., 2011) and PARSER (Perruchet & Vinter, 1998). A point of debate in computational research is the selection of parameter values. The MDLChunker has no free parameter, which is construed as a major advantage by its proponents. In PARSER, the rates of decay and forgetting (and a few other, minor parameters) may be tuned to comply with the materials or the study population. The general strategy adopted in earlier studies (e.g., Frank et al., 2010; Giroux & Rey, 2009; Perruchet & Peereman, 2004; Perruchet, Tyler, Galland, & Peereman, 2004) has been to first apply the parameters used in the initial study (Perruchet & Vinter, 1998), which often turn out to be well-fitted for other objectives. All the subsequent simulations have been performed with these standard parameters.

Participants in Endress and Mehler's experiments were exposed to 75 repetitions of each word (this amounts approximately to 5 min of speech flow) or more. We explored the effect of 5, 10, 15, 25, 35, 55, 75, and 100 repetitions on models' performance. Fig. 1 reports the mean performance of the models in the Endress and Mehler's forced-choice test. Because the test used in these experiments also comprised part-words, that is trisyllabic units spanning word boundaries, the scores were computed for both word/phantom-word and word/part-word pairs. A response was generated for each pair, based on the ratio between the weights of the word and the nonword in the internal lexicon of the models. If the weight of the word was stronger than the weight of the nonword, then the score was set to 1. If both weights were equal, the score was set to 0.5, and if the weight of the word was smaller than the weight of the nonword, the score was set to 0. Each point from the curves was averaged over 100 runs, with each run using a different language (with regard to word order). All simulations were performed with U-learn (Perruchet, Robinet, & Lemaire, submitted for publication), which is freely available to the following URL: http://leadserv.u-bourgogne.fr/~perruchet/.

PARSER performed above chance, $t(99) = 15.59$, $p < .001$, with only five repetitions of each word, while the MDLChunker was still at chance. However, the MDLChunker's predictions were above chance, $t(99) = 3.79$, $p < .001$, with only ten repetitions. With 25 repetitions, the performances of the two models were nearly identical, then the MDLChunker moved toward asymptote slightly quicker than PARSER. These observations hold for both word/phantom-word and word/part-word pairs, which do not substantially differ.

A more detailed analysis of these predictions and their implications for the models is postponed to the general discussion. For our current concern, the major conclusion is that irrespective of their specific instantiation, chunk-based models extracted the words of the Endress and Mehler's language without any difficulty. Both models reached asymptotic performances with the shortest amount of exposure used in Endress and Mehler's experiments. Still more strikingly, PARSER performed above chance with word/phantom-word pairs with only five repetitions of each word during familiarization, whereas participants in Endress and Mehler's Experiment 1d failed to do so with 600 repetitions. Neither PARSER nor the MDLChunker are aimed at fitting the time course of human learning precisely, and it has been observed from the outset that models tend to outperform human participants (e.g., Perruchet & Vinter, 1998). But even taking this observation into account, the discrepancy between models' predictions and the Endress and Mehler's data is so drastic that it raises a major challenge for chunk-based models. Before speculating further on how to deal with this challenge, however, we have to raise a preliminary question: Is the failure of the Endress and Mehler's participants to select (heard) words over (unheard) phantom-words a reliable outcome?

The present study

The following experiments were aimed at replicating Endress and Mehler's (2009)'s study. We focused on their first experiment, in which subjects were familiarized with a speech flow comprising no prosodic marker of word boundaries during 5 min (i.e., the shorter duration of exposure explored by the authors). To anticipate, Experiment 1 revealed that participants selected words over phantom-

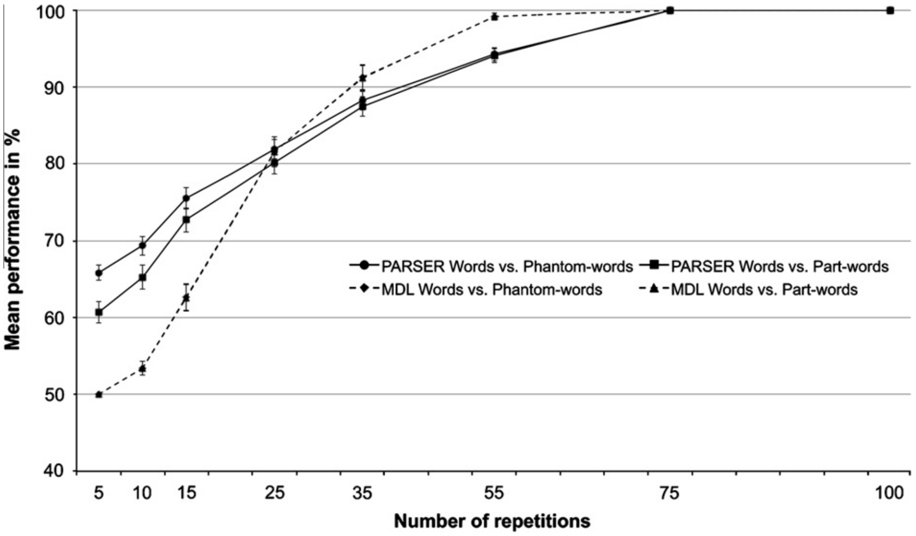

Fig. 1. Mean performances of MDLChunker (Robinet et al., 2011) and PARSER (Perruchet and Vinter, 1998) on the two types of test trials (words vs. phantom-words and words vs. part-words) as a function of the number of repetitions. Note that the two curves for the MDLChunker overlap (see Footnote 2). Error bars represent standard errors.

words in those conditions, in striking contrast with Endress and Mehler's results. Experiments 2 and 3 were intended to examine whether procedural differences could account for the discrepancy between our results and those of Endress and Mehler.

Experiment 1

Method

Participants

A total of 40 undergraduate students from the University Paris-Descartes, France, participated in the experiment in partial fulfillment of a course requirement. All subjects were native French speakers. Participants were randomly assigned to one of the two experimental groups (Language 1, $N = 20$; Language 2, $N = 20$).

Materials

To control for possible preferences for some acoustic utterances, we created two counterbalanced versions of the artificial language. Language 1 (including words and phantom-words) was strictly identical to the language used in Endress and Mehler's (2009) Experiment 1 (Appendix A1). For Language 2, a complete switching between words and phantom-words was not possible, given there are more words than phantom-words, but, as a close approximation, the two phantom-words from Language 1 served as words in Language 2, while two randomly selected words from Language 1 served as phantom-words in Language 2 (Appendix A2).

As in Endress and Mehler, the speech was synthesized through the MBROLA (Multiband Resynthesis Overlap Add) speech synthesizer (http://tcts.fpms.ac.be/synthesis/; Dutoit, Pagel, Pierret, Bataille, & Van Der Vrecken, 1996) with the fr2 diphone database. The mean syllable duration was 232 ms. The resulting WAV files were modified using CoolEdit. Progressive fades in and out were applied to the first and last 5 s of each part of stream to avoid word boundary cues. The speech stream was played through headphones connected to a Macintosh computer.

Procedure

Participants first completed a pre-training phase consisting of 10 trials. Two syllables were played on each trial, one of which was 'so'. The task was to indicate whether 'so' was in first or second location by pressing on one of two predefined keys. This was intended to familiarize the participants with the final forced-choice test.

Then participants were told that they would be exposed to an imaginary language. They were told to carefully listen to the speech flow, mimicking the attitude they may have when they are listening to music. Each of the six words occurred 75 times. The words were pseudo-randomly ordered for each participant, without immediate repetition. This phase of familiarization to the language lasted about 5 min. At the end, participants were told that they would be presented with pairs of items, and that they would have to judge, for each pair, which item was a word of the imaginary language. The two trisyllabic items of each pair were separated by a 500-ms silent interval.

In Endress and Mehler, word/phantom-word pairs were mixed with word/part-word pairs during the test. We reasoned that adding word/part-word pairs could bias the choices on the word/phantom-word pairs, which were of primary interest. Indeed, selecting the words in the word/part-word pairs is relatively easy, as attested by the high rate of success in Endress and Mehler's experiments. Let us assume that a word/part-word pair and a word/phantom-word pair involving the same word occur in close temporal succession. Participants may correctly select the word in the first pair, then continuing to select the word in the second pair only for the sake of response consistency. Such sequential effects would be unable to account for participants' failure to discriminate words from phantom-words in Endress and Mehler's experiment, given these effects should enhance the selection of the word in the word/phantom-word pairs. However, we wanted to avoid that positive results, if observed, could be ascribed to sequential effects. As a consequence, only word/phantom-word pairs were played. Given that there were two phantom-words, only two words out of the six were used during the test, to avoid any possibility of gaining information from the relative frequency of test items. The two words selected for Language 1 served as phantom-words for Language 2 and vice versa, so that participants ascribed to both languages were exposed to exactly the same test pairs, with the word/phantom-word status of the items being reversed. There were four word/phantom-word pairs (Appendix A). Each test pair was presented twice in different item orders, resulting in eight pairs of items, the order of which was randomized for each subject.

Results and discussion

Participants showed a significant preference for words over phantom-words, $M = 63.12$, $SD = 24.01$, $t(39) = 2.93$, $p = .001$, *Cohen's* $d = 0.546$. This finding strikingly departs from Endress and Mehler's results. Given that only one of our languages was borrowed from Endress and Mehler, a possibility is that this effect comes from the counterbalanced, new version of the language. Our results do not lend support to this possibility: Language 1 and Language 2 led to the very same mean performance, $M = 63.12$, $SD = 27.05$ and $M = 63.12$, $SD = 21.26$, respectively.

A possible explanation for this departure could stem from the differences regarding the instructions given to participants. Participants in Endress and Mehler's experiments "were told that they would listen to a monologue in an unknown language ("Martian"), and were instructed to try to find the words in the monologue." (Endress & Mehler, 2009; p. 355). By contrast, we used incidental learning instructions in Experiment 1. We did not borrow Endress and Mehler's intentional instructions all simply for the sake of compliance with the most common practice: In an overwhelming proportion of studies on word segmentation in adults, any attempt to analyze the speech flow is tacitly or explicitly discouraged (e.g., Perruchet & Tillmann, 2010; Saffran et al., 1996; Toro, Sinnett, & Soto-Faraco, 2005). Although the motivation for using incidental learning instructions as a standard is rarely made explicit, we guess that this practice naturally derives from the primary interest of the contributors to this literature, which is the formation of the lexicon in infants. Inferring how infants process a continuous speech flow is largely a matter of speculation, but, it seems relatively safe to posit that infants do *not* process the speech flow as adults do when they are intentionally searching for the structure of the language.

Using intentional learning instructions may have influenced subjects' performance in Endress and Mehler (2009)'s experiments. We are unaware of any systematic comparison between incidental and intentional instructions in word-segmentation research, but in the related literature on implicit learning, instructions asking participants to search for rules is known to have detrimental consequences with regards to more incidental instructions (Reber, 1976), especially when the task is complex (Reber, Kassin, Lewis, & Cantor, 1980). In a word-segmentation paradigm, a possibility is that participants under intentional learning instructions look for bigram statistics, storing a syllable and its successor in working memory and tracking whether the next occurrence of the same syllable will have the same or a different successor. This or another hypothesis-testing operation may prevent the online attentional processing of sequential information, which may be required for the creation of longer chunks of syllables.

Experiment 2 was aimed at exploring the role of instructions in word segmentation, by directly comparing the effects of incidental and intentional instructions in independent groups of subjects. Our hypothesis was that intentional learning instructions might have prevented word formation in the participants of Endress and Mehler's experiments.

Experiment 2

Method

Participants

Forty undergraduate students at the University of Bourgogne in Dijon, France, participated in the experiment in partial fulfillment of a course requirement. All were native French speakers. Participants were randomly assigned to incidental or intentional instructions, and within each condition, to one of two counterbalanced languages (with $N = 10$ in each cell).

Materials and procedure

For the incidental group, the pre-training phase, the familiarization phase and the test phase were exactly identical to those of Experiment 1. For the intentional group, only the instructions differed. These instructions were as similar as possible to the instructions reported in Endress and Mehler (2009). Before the familiarization phase, participants were told that they would have to listen to a monologue in an unknown language (in "Martian") and were instructed to try to find the words in the monologue. At test, participants had to choose between the two items of each pair the one that was more likely to be a Martian word.

Results and discussion

Participants again preferred words over phantom-words under incidental instructions, $M = 62.50$ $SD = 19.45$, $t(19) = 2.87$, $p = .010$, *Cohen's* $d = 0.643$, hence replicating Experiment 1. However, contrarily to our prediction, the effect was also observed under intentional instructions, $M = 63.13$, $SD = 24.83$, $t(19) = 2.36$; $p = .029$, *Cohen's* $d = 0.538$. An ANOVA was carried out with Instructions (incidental, intentional) and Language (language 1, language 2) as between-subject variables. There was no main effect of Instructions, $F(1,36) = 0.008$, $p = .931$. Likewise, there was no main effect of Language, $F(1,36) = 0.61$, $p = .439$, and no significant Instructions × Language interaction, $F(1,36) = 0.008$, $p = 931$.

Given that participants from the incidental and intentional groups were exposed to the same set of test pairs, their pattern of responses can be directly compared. We computed the correlation over the eight pairs of test items between the two conditions. This correlation was positive and significant, $r(6) = .739$, $p = .036$, indicating that participants from the two groups showed similar patterns of responses to the test pairs.

These results have contrasting implications. On the one hand, they strengthen the general contention that considering statistical structure is sufficient to extract word-like units, by replicating the results from Experiment 1 on incidental instructions, and generalizing the conclusion to intentional instructions. But on the other hand, they rule out our hypothesis that the failure of participants in Endress and Mehler (2009) to show a preference for words over phantom-words would be due to the use of intentional instructions. In fact, manipulating the instructions had no effect, irrespective of whether mean performances or response patterns were considered. These results are interesting on their own, given that there was no earlier comparison between incidental and intentional instructions in earlier word-segmentation research. However, for our present concern, this raises a new question: Is there one or several other procedural differences that could account for the discrepancy between our results and those of Endress and Mehler?

Excluding contextual elements that are typically not reported in experimental papers (e.g., the gender of the experimenter), the only discernable difference is related to the list of test pairs. Experiments 1 and 2 involved only word/phantom-word pairs, whereas in Endress and Mehler, word/phantom-word pairs were mixed with a larger number of word/part-word pairs. As detailed in the Method section of Experiment 1, removing the word/part-word pairs from the test was intended to prevent sequential effects that could have artificially *enhanced* the selection of the word in the word/phantom-word pairs. However, it cannot be excluded that mixing word/phantom-word pairs with word/part-word pairs could *impair* the performance on word/phantom-word pairs, if only because of increased interference and complexity. To examine this possibility, Experiment 3 is a replication of Experiment 1, except that the test involved both word/phantom-word and word/part-word pairs, as in Endress and Mehler.

Experiment 3

Method

Participants

A total of 28 undergraduate students from the University of Bourgogne in Dijon, France, participated in the experiment in partial fulfillment of a course requirement. All subjects were native French speakers. Participants were randomly assigned to one of two experimental groups (Language 1, $N = 14$; Language 2, $N = 14$).

Materials and procedure

The pre-training and the familiarization phases of Experiment 3 were identical to those of Experiment 1. However, during the test phase, there were six word/phantom-word pairs, and 12 word/part-word pairs, as in Endress and Mehler. If all words are represented as ABC, part-words were either of the BCA (six part-words) or CAB (six part-words) types. Each test pair was presented twice in different item orders, resulting in 36 pairs the order of which was randomized for each subject (Appendix B).

Results and discussion

The main results are shown in Fig. 2. Unsurprisingly, participants showed a significant preference for words over part-words, $M = 71.57$, $SD = 20.79$, $t(27) = 5.49$, $p < .001$, Cohen's $d = 1.04$. There was no difference according to whether BCA or CAB part-word types were considered, $t(27) = 0.49$, $p = .631$, and there was no difference between Languages, $t(26) = 1.43$, $p = .165$.

Despite the presence of intermixed test pairs including a part-word, participants again preferred words over phantom-words, $M = 59.82$, $SD = 17.72$, $t(27) = 2.93$, $p = .007$, Cohen's $d = 0.554$. As in the prior experiments, the scores were closely similar for the two counterbalanced languages, $M = 60.12$, $SD = 18.25$ vs. $M = 59.52$, $SD = 17.86$, $t(26) = 0.09$, $p = .931$.

When the scores in the two kinds of test pairs are directly compared, it appears that participants were significantly better at selecting the words when they were paired with the part-words than when they were paired with the phantom-words, $t(13) = 3.82$, $p = .002$. In terms of Cohen's d, the size of the former effect falls within the "large" range, while the latter falls within the "medium" range in our three experiments. Discussion of this result is postponed to the next section.

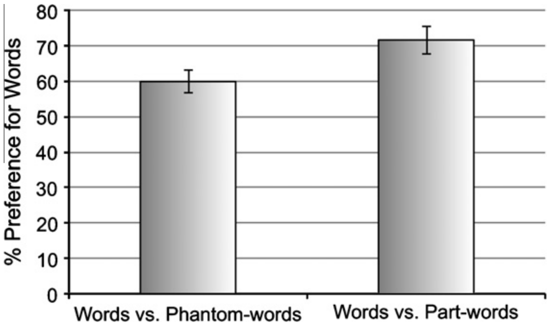

Fig. 2. Proportion of correct responses on the two types of test trials (words vs. phantom-words and words vs. part-words) in Experiment 3. Error bars represent standard errors.

General discussion

In three experiments, participants showed a significant preference for words over phantom-words after only 5 min of exposure to the language. Overall, the effect appears remarkably stable, with the rate of correct responses on eight independent groups (considering the two counterbalanced languages separately) ranging from 59.52% to 66.25% (with chance set to 50%). Before discussing the theoretical implications of these results, we will deal with two preliminary issues in turn. First, how can the differences with the Endress and Mehler's results be explained? Second, is the preference for phantom-words over part-words observed in Experiment 3 consistent with chunk-based models of word segmentation?

Accounting for the differences with Endress and Mehler, 2009 results

Our results strikingly depart from those of Endress and Mehler, who saw chance performance in several experiments, even though some of them involved a much longer duration of exposure to the language (until 40 min in Experiment 1d). We have examined and rejected two possible explanations for the failure of Endress and Mehler to get the same effect. A first possibility is that Endress and Mehler's results were due to the use of rather unusual intentional learning instructions. This hypothesis was clearly ruled out in Experiment 2, which showed a significant preference for words over phantom-words under the very same intentional instructions as used by Endress and Mehler. Experiment 3 led us to reject a second possibility, according to which performance on the word/phantom-word pairs depended on whether the test also includes word/part-word pairs. A reliable effect was still present when the test comprised the same word/part-word pairs as in Endress and Mehler. An additional possibility would be that the acoustical properties of words and phantom-words have biased the Endress and Mehler's results, which were based on a unique language for each experiment. In all the experiments above, we used two versions of the artificial language, with one version being identical to the language of Endress and Mehler, and the other being created by reversing words and phantom-words. Performance on the two languages was nearly identical in each experiment, ruling out the presence of acoustical biases as a potential explanation of the departure.

At this juncture, after a careful scrutiny of the Endress and Mehler's procedure, the only remaining explanation we envision for the discrepant results stems from the observation that both sets of experiments involved (inevitably) different samples of participants. Note that we get the same results with students from our university (Experiments 2 and 3), and with students from another university (Experiment 1), who differed along several characteristics (e.g., students from our university may have been exposed to artificial languages over the few past years, whereas no experiment of this type had been carried out in the other university). This makes it very unlikely that our positive results are dependent on some idiosyncratic characteristics of the studied samples. However, one point deserves particular attention. Although both Endress and Mehler and ourselves used the same French diphone database to create the synthesized languages, our experiments were carried out with French participants, while Endress and Mehler ran their experiments with Italian participants.

Endress and Mehler (2009) wrote: "Pilot tests [...] showed that Italian native speakers find synthesized speech with the fr2 voice more intelligible than speech synthesized with the available Italian diphone databases" (Endress & Mehler, 2009, p. 354). This observation, however, does not tell much about the resulting level of intelligibility. A reasonable hypothesis is that perceptual discrimination of the speech flow is more difficult when the diphone database used for language synthesis was not extracted from the participants' mother language, as in Endress and Mehler, than when it was, as in our experiments. If this hypothesis is correct, then the differences between Endress and Mehler's results and our results should not be circumscribed to the word/phantom-word pairs, but should extend to word/part-word pairs. In our Experiment 3, which is the only experiment comprising such pairs, participants who were exposed to the same language as used in Endress and Mehler choose the words over the part-words on 77.08% of the trials. The score in Endress and Mehler's Experiment 1a, which involves the same duration of exposure to the same language, was 69.9%. Admittedly, the difference is not drastic, but there is at least a numerical support for the idea that Italian participants in Endress and Mehler may have experienced some perceptual difficulties with the French database.

Why are part-words easier to reject than phantom-words?

When word/part-word pairs were introduced among the test items in Experiment 3 for the sake of using the same test lists as Endress and Mehler (2009), participants were significantly better at selecting the words in these pairs than in the word/phantom-word pairs. The same effect was obtained in all the experiments of Endress and Mehler, who construed this part of their results as a proof that "participants learn the TP-structure of the streams" (Endress & Mehler, p. 358). They based their reasoning on the premise that the mean TP between each pair of syllables was lower for the part-words than for the phantom-words. This premise looks as sensible, given that phantom-words were matched with the words in this regard. However, a close examination of the data suggests that the support this effect provides for TP computations could be less compelling than Endress and Mehler contended. Given the particularities of the Endress and Mehler's materials, the mean pairwise TPs computed on part-words and phantom-words (or words) were in fact quite close one each other. We obtained .41 and .50, respectively. Whether detecting a so small difference is possible within 5 min of exposure to the language and sufficient to generate a so large and robust effect during the test is questionable. Moreover, the predictions based on the exploitation of TPs depend on the events that are considered for TP computations. In the calculation above, only pairwise relations were considered (i.e., B|A, C|B, and C|A for an ABC item), as in Endress and Mehler. If second order conditionals were taken into account (i.e., C|AB), the effect would be inverted. Indeed, the transitional probability C|AB turns out to be notably higher for the part-words (TP = .48) than for the phantom-words (TP = 0), hence making the phantom-words easier to reject than the part-words. Thus the support the observed result brought out for models based on the computation of TPs is more apparent than real.

However, irrespective of the ability (or inability) of the models relying on TP computations to account for it, this result seemingly provides a strong challenge for chunk-based models. As shown in the simulated data reported in Fig. 1, both MDLChunker and PARSER failed to predict the observed pattern: The mean performances of the models for the word/part-word pairs never exceeded those for the word/phantom-word pairs, whatever the amount of training.[2] The reason is straightforward: The principles underpinning these models prevent the formation of any unit that has never been processed during the familiarization phase. Phantom-words have never been encountered, while part-words have been encountered, although less frequently than words. As a consequence, the rate of selection of words when they are paired with phantom-words can only be *higher* than when they are paired with part-words, which stands in contradiction with the observed pattern. Is there a way to reconcile chunk-based models with this particular result?

Addressing this issue gives us the opportunity to recall that the primary objective of chunk-based models is to reproduce the internal lexicon of the learner. They are not aimed at describing how a learner having acquired a given lexicon proceeds to select a response in a forced-choice recognition task involving a word and another item. The algorithm we have used in our simulations above to infer a score from a mental lexicon is an oversimplification, because it does not take into account the characteristics of the test items, like their frequency or the similarity of the items within a pair, which may affect the recognition scores while the lexicon is kept unchanged.

To illustrate, let us consider the test used in Experiment 3, which was borrowed from Endress and Mehler's Experiment 1. A given word was paired, on successive trials, with two part-words (BCA or CAB) and one phantom-word. This design entails that part-words occurred three times less of-

[2] It may be seen in Fig. 1 that for the MDLChunker, the curves representing the predicted rate of correct responding in the word/phantom-word pairs and in the word/part-word pairs are nearly superimposed. This reflects the fact that the MDLChunker creates virtually no part-word. For PARSER, the two curves are dissociated in the first stages of training, because PARSER creates a few part-words, which are removed from the lexicon with subsequent exposure to the language. This difference reflects the mode of chunk creation in the two models: Chunks are created by the MDLChunker only if this creation has positive consequences on the coding of stored data (a condition that part-words do not fulfilled), while in PARSER, chunks are created on a random basis then subsequently selected through decay and interference.

ten than words and phantom-words across the 36 test trials. A possibility is that phantom-words became more familiar than part-words during the test, due to their higher frequency, hence making their discrimination with words harder. Chunk-based models are in no way prepared to deal with this kind of situation.

Let us assume that the same pattern of results would persist after the frequency of part-words and phantom-words played during the test has been controlled. A simple thought experiment is sufficient to give evidence that no strong conclusions could be drawn, nevertheless. Consider a lexical decision task in which one would have to decide whether the following items are English words: *potention*, *nalinten*, and *tentialdi*. None are words, but it looks very likely that rejecting *potention* would be more difficult (maybe with a longer latency or an increased probability of errors) than rejecting *nalinten*, and *tentialdi*. Now, *potention* is a ''phantom-word'' (composed from the real words *potential*, *intention*, and *position*), which has likely never been heard in the past, while *nalinten* and *tentialdi* are part-words, which have high chance of having already been heard (e.g., in ''*original intention*'' and ''*potential diversion*'' respectively). Presumably, the relative difficulty of distinguishing words from phantom-words should be referred to some well-known phenomena in the psychological literature, notably the false recognition of unstudied prototypes in Posner and Keele's studies (e.g., 1970), or the false recall or recognition of lures associated with the studied items in the DRM tasks (Roediger & McDermott, 1995). Further studies are needed to clarify this issue. For the present concern, however, these considerations are sufficient to make the point that the differences observed as a function of whether a word is paired with a phantom-word or a part-word can hardly be conceived as a challenge for chunk-based models.

As an aside, at this juncture, the reader may wonder why the just mentioned explanation could not be applied to the main results of Endress and Mehler, namely the indistinctiveness of words and phantom-words for the participants. We fully agree that in so far as phantom-words are assimilated with the prototypes of word classes, a prototype effect could account for the lack of preference for the words themselves (the standard prototype effect would even predict a preference for phantom-words over words). This interpretation of the Endress and Mehler 's results would be fully compatible with a chunk-based approach. However, this leaves the question open: Why would a prototype effect be stronger in Endress and Mehler's experiments than in our own study? Without any hint for a response to this question, we suggest that the low level of perceptual discrimination of the speech flow in Endress and Mehler's experiments remains the best hypothesis to-date to account for their null results, as proposed in the prior section.

The surprising power of statistical learning

Let us return now to the main result of the present series of experiments, namely the preference for words over phantom-words after a few minutes of exposure to the language. It is worth stressing that our disagreement with Endress and Mehler is related to their empirical findings, not to the logic of their reasoning. On the contrary, we fully acknowledge that the method they propose is a very ingenious way of teasing apart two opposite conceptions regarding the exploitation of the statistical information embedded in a speech flow. Endorsing their rational and borrowing their methodology, we were able to demonstrate that exposure to a speech flow comprising only statistical cues is sufficient to create word-like units. Training with unsegmented speech results in the formation of word-like units, rather than in a string of sounds linked by TPs varying on a continuous dimension, or in a set of fragments that does not map the actual constituents of the language.

Some prior evidence for the same conclusion can be found in the literature (Giroux & Rey, 2009; Graf Estes, Evans, Alibali, & Saffran, 2007; Saffran & Graf Estes, 2006; Saffran & Wilson, 2003). For instance, Graf Estes et al. (2007, Exp 2) played to 17-month-old infants a continuous language comprising four invented words, then introduced an object-labeling task. Novel objects were paired either with the invented words or with part-words. Infants learned the labels in the former case, but not in the latter, despite the fact that words and part-words were frequency-balanced. These results are clearly consistent with the hypothesis that statistical segmentation processes generate word-like units. However, critics may still argue that words were built after the exposure to the artificial language to cope with the object-labeling task. In this view, the computation of TPs during the listening phase would help subsequent word segmentation, but would not lead to word extraction on its own. The findings reported in the present paper provide certainly the most compelling evidence to date for the view that statistical information is sufficient to extract the words as functional units from a continuous speech flow.

Needless to say, acknowledging the surprising power of statistical learning does not imply that statistical structure is the only (and even the main) source of information to be exploited in word discovery. As detailed in Introduction, a huge number of studies gives evidence for the contribution of phonological, prosodic, and contextual cues (e.g., Creel et al., 2006; Curtin et al., 2005; Dahan & Brent, 1999; Johnson & Jusczyk, 2001; Onnis et al., 2005; Thiessen & Saffran, 2007), which have proven to interact with statistical cues (e.g., Perruchet & Tillmann, 2010; Shukla et al., 2007). The Endress and Mehler's (2009) study, which shows that introducing pauses between words or lengthening the final syllables of the words helps word segmentation, provides additional supports to the role of prosodic factors. What our study clearly demonstrates, however, is that in striking contradiction to Endress and Mehler's claims, prosodic cues are in no way necessary to extract word-like units from the speech flow.

A final note about word segmentation models

This paper was not aimed at ruling out a bracketing approach, in which words are inferred from the discovery of the boundaries defined as the points where the predictability of the next element is the lowest. In particular, we did not intend to rule out the idea that participants might have computed TPs, although the observation that words were preferred over phantom-words despite they were matched with regards to the TP structure provides some challenge for this view. Likewise, the analysis of Endress and Mehler's materials showed that the strong preference of words over part-words, which was observed both in Endress and Mehler's experiments and in our Experiment 3, was far less consistent with the TP structure than Endress and Mehler claimed.[3] However, providing compelling evidence against the exploitation of TP structures would have required further experimental manipulations.

Our primary motivation for examining the Endress and Mehler's results was their inconsistency with the chunk-based models of statistical learning. These results indeed suggested that learners compute TPs without extracting the words, whereas chunk-based models posit that learners extract the words without computing TPs. A possible account of Endress and Mehler's results within a chunk-based framework relied on the hypothesis that the particular structure of their languages could have made their decomposition into words especially difficult, but our simulations with the MDLChunker (Robinet et al., 2011) and PARSER (Perruchet & Vinter, 1998) showed that both models extracted the words quickly and efficiently in these conditions. The data reported in the three experiments above restore the viability of chunk-based models, which was downsized by the Endress and Mehler's apparent counterevidence. More specifically, we argue below that the simulations performed in the present study brings added value to PARSER.

Why should PARSER be privileged over the MDLChunker, given that both models made closely similar predictions in our simulations (see Fig. 1)? Admittedly, our contention does not stem from the relative ability of the two models to extract chunks. As shown in earlier comparative studies (Frank et al., 2010; Robinet et al., 2011) and confirmed above, the predictions from the two models do not substantially differ. But this is precisely this lack of difference, which makes PARSER performance remarkable. Indeed, the MDLChunker uses powerful and specially designed mathematical algorithms, which have a very low psychological plausibility, unless assuming a very smart cognitive unconscious. But the predictions of a MDL model may be used as a benchmark against which the efficiency of other, more plausible models, can be assessed. Now, it turns out that, without any adjustment of its parameters, PARSER, which relies on simple and ubiquitous psychological mechanisms, performed as well as the MDLChunker. As shown in Fig. 1, with a very small corpus, PARSER even outperformed the MDLChunker.

This achievement of PARSER is all the more worth noting as the model has proven to be able to account for data for which it was not a priori prepared. PARSER was initially designed to account for Saffran, Newport et al.'s (1996) results, which are consistent with the mere exploitation of raw frequencies. The model nevertheless turned out to be able to reproduce the effect of TPs (Aslin et al., 1998; Perruchet & Pacton, 2006; Perruchet & Peereman, 2004). Moreover, while Aslin et al. only considered forward TPs, subsequent studies (Pelucchi, Hay, & Saffran, 2009; Perruchet & Desaulty, 2008) showed that participants also relied on backward TPs to segment a continuous speech flow. Again, PARSER was able to reproduce the same outcome, without any change or parametric adjustment with regard to the initial model (Perruchet & Vinter, 1998). Moreover, PARSER turned out to be able to predict the interaction between statistic and acoustic or contextual factors observed in Perruchet and Tillmann (2010). The present study suggests that the model performs as well as the MDLChunker, which relies on a specially designed and powerful mathematical algorithm. We hope that cumulative evidence lending support to PARSER will provide a sufficient motivation to reconsider the current prevalence of explanations based on the notion of TP computations in the field of word segmentation.

Acknowledgments

This work has been supported by the Centre National de la Recherche Scientifique (CNRS, UMR 5022), and by the Université de Bourgogne. We are grateful to Sébastien Pacton, Charlotte Carrion, Cédric Foucault, Pascal Morgan, and Aurélie Quévreux for help with the experiments.

Appendix A: Materials used in Experiments 1 and 2

Group 1

Familiarization phase	Test pairs	
Words	Words	Phantom-words
Familiarization phase	Test pairs	
Words	Words	Phantom-words
tazepi	tazepi	tazeRu
mizeRu	mikula	tazeRu
tanoRu	tazepi	fekula
fekupi	mikula	fekula
mikula		
fenola		

Group 2

Familiarization phase	Test pairs	
tazeRu	tazeRu	tazepi
fezepi	fekula	tazepi
tanopi	tazeRu	mikula
mikuRu	fekula	mikula
fekula		
minola		

[3] We allude here to the analysis reported above comparing performances in the word/part-word pairs and in the word/phantom-word pairs. However, this analysis is endowed with more general implications. In fact, words were consistently preferred over part-words even though the mean pairwise TPs computed on words and part-words were much closer one each other (.50 and .41, respectively) than in most comparable studies. This casts some doubt on the fact that TP computations might have been responsible for the selection of words over part-words in this specific paradigm, and by way of generalization, in all prior studies on word segmentation of artificial languages.

Appendix B: Test pairs used in Experiment 3

Group 1

Word/Phantom-word trials		Word/Part-word trials			
Words	Phantom-words	Words	Part-words (BCA)	Words	Part-words (CAB)
Word/Phantom-word trials		Word/Part-word trials			
Words	Phantom-words	Words	Part-words (BCA)	Words	Part-words (CAB)
tazepi	tazeRu	tazepi	zepimi	tazepi	pitano
mizeRu	tazeRu	mizeRu	zeRufe	mizeRu	Rufeno
tanoRu	tazeRu	tanoRu	noRumi	tanoRu	Rufeku
fekupi	fekula	fekupi	kupita	fekupi	pimiku
mikula	fekula	mikula	kulafe	mikula	lataze
fenola	fekula	fenola	nolata	fenola	lamize

Group 2

Word/Phantom-word trials		Word/Part-word trials			
tazeRu	tazepi	tazeRu	zeRufe	tazeRu	Rutano
fezepi	tazepi	fezepi	zepimi	fezepi	pimino
tanopi	tazepi	tanopi	nopife	tanopi	pimiku
mikuRu	mikula	mikuRu	kuRuta	mikuRu	Rufeku
fekula	mikula	fekula	kulami	fekula	lataze
minola	mikula	minola	nolata	minola	lafeze

References

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

Bortfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech stream segmentation. *Psychological Science, 16*, 298–304.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*(1–2), 93–125.

Creel, S. C., Tanenhaus, M. K., & Aslin, R. N. (2006). Consequences of lexical stress on learning an artificial lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 15–32.

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

Dahan, D., & Brent, M. R. (1999). On the discovery of novel word-like units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General, 128*, 165–185.

Dutoit, T., Pagel, N., Pierret, F., Bataille, O., & Van Der Vrecken, O. (1996). The MBROLA Project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proc. ICSLP'96, Philadelphia* (Vol. 3, pp. 1393–1396).

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.

Frank, M. C., Goldwater, S., Griffiths, T., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition, 117*, 107–125.

Giroux, I., & Rey, A. (2009). Lexical and sub-lexical units in speech perception. *Cognitive Science, 33*, 260–272.

Goodsitt, J. V., Morgan, J. L., & Kuhl, P. K. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Development, 20*(2), 229–252.

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

Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44*, 548–567.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134*(4), 477–500.

Onnis, L., Monaghan, P., Chater, N., & Richmond, K. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language, 53*, 225–237.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition, 113*(2), 244–247.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition, 36*(7), 1299–1305.

Perruchet, P., Robinet, V., & Lemaire, B. (submitted for publication). U-Learn: Finding optimal coding units from unsegmented sequential databases.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: Two approaches, one phenomenon. *Trends in Cognitive Sciences, 10*, 233–238.

Perruchet, P., & Peereman, R. (2004). The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics, 17*(2–3), 97–119.

Perruchet, P., & Tillmann, B. (2010). Exploiting multiple sources of information in learning an artificial language: Human data and modeling. *Cognitive Science, 34*, 255–285.

Perruchet, P., Tyler, M. T., Galland, N., & Peereman, R. (2004). Learning non-adjacent dependencies: No need for algebraic-like computations. *Journal of Experimental Psychology: General, 133*, 573–583.

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

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83*, 304–308.

Reber, A. S. (1976). Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 2*(1), 88–94.

Reber, A. S., Kassin, S. M., Lewis, S., & Cantor, G. (1980). On the relationship between implicit and explicit modes in the learning of a complex rule structure. *Journal of Experimental Psychology: Human Learning and Memory, 6*(5), 492–502.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*(5), 465–471.

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDLChunker: A MDL-based cognitive model of inductive learning. *Cognitive Science, 35*, 1352–1389.

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 803–814.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.

Saffran, J. R., & Graf Estes, K. M. (2006). Mapping sound to meaning: Connections between learning about sounds and learning about words. In R. Kail (Ed.), *Advances in child development and behavior* (pp. 1–38). New York: Elsevier.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606–621.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*(2), 101–105.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy, 4*, 273–284.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(4), 592–608.

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology, 54*, 1–32.

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

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.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: Acquisition of stress-based strategies for word segmentation. *Language Learning and Development, 3*, 75–102.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97*(2), B25–B34.

Tyler, M. D., Perruchet, P., & Cutler, A. (2006). A cross-language comparison of the use of stress in word segmentation. *Journal of the Acoustical Society of America, 129*(5), 3087.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences, 8*(10), 451–456.