Acta Psychologica 149 (2014) 1–8

Contents lists available at ScienceDirect

Acta Psychologica

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

New evidence for chunk-based models in word segmentation

Pierre Perruchet [a,*], Bénédicte Poulin-Charronnat [a], Barbara Tillmann [b], Ronald Peereman [c]

[a] Université de Bourgogne, LEAD/CNRS, UMR5022, Pôle AAFE, 11 Esplanade Erasme, 21000 Dijon, France
[b] CNRS, UMR5292, INSERM U1028, Lyon Neuroscience Research Center, Auditory Cognition and Psychoacoustics Team, Université of Lyon I, Lyon, France
[c] Laboratoire de Psychologie et Neurocognition, CNRS UMR5105, Université Grenoble Alpes, Bâtiment Sciences de l'Homme et Mathématiques, BP47, 38040 Grenoble Cedex 9, France

ARTICLE INFO

Article history:
Received 24 June 2013
Received in revised form 23 January 2014
Accepted 27 January 2014
Available online 12 March 2014

PsycInfo-codes:
2340
2343
2346

Keywords:
Word segmentation
Chunking
Modeling
Artificial language

ABSTRACT

There is large evidence that infants are able to exploit statistical cues to discover the words of their language. However, how they proceed to do so is the object of enduring debates. The prevalent position is that words are extracted from the prior computation of statistics, in particular the transitional probabilities between syllables. As an alternative, chunk-based models posit that the sensitivity to statistics results from other processes, whereby many potential chunks are considered as candidate words, then selected as a function of their relevance. These two classes of models have proven to be difficult to dissociate. We propose here a procedure, which leads to contrasted predictions regarding the influence of a first language, L1, on the segmentation of a second language, L2. Simulations run with PARSER (Perruchet & Vinter, 1998), a chunk-based model, predict that when the words of L1 become word-external transitions of L2, learning of L2 should be depleted until reaching below chance level, at least before extensive exposure to L2 reverses the effect. In the same condition, a transitional-probability based model predicts above-chance performance whatever the duration of exposure to L2. PARSER's predictions were confirmed by experimental data: Performance on a two-alternative forced choice test between words and part-words from L2 was significantly below chance even though part-words were less cohesive in terms of transitional probabilities than words.

1. Introduction

Language acquisition initially proceeds from auditory input, and linguistic utterances usually consist of sentences linking several words without clear physical boundaries. The question thus arises: How do infants become able to segment a continuous speech stream into words? Recent psycholinguistic research has identified a number of potentially relevant factors. Analyses of natural languages have shown that a number of acoustical, prosodic, and statistical features are correlated with the presence of word boundaries, and could therefore be used as cues for segmenting the speech signal into words. There is large evidence that these cues are used at a various extent according to the age of the learners and the specific structure of the language (Thiessen & Saffran, 2003), and that they interact in complex ways (Creel, Tanenhaus, & Aslin, 2006; Onnis, Monaghan, Richmond, & Chater, 2005; Perruchet & Tillmann, 2010).

In this paper, we focus on statistical cues, such as they were revealed in the seminal studies by Saffran and collaborators. For instance, Saffran, Aslin, and Newport (1996) used an artificial language consisting of four trisyllabic words, such as *golatu* and *daropi*. In the familiarization phase, 8-month-old infants listened to a sequence of words, which were read by a speech synthesizer in random order in immediate succession, without pauses or any other prosodic cues. In the following test phase using a familiarization-preference procedure, the infants were presented with repetitions of either words or trisyllabic "part-words", such as *tudaro*, consisting of the final syllable of a word joined to the first two syllables of another word. Infants showed longer listening times for part-words, suggesting that they were perceived as novel sequences. This and other studies (e.g., Aslin, Saffran, & Newport, 1998) are usually interpreted as indicating that infants exploit the transitional probabilities (TPs) between syllables, because word-internal TPs are stronger than TPs between the syllables that compose the part-words (i.e., containing word-external TPs).

1.1. Two competing hypotheses

The prevalent interpretation for this remarkable outcome is that participants perform statistical computations (e.g., Aslin et al., 1998; Endress & Mehler, 2009). The reasoning is straightforward: If learners' behavior turns out to be sensitive to a given statistical property of the input, then this implies that learners somehow compute the relevant statistics. Typically, learners are assumed to compute the TPs between successive syllables (Saffran, Newport, & Aslin, 1996). In a competing approach, the sensitivity to statistics is a mandatory consequence of the engagement of other cognitive processes. Instead of inferring the words from the prior computation of TPs, the general strategy shared

* Corresponding author at: Université de Bourgogne, LEAD/CNRS, Pôle AAFE, 11 Esplanade Erasme, 21000 Dijon, France.
E-mail address: pierre.perruchet@u-bourgogne.fr (P. Perruchet).

http://dx.doi.org/10.1016/j.actpsy.2014.01.015
0001-6918/

by all chunk-based models is that many potential chunks are created, then selected as a function of their relevance (e.g., Brent & Cartwright, 1996; Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Perruchet & Vinter, 1998; Robinet, Lemaire, & Gordon, 2011; Servan-Schreiber & Anderson, 1990).

Although relying on very different processes, the two accounts coined hereafter as the *TP-based* approach and the *chunk-based* approach, respectively, appear to be surprisingly difficult to dissociate. We propose below an experimental design leading to contrast the predictions of these two approaches. Before introducing to this new design, however, a finer description of the chunk-based model that will be considered here, namely PARSER (Perruchet & Vinter, 1998), is in order.

1.2. PARSER model

Let us consider the famous Saffran, Newport et al. (1996) study in which six trisyllabic words, *babupu*, *bupada*, *dutaba*, *patubi*, *pidabu*, and *tutibu*, were repeated in random order. The speech flow may begin as:

(1) *babupututibubabupudutabapatubibupadapatubidutababababu pupidabu…*

PARSER postulates that (1) will be perceived as, for example:

(2) *babu putu ti buba bupudu ta bapa tubi bupada pa tubi duta bababu pupi dabu…*

where spaces stand for subjective boundaries. These boundaries are introduced as a consequence of attentional mechanisms, which naturally segment the sensory input into small disjunctive parts of various lengths. [1] The randomly determined fragments are created as provisional chunks as they appear in the language. Clearly, a few of them are relevant to the structure of the language (*bupada* is a word, and *babu*, *tubi*, and *duta* are components of words) and others are irrelevant. How does the model operate a selection without calling to sophisticated computations? In PARSER, the fate of a new chunk depends on the probability for this new chunk to be encountered later. The relevant units emerge through a selection process based on forgetting, which leads to eliminate the less cohesive parts among all parts generated by the initial chunking of the material. For instance, *bababu* is doomed to forgetting, because it will reoccur only when *dutaba* is followed by *babupu*. By contrast, *babu* and *bupada* have more chance of resisting to forgetting because they will be strengthened on each occurrence of *babupu* and *bupada* respectively, whatever the surrounding words.

Forgetting, in PARSER, is the end-product of both decay and interference. If forgetting was only due to decay, PARSER would be only sensitive to the raw frequency: The candidate units resisting to forgetting would be those that occur the most frequently in the speech flow. Interference allows the model to be sensitive to more sophisticated measures of contingency. To illustrate, *putu*, which straddles a word boundary, has been processed as a unit in (2). The weight of this unit will be decreased each time another interfering unit will be perceived. This is the case with the units *bupudu* and *pupi* in (2), because *pu* is present and followed by another syllable as *tu*. The resulting effect is nothing else here than the classical effect of retroactive interference, whereby learning AC has a more detrimental influence on the retrieval of a previously learned pair AB than learning a list of unrelated items (e.g., DE). It is clear that, overall, *putu* will receive more interference than a within-word component, given that *pu*, as a final syllable of a word, may be followed by several different syllables. This example illustrates that increasing the sources of interference and decreasing TPs are two sides of the same coin, because both result from an increased number of possible adjacent events (Perruchet & Poulin-Charronnat, 2012a). As a consequence, implementing interference as a mechanism of chunk selection in PARSER makes the model responsive to TPs.

Crucially, once a new chunk has been created on the basis of its internal consistency, it plays the role of a new primitive, which constrains the coding of the incoming information as did the initial primitives (i.e., the syllables). For instance, once *bupada* has been built as a perceptual primitive for the model, the following percept necessarily begins with the following word, hence increasing the probability of discovering this word (i.e., *patubi* from (1)). In this way, PARSER naturally accounts for the fact that known words help to discover new words (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005; Dahan & Brent, 1999), as analyzed in Perruchet and Tillmann (2010).

1.3. The present study

The rationale of the present study directly follows from the principle stated just above. We noted that knowing *bupada* helps to discover *patubi* when exposed to *bupadapatubi*. However, a more general claim is that the probability of creating a new unit depends on the units already present in the lexicon, *whether relevant or not*. If *dapa* has been created, this will trigger the formation of chunks such as *bupa* or *tubi*, which are not words, when exposed to *bupadapatubi*. More generally, if a wrong unit has been created, this will trigger the formation of other wrong units. This happens only rarely in natural settings, given that decay and interference tend to select the relevant units (i.e., the words) of the language. But the phenomenon can be artificially induced in controlled conditions through the prior presentation of irrelevant units. This offers the opportunity of manipulations leading to predict opposite effects in a chunk-based framework and in a TP-based framework, which does not exploit such a principle.

In keeping with this strategy, the present study examines how familiarization with a first language, L1, affects the segmentation of a second language, L2. In the following experiment, L2 was composed of three trisyllabic words, ABC, DEF, and GHI (each letter stands for a syllable), which were randomly concatenated without immediate repetition. L1 was composed of bisyllabic words, which were played as isolated utterances. In the main experimental condition (the *overlapping* condition), the words of L1 reoccurred as word-external transitions in L2 (e.g., *CD* occurred in L2 when *ABC* was followed by *DEF*). In a *control* condition, the pairs of events composing the words of L1 never occurred in L2 (e.g., *CA* could not occur, because repetitions of words were not allowed). In a subsequent two-alternative forced choice (2AFC) test, participants were exposed to pairs composed of a word and a part-word of L2 (see Table 1). For each pair, participants had to decide which item seemed more like a word of the imaginary language they were exposed to before.

The underlying intuition was that a TP-based approach should predict above-chance performance in the 2AFC test, whatever L1. This is because, as shown in Table 1, none of the pairs of syllables played in L1 was included in the test items, whether words or part-words, and this was true for both the overlapping and control conditions. PARSER should also predict above-chance performance in the control condition. Indeed, because the chunks built from L1 are no longer present in L2, they will be progressively forgotten, and learners have only to build new chunks from L2. However, crucially, L1 chunks continue to be perceived during L2 presentation in the overlapping condition, and because they are between-word transitions in this new context, they could misguide the segmentation of L2, as explained above. As a consequence, the score in the overlapping condition should be lower than the score in

[1] Certainly the subjective experience of the beginning listeners would be rather the perception of a continuous and unintelligible speech flow, from which a sequence of a few syllables pops out from time to time. This does not change the rationale of the model. Simulations have shown that PARSER was able to reproduce the performance of actual participants while processing only 3 to 5% of the syllables of the languages (Perruchet & Vinter, 1998, Study 2). For instance, only *putu* or any other bisyllabic items may have popped out from Sequence (1), without hampering the ability of the model to account for human performance.

Table 1
Design for the simulations and the experiment.

L1		L2	Part-words in test
Overlapping	Control		
			CGH
CD	CA	ABC	FAB
FG	FD	DEF	IDE
IA	IG	GHI	BCG
			EFA
			HID

Note. Participants were familiarized with a first language (L1), in which the words were played in isolation. In a second phase, participants were exposed to a continuous speech stream (L2), which differed from L1 to various extents (Overlapping = the words of L1 were between-word transitions in L2; Control = L1 and L2 shared no pair of syllables). The final phase was a 2AFC test contrasting the words from L2 to part-words. The letters are used as placeholders for randomized syllable instantiations.

the control condition, and potentially below chance. The next section presents computational simulations, which lead to complete and refine these speculative predictions.

2. Simulations

2.1. A TP-based model

From Saffran, Newport et al. (1996), and in most subsequent studies from the same laboratory, predictions of participants' performance were based on a comparison of the TPs between syllables composing the words and the part-words, such as inferred from the composition of the language. For instance, in a language comprising six words of equal frequency, the TP between a terminal syllable of a word and the initial syllable of another word is said to be .20 because there are five possible successors (word repetitions are not allowed), and this value is invariant whatever the duration of the language.

In our study, the succession of two languages involving (partially) the same syllables makes the theoretical assessment of the TPs for L2 a little more complex. Indeed, at least some of the TPs for L2 are affected by the structure of L1, but to a various extent depending on the length of L2. To consider the endpoints, at the very beginning of L2, the TPs are those of L1, while the influence of L1 over the TP pattern of L2 approaches nullity as L2 goes to an infinite length. The concrete values for a given experiment are in-between. In the following calculations, the number of repetitions of the words [2] from L1 was set to a fixed value, which was selected to be well-suited for the subsequent experiment (each word was played 28 times, which generates two minutes of oral production). For L2, the number of item repetitions was varied along a large range, roughly following an exponential function (each word occurred 16, 24, 36, 54, 80, 120, 180, or 270 times). The pairwise TPs were computed from frequency counts performed on actual instances of languages as indicated in the following standard equation. For a pair xy:

$$p(y|x) = p(xy)/p(x) \approx \text{freq}(xy)/\text{freq}(x).$$

Because the TPs for the part-words may slightly vary as a function of word order, the values were averaged over 100 randomly generated languages.

For each trisyllabic test item, XYZ, a value was computed as the mean of TPs for the pairs XY, YZ, and X_Z. Note that Saffran et al. considered only the first two pairs. We added X_Z, as did Endress and Mehler (2009), in keeping with recent studies showing that distant dependencies can be learned under some conditions (e.g., Gomez, 2002; for a brief review, see Perruchet, Poulin-Charronnat, & Pacton, 2012). Without L1, the theoretical values should be 1 for the words of L2, and .66 for the mean TPs of part-words (e.g., for the part-word CGH, $p(G|C) = .5$, $p(H|G) = 1$, and $p(H|C) = .5$). The question is how the prior exposure to L1 affects these values, and in particular, whether the influence of L1 differs as a function of conditions (overlapping vs. control).

Fig. 1 reports the mean TPs for words and part-words for each of the two conditions and for each length of L2. Only two curves are apparent out of the expected four ones, because the values for the overlapping and the control conditions were almost perfectly superimposed for each length of L2. The TPs for both words and part-words asymptotically converged towards their theoretical values (1 and .66 for words and part-words, respectively), but they started from lower values. Importantly, for each length of L2, the TPs for words were substantially higher than the TPs for part-words (the two curves evolved roughly in parallel; all $ps < .001$).

As in studies by Saffran and colleagues, the quantitative assessment of TPs was not devised to be translated into a quantitative score in a 2AFC test. However, the pattern of TPs is straightforward enough to allow for clear-cut qualitative predictions. The underlying reasoning is that, to quote Endress and Mehler (2009), "participants are more familiar with items with stronger TPs than with items with weaker TPs" (p. 352), and therefore the participants' choice in the 2AFC test should be guided by the mean TP for each item of the pair (see also Frank, Goldwater, Griffiths and Tenenbaum, 2010; Kurumada, Meylan and Frank, 2013). In keeping with this reasoning, two predictions follow. First, no difference is expected between the overlapping and the control group. Second, the pattern of TPs should ensure the selection of words over part-words and hence above-chance performance in the 2AFC test in all cases, even after the shortest duration of training.

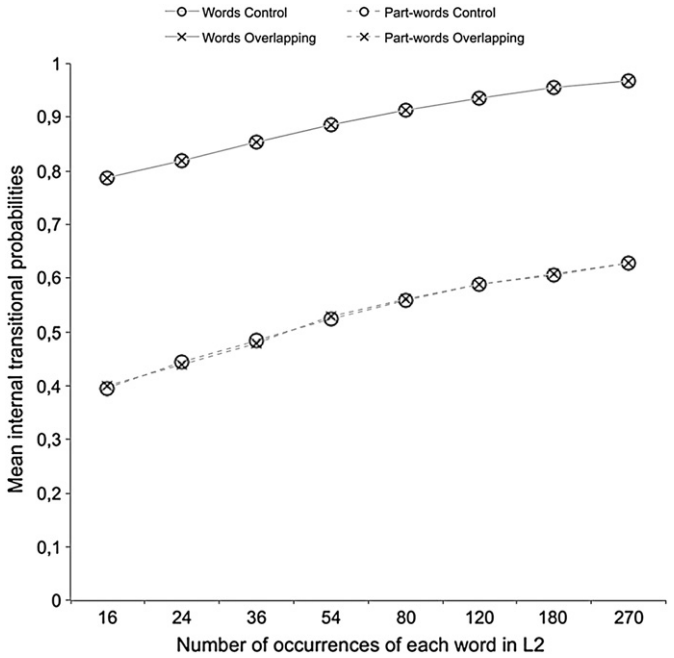

Fig. 1. Mean internal TPs for the "words" (full line) and the "part-words" (dotted line) described in Table 1, as a function of the length of L2. Data for the overlapping condition and the control condition are almost exactly superimposed. For the words, the TPs computed over 100 languages did not vary at all (all languages comprised the same number of words), and for the part-words, the variations due to the specific word order were too small to be represented (all standard errors <.004).

[2] Artificial languages were continuous in an overwhelming proportion of earlier studies, so that there is no consensus about the best way to code the spaces between the words of L1. In the data reported in Fig. 1, the space between words was coded as a pseudo-syllable. Kurumada, Meylan, and Frank (2013) did not proceed this way, and made counts only within each utterance. When calculated as in Kurumada et al., all the values reported in Fig. 1 were higher, but the conclusions remain strictly identical.

2.2. PARSER

The very same languages were entered into PARSER.[3] L1 was entered as individual words, a procedure that prevents the formation of provisional chunks that would span over successive words. L2 was shown as a continuous sequence of words.

A point of debate in computational research is the selection of parameter values. In PARSER, the rates of decay and interference (and a few other, minor parameters) may be tuned to comply with the materials or the study population. However, the general strategy adopted in earlier studies (e.g., Frank et al., 2010; Giroux & Rey, 2009; Kurumada et al., 2013; 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 turns out to be appropriate for studies with other objectives. Because we were not interested in maximizing the quality of fit, but instead in examining whether some crucial predictions were actually constitutive of the model, all the subsequent simulations have been performed with these standard parameters.

As for the TP-based model, the length of L2 was varied, with each word occurring 16, 24, 36, 54, 80, 120, 180, or 270 times. Fig. 2 reports the probability for a word and a part-word to be found by PARSER (i.e., to be in the model's lexicon at the end of familiarization) averaged over 100 runs for each condition, for each length of L2. For the control condition (full dots), words are much more likely to be in PARSER's lexicon than part-words. [4] This ensures that words will be selected much more often than part-words in a 2AFC test. [5] By contrast, for the overlapping condition (empty dots), the probability is in fact higher for a test part-word than for a word to be in PARSER's lexicon (except for the longest exposure to L2; the difference reached significance, $ps < .05$, for L2 comprising 24, 36, and 80 occurrences of each word). Two predictions follow. First, PARSER predicts lower performances in the overlapping condition than in the control condition. Second, given the relative scores for words and part-words in the overlapping condition, the score in a 2AFC test should be *below chance-level* in this condition. As anticipated, these predictions are opposite to the predictions of a TP-based model.

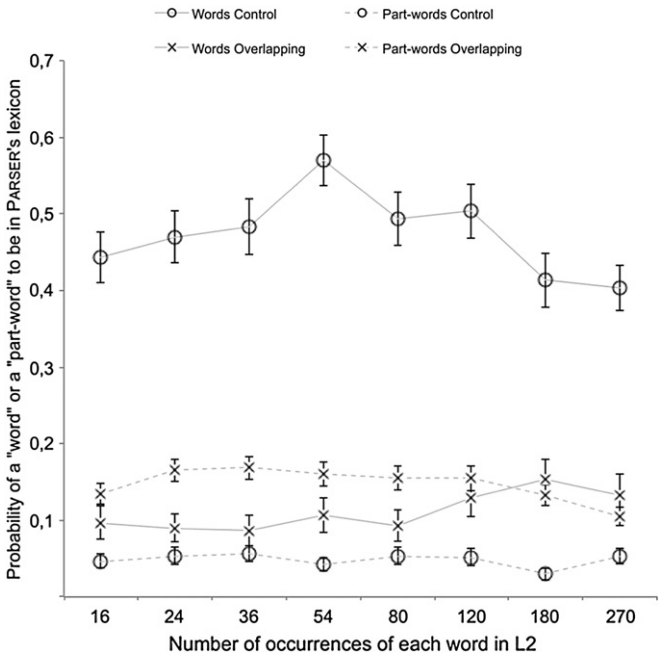

Fig. 2. Probability for a "word" (full line) and a "part-word" (dotted line) from Table 1 to be in PARSER's lexicon in the control condition (circles) and in the overlapping condition (cross), as a function of the length of L2. Error bars represent the standard error of the mean.

3. Experiment

In the present experiment, L1 was played as a succession of separate words. In a recent study devised to examine the influence of prior knowledge on the subsequent segmentation of a continuous speech stream, Lew-Williams and Saffran (2012) played a list of words separated by a short pause during the pre-exposure phase. We used a similar strategy to be sure that the words composing L1, and only the words, were learned. L2 was played as a continuous stream of syllables.

3.1. Method

3.1.1. Participants

Forty-four undergraduate students from the University of Bourgogne took part in the experiment for partial fulfillment of a course requirement. They were randomly assigned to one of the two conditions with 22 participants per condition.

3.1.2. Material

Two languages, L1 and L2, were prepared for each participant. L1 differed according to the conditions (overlapping vs. control). L2 was common to all participants, and was obtained by randomly concatenating the words *ABC*, *DEF*, and *GHI* (the letters are used as placeholders for randomized syllable instantiations), except that the same word never occurred twice in immediate succession. The CV syllables /*bi*/, /*kĩ*/, /*do*/, /*te*/, /*dã*/, /*pa*/, /*tu*/, /*gy*/, and/*põ*/ were ascribed to the letters used in Table 1. Participants in the overlapping condition and in the control condition were yoked, in such a way that the syllable/letter matching was the same for each pair of participants, to ensure that any difference between conditions cannot be attributed to an a priori preference for specific items. However, the syllable/letter matching differed for each pair of yoked participants, to prevent the influence of irrelevant factors (e.g., phonological preferences, similarity with words of the natural language).

The speech was synthesized using the MBROLA speech synthesizer (http://tcts.fpms.ac.be/synthesis/; Dutoit, Pagel, Pierret, Bataille, & Van Der Vrecken, 1996) with the FR2 diphone database. The mean syllable

[3] All the simulations reported in this paper were run under U-learn (Perruchet, Robinet, & Lemaire, submitted for publication), which is freely available at the following URL: http://leadserv.u-bourgogne.fr/~perruchet/. The reported results can be exactly reproduced by setting "Current run" as the random seed.

[4] In the control condition, the number of words in the lexicon (slightly) increased between 16 and 54 repetitions, but, surprisingly decreased with further training. This indicates that some words, once acquired, disappear from PARSER's lexicon with further practice. In fact, the phenomenon occurs for a straightforward reason: PARSER is a model of unsupervised learning, and there is no internal controller to tell the chunking process to stop once a word has been discovered. For instance, if *ABC* and *DEF* frequently occur in succession in the language, this may lead to the creation of the *ABCDEF* unit, which in turn concurs to eliminate *ABC* and *DEF* from the lexicon. In most cases, the creation of multiword units does not happen because they are not cohesive enough (and in natural language, words have distinct referents, which help to maintain them as distinct units). However, the small number of words in the present study (and the fact they have no referent) makes possible the creation of units embedding several words. Increasing the forgetting rate in PARSER would have prevented, or at least reduced such an undesirable effect.

[5] The chance of finding a part-word when drawing at random a 3-syllable unit from a language composed of trisyllabic words is higher than the chance of finding a word. However, the score displayed in Fig. 2 is the probability of being in PARSER's lexicon for *a given test part-word*, not for all the test part-words listed in Table 1, and even less for all potential part-words in L2. By the same token, this score is independent of the number of test part-words (the fact that there is twice more test part-words than words is inconsequential). Given that the 2AFC test contrasts a given word to a given part-word, the reported probabilities can be directly translated into a score of discrimination: A higher probability for the words than for the part-words to be in PARSER's lexicon entails that the score in a 2AFC test should be above chance-level, and likewise, a higher probability for the part-words than for the words to be in PARSER's lexicon entails that the score in a 2AFC test should be below chance-level.

duration was 232 ms. Progressive fade-in and fade-out were applied to the first and last 5 s of L2. Both languages, as well as the test items, were played through headphones connected to a PC computer.

3.1.3. Procedure

Participants were told that they would have to listen to a few words of an imaginary language. Each of the three words of L1 was repeated 28 times. The words were separated by short pauses, which varied randomly in duration from 500 ms to 1500 ms. The total duration of L1, including the pauses, was approximately 2 min.

Then participants were told that they would have to listen to a sample of another imaginary language. The motivation for presenting L2 as another language was to prevent strategies consisting in the explicit search of L1 words, or more generally the search of bisyllabic items, while listening to L2. Participants were asked to avoid engaging in analytic, problem-solving activities. They were exposed to L2 as a continuous speech stream. The words were played in random order for about 6 min, each word occurring 180 times. As an aside, this number of repetitions exceeds the values for which PARSER predicts below-chance performance in the overlapping condition, as shown in Fig. 1. The paradox is only apparent, however, because it has been repeatedly observed that PARSER tends to outperform human participants when trained with the same corpus (e.g., Perruchet & Vinter, 1998). Using a longer corpus for human participants is thus theoretically motivated, and would make below-chance scores still more remarkable if they are obtained in these conditions.

After exposure to the speech stream, participants were told that they would be presented with pairs of items, and that they would have to judge, for each pair, which item seemed more like a word of the imaginary language they were exposed to before. There were 36 pairs of items, composed of the repetition of 18 different pairs of items. Among the 18 pairs, 12 pairs comprised one word and one part-word. Each of the three words was crossed with four different part-words, selected among the six part-words shown in Table 1. This resulted in a difference of frequency between words and part-words because each word occurred four times, while each part-word occurred only twice. This may be a source of bias, because participants may select the words on the basis of their familiarity induced during the test. To prevent this bias, six additional pairs were composed of two part-words, in such a way that overall, each word and each part-word were presented equally often (i.e., four times). Each pair of two part-words contrasted one 3-1-2 item (i.e., an item composed from the last syllable of a word and the first two syllables of another word; the first three test items in Table 1 are 3-1-2) and one 2-3-1 item (i.e., an item composed from the last two syllables of a word and the first syllable of another word; the last three test items in Table 1 are 2-3-1). The members of each pair were separated by a 500-ms interval. The order of the items within a pair, as well as the order of the pairs in the test sequence, was randomized, with a different randomization for each participant.

3.2. Results

Although we had no specific expectation about the test pairs contrasting two part-words, the data were analyzed. There was no significant difference between the overlapping and the control groups, $t(42) = 1.21$, $p = .232$, and overall the 3-1-2 and 2-3-1 part-words were selected equally often (49.24% vs. 50.76% respectively, $t(43) = 0.268$, $p = .79$). The following analyses concern the word/part-word pairs.

The rate of correct responses is shown in Fig. 3 as a function of condition. Overall, the pattern of results confirmed PARSER's main predictions. The mean performance for the overlapping group was lower than for the control group, $F(1, 42) = 7.397$, $p = .009$, $\eta 2; p = .15$.

Crucially, the mean score of the overlapping group was significantly below chance, $M = 44.13$, $SD = 13.22$, $t(21) = 2.083$, $p = .049$, $d = 0.44$. To examine whether this score was in the range predicted

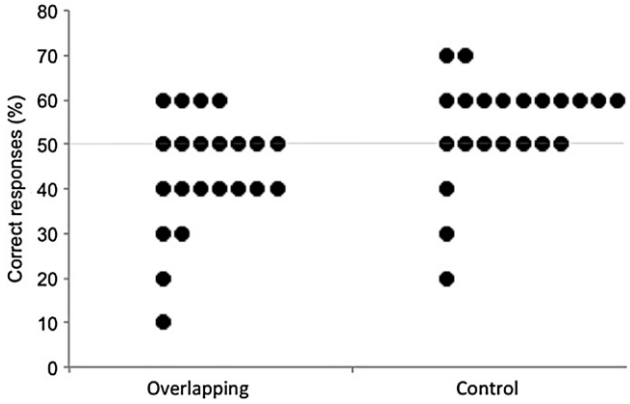

Fig. 3. Percentage of correct responses as a function of whether the probabilistic structure of L1 and L2 partially overlapped (overlapping) or not (control). Each dot represents a participant, and the corresponding value on the Y-axis indicates the center of the class interval in which the score of this participant falls (e.g., 40 stands for the [35, 45[interval). The gray line indicates chance level.

by PARSER, a score in the 2AFC test was computed from the simulated data reported in Fig. 2. As in Perruchet and Poulin-Charronnat (2012b), a response was generated for each pair of test items, as a function of the representation of the word and the part-word in the internal lexicon of the model. If neither the word nor the part-word was present, then the score was set to 50. If only the word was present, the score was set to 100, and if only the part-word was present, the score was set to 0. If both the word and part-word were in the internal lexicon of the model, the response was set to either 100 or 0 as a function of whether the weight of the word was higher or lower (respectively) than the weight of the part-word. A mean score was computed for each run, and data were averaged over 100 runs for each length of L2. The mean scores for the overlapping condition were 45.75, 45.75, 46.80, and 46.92 for L2 comprising 24, 36, 54, and 80 occurrences of each word, respectively. These values are close from the observed performance, despite the fact that no parameter adjustment was brought out to improve the quality of fitting.

Participants' scores for the control condition were only marginally above chance, $M = 54.36$, $SD = 11.68$, $t(21) = 1.749$, $p = .094$. Neither the TP-based model, nor PARSER was predictive of a so low score. Indeed, the TPs for words were substantially higher than the TPs for part-words even after the shortest exposure to the language (around .8 and .4 respectively, see Fig. 1). Likewise, using the method described above to infer a score in the 2AFC task from the items present in PARSER's lexicon (and reported in Fig. 2), PARSER produced more than 70% correct responses in the control condition after the shortest training duration. A possibility is that the occurrence of pairs of part-words in the 2AFC test had a detrimental influence on the responses to other word/part-word pairs, due to the fact that no correct response could be provided.

Another, nonexclusive possibility is that learning L2 was made difficult due to the prior presentation of another language (partially) sharing the same syllables. Earlier studies in which participants heard in succession two artificial speech flows differing in their statistical structure suggest that such a factor may have played a substantial influence. Gebhart, Aslin, and Newport (2009) showed that, when there was no explicit cue to identify the shift in languages, only the first of the two structures was learned. However, participants are able to learn two languages without any deficit when each language is individuated by strong contextual cues, such as talker's voice (Gebhart et al., 2009; Weiss, Gerfen, & Mitchel, 2009). Our paradigm lies somewhere in between these two extremes: L1 and L2 were produced by the same synthetic voice, but the shift between the two languages was clearly marked, and participants were informed that they will hear another imaginary language at the outset of L2. In this regard, our situation is

close to that of Franco, Cleeremans, and Destrebecqz (2011). Under conditions similar to ours, Franco et al. showed that L2 was learned, but to a lesser extent than when the same language was played in isolation.

Besides the general consequences of hearing two languages in immediate succession, the low score of participants in the control condition may also be due to specific properties of L1. Given that L1 was segmented into words, it provided some information about positional information (e.g., which syllable may begin a word). Participants may have exploited this kind of knowledge when running the 2AFC test. Now, as may be seen in Table 1, the test part-words either began or ended as the words of L1, whereas this was not the case for the test words. This may have favored the selection of part-words.

Whatever the explanation given for the low scores of participants in the control group, it is worth stressing that it leaves our main conclusions unchallenged. The interpretation of the effects observed in the overlapping group remains intact. Even the conclusions issued from a comparison between the overlapping and the control groups are not affected, because the factors evoked above as potentially detrimental for the learning of L2 were identical across the two groups. Notably, the words from L1 shared the same individual syllables at the same position in the overlapping and control conditions, and therefore the possible exploitation of positional information is unable to account for the observed difference in the two conditions.

4. Discussion

Although relying on very different processes, the *TP-based* approach and the *chunk-based* approach to word segmentation appear to be surprisingly difficult to dissociate. We propose here a procedure testing the influence of a first language, L1, on the segmentation of a second language, L2. When L1 and L2 overlapped in such a way that the words of L1 became word-external transitions of L2, predictions were opposite according to whether learning is thought of as mediated by the computation of TPs or by the creation of chunks. The direct computation of TPs predicted a preference for the words over the part-words of L2 in a 2AFC test, with predicted scores being identical to the scores of a control group, in which L1 was unrelated to L2. Simulations run with PARSER (Perruchet & Vinter, 1998), a chunk-based model, also predicted a preference for words over part-words in the control group. However, when L1 overlapped with L2, the segmentation of L2 was found to be impaired until reaching below-chance level, at least before extensive exposure to L2 reverses the effect. Overall, the results obtained in the present experiment with adult participants confirmed PARSER's predictions.

A pilot study based on a similar experimental design was presented at a workshop on implicit learning (Perruchet, 2004), and Franco and Destrebecqz (2012) explored a variant of the method. They reported results that do not confirm PARSER's predictions. If any, their data fit better with the predictions of a TP-based approach. The evidence provided by Franco and Destrebecqz can hardly be construed as a challenge for our conclusions, however, mainly because they did not use artificial languages in their experiments. The authors used a sequence of nonlinguistic visual targets that could appear at one of several possible positions on a touch-sensitive screen monitor. Participants were instructed to press the location of each target as fast as possible. Among many other consequences, the timing was considerably slowed down with regard to artificial language studies, with the "words" in Franco and Destrebecqz (i.e., a sequence of three to-be-responded visual stimuli) lasting more than 2 s on average. Moreover, while the duration of words is constant for a given experiment in an overwhelming proportion of artificial language studies, the duration of "words" in the Franco and Destrebecqz's Serial Reaction Time (SRT) task varied, given that it included participants' RTs on each of the stimulus. Hoch, Tyler, and Tillmann (2013) observed that learning was better when the artificial language is composed of regular-length units rather than irregular-length units. Finally, responding to each event might have further impeded the formation of multi-event units in Franco and Destrebecqz's experiments. To conclude, Franco and Destrebecqz's results do not challenge our conclusions for the issue of word segmentation, but they raise an interesting question regarding the generality of the mechanisms of chunking over different domains. Further explorations are needed on this point.

The present study focused on PARSER. This does not entail that PARSER is the only chunk-based model of segmentation that is able to account for the data we report in our experiment. We run pilot simulations with a model relying on Minimum Description Length (MDL) principles, the MDLChunker of Robinet et al. (2011, see Footnote 3). Overall, the predictions of the MDLChunker concerning the better performance in the control condition were similar to those of PARSER. The MDLChunker also predicted below-chance scores for the overlapping group, although this prediction was limited to the very early stage of familiarization with L2. The predictions of other MDL-based models (e.g., Frank et al., 2010; Orbán, Fiser, Aslin, & Lengyel, 2008), and the predictions of a recent chunk-based connectionist model (French, Addyman, & Mareschal, 2011) should be also examined through systematic and large-scale simulations. Pending further investigations, a conservative conclusion could be that our results do not reveal specific advantages of PARSER over all other models, but more generally the ability of chunk-based models to account for the data.

The question that remains to be considered is now: To what extent are our data actually inconsistent with the predictions of a TP-based model? It could be objected, for instance, that our design was inappropriate to reveal an effect of L1 on the processing of L2. After hearing L1, participants in our experiment were instructed that they will have to listen to *another* imaginary language to prevent explicit strategies, and this could have led them to process the two languages along independent pathways. The literature on bilingualism, however, makes this hypothesis implausible. Many studies have reported cross-linguistic effects in bilinguals in both visual and auditory word recognition (Dijkstra, 2005; Kroll & Dussias, 2013, for reviews). The orthographic and phonological characteristics of the native language have been shown to influence visual word recognition in L2 (Ota, Hartsuiker, & Haywood, 2010; Wang, Koda, & Perfetti, 2004). Still more relevant for our concern, the similarity of the phonotactic structures between L1 and L2 affects L2 learnability (Ellis & Beaton, 1993) and the phonotactic structure of L1 contributes to continuous speech segmentation in L2 (Weber & Cutler, 2006). Also, artificial words extracted from an artificial speech flow have been shown to be quickly lexicalized (i.e., became able to influence natural language processing), even though the artificial speech flow was not construed as an excerpt of natural language by the listeners (e.g., Fernandes, Kolinsky, & Ventura, 2009).

Another objection would be that because the words composing L1 were shown in isolation, the learners did not engage the mechanisms at play when faced with a continuous speech flow. In particular, there would be no need to compute the TPs between the constituent syllables of L1, and hence, genuine statistical learning would start only with exposure to the continuous speech flow of L2. However, assuming that L1 was not influential, the scores of the participants in the overlapping condition should have been above chance and similar to the scores of the control group, given that participants heard exactly the same L2 in the two conditions. The data clearly show that this was not the case: the segmentation of L2 was affected by the nature of L1.

Another argument against our conclusions could be that calculating piecemeal TPs for the test items, as done above, is irrelevant, and that what needs to be considered is the closeness in the overall distribution of TPs for L1 and L2. A natural intuition would be that updating the TPs computed from L1 to fit with L2 will be all the easier as the TPs of L2 are closer to those of L1. Analyzing Table 1 indicates that the two distributions of TPs (L1, L2) are closer to each other in the overlapping condition than in the control condition. Indeed, for the overlapping condition, some of the TPs that were set to 1 in L1 (e.g., CD) become .50 in L2 (C can be followed by D and G). By contrast, for the control condition, all TPs change drastically, from 1 to 0 or from 0 to 1. It follows that

learning L2 should be easier for the participants trained in the overlapping condition (the TP distributions of L1 and L2 partially overlap) than for those trained in the control condition. Empirical data revealed the exact opposite, ruling out the objection.

A last possibility was suggested by a reviewer of an earlier version of this paper. The starting point consists in reversing the postulate above, and to posit that segmenting L2 will be all the easier as the TPs of L2 are *farther away* from those of L1. Far from being an ad-hoc amendment, this alternative postulate is reminiscent of a view that has a long and venerable history in the area of conditioning and associative learning, since Kamin (1969) and the importance allocated to the concept of "surprise". Related views have been developed in other research domains, such as the conflict monitoring theory (e.g., Botvinick, Braver, Barch, Carter, & Cohen, 2001). One interpretation is that learning depends on the amount of attention devoted to the materials, and the surprise provoked by an unexpected or conflicting event is undoubtedly more prone to capture learner's attention than a perfectly predictable event (e.g., Pearce & Hall, 1980). Given that, as analyzed above, the TP distribution for L2 is more surprising for the participants in the control group than for the other participants, this leads to predict better performance in the control condition than in the overlapping condition. This prediction fits well with our data. However, this line of reasoning remains unable to account for the below-chance score of the overlapping group. The amount of surprise or conflict is assumed to modulate the speed to which participants tune themselves to the probabilistic structure of L2. Even if one takes for granted that the low level of attention of participants faced with the overlapping L2 slows down this tuning, or even prevents it altogether, there is no way to explain that learners *reverse* the pattern of TPs present in L2.

To sum-up, we fail to see how the reported data could be encompassed within a pure TP-based approach. This does not mean that these data invalidate this approach, for at least two reasons. Firstly, the predictions of a TP-based model were drawn from direct TP computations. Simulations with Simple Recurrent Networks (SRNs) have been sometimes conceived as an alternative to TP computations to test the same framework (e.g., Christiansen, Allen, & Seidenberg, 1998). It looks as somewhat unlikely that an SRN, which basically computes statistics (Redington & Chater, 1998), achieves to predict that the processing of L1 could elicit the observed preference for the part-words over the words of L2 in the overlapping condition, given that (1) these test items share no pair of syllables with the words of L1 and (2) part-words are both less frequent and less cohesive in terms of word-internal TPs than are words. Although Franco and Destrebecqz (2012) did not perform actual simulations, they also claimed that an SRN would not predict this kind of data pattern, as we observed here. However, given the huge number of parameters that can be manipulated in SRN models, it cannot be asserted that the observed data pattern is definitely out of reach of this class of models without running extensive computational investigations.

Secondly, even if an SRN turns out to be unable to account for our data, this would not mean that statistical computations did not occur. Indeed, advocates of a TP-based view do not claim that statistical computations are the only mechanisms involved in word segmentation. It remains possible to argue that TPs are computed, but that their effects are overshadowed or reversed by the action of additional mechanisms, such as those involved in chunk-based models. The unique advantage of chunk-based models is that they directly account for the present result pattern without any ad-hoc adjustments. In this regard, the data reported in this paper strengthen other supporting evidence for chunk-based models in the domain of word segmentation (e.g., Frank et al., 2010; Giroux & Rey, 2009; Kurumada et al., 2013; Perruchet & Poulin-Charronnat, 2012b; Perruchet & Tillmann, 2010).

Acknowledgments

This work has been supported by the Centre National de la Recherche Scientifique (CNRS, UMR5022, UMR5105, and UMR5292), by the Université de Bourgogne, the Université de Lyon I, and the Université Pierre Mendès-France à Grenoble.

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.
Botvinick, M. M., Braver, T. S., Barch, D.M., Carter, C. S., & Cohen, J.D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652.
Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.
Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language & Cognitive Processes*, *13*, 221–268.
Creel, S.C., Tanenhaus, M. K., & Aslin, R. N. (2006). Consequences of lexical stress on learning an artificial lexicon. *Journal of Experimental Psychology: Learning*, *32*, 15–32.
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.
Dijkstra, A. (2005). Bilingual visual word recognition and lexical access. In J. F. Kroll, & A.M. B. De Groot (Eds.), *Handbook of bilingualism: psycholinguistic approaches* (pp. 178–201)Oxford: Oxford University Press.
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).
Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning*, *43*, 559–617.
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.
Fernandes, T., Kolinsky, R., & Ventura, P. (2009). The metamorphosis of the statistical segmentation output: Lexicalization during artificial language learning. *Cognition*, *112*, 349–366.
Franco, A., Cleeremans, A., & Destrebecqz, A. (2011). Statistical learning of two artificial languages presented successively: How conscious? *Frontiers in Psychology*, *229*, 1–12 (Article).
Franco, A., & Destrebecqz, A. (2012). Chunking or not chunking? How do we find words in artificial language learning? *Advances in Cognitive Psychology*, *8*, 144–154.
Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107–125.
French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*, 614–636.
Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing structures in mid-stream: Learning along the statistical garden path. *Cognitive Science*, *33*, 1087–1116.
Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, *33*, 260–272.
Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431–436.
Hoch, L., Tyler, M.D., & Tillmann, B. (2013). Regularity of unit length boosts statistical learning in verbal and nonverbal artificial languages. *Psychonomic Bulletin and Review*, *20*, 142–147.
Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B.A. Campbell, & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296)New York: Appleton-Century-Crofts.
Kroll, J. F., & Dussias, P. E. (2013). The comprehension of words and sentences in two languages. In T. Bhatia, & W. Ritchie (Eds.), *The handbook of bilingualism and multilingualism* (pp. 216–243) (2nd ed.). Malden, MA: Wiley-Blackwell Publishers.
Kurumada, C., Meylan, S., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*, 439–453.
Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition*, *122*, 241–246.
Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, *53*, 225–237.
Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 2745–2750.
Ota, M., Hartsuiker, R. J., & Haywood, S. L. (2010). Is a FAN always FUN? Phonological and orthographic effects in bilingual visual word recognition. *Language and Speech*, *53*, 383–403.
Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–552.
Perruchet, P. (2004). Is it possible to deny unconscious cognition? *Workshop on implicit learning, December 2, Dijon, France*.
Perruchet, P., & Peereman, R. (2004). The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics*, *17*, 97–119.
Perruchet, P., & Poulin-Charronnat, B. (2012). Word segmentation: Trading the (new, but poor) concept of statistical computation for the (old, but richer) associative approach. In P. Rebuschat, & J. N. Williams (Eds.), *Statistical learning and language acquisition* (pp. 119–143)Berlin: De Gruyter Mouton.
Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*, 807–818.

Perruchet, P., Poulin-Charronnat, B., & Pacton, S. (2012). Learning nonadjacent dependencies. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1941–1944) New-York: Springer-Verlag.

Perruchet, P., Robinet, V., & Lemaire, B. (2014). *U-Learn: Finding optimal coding units from unsegmented sequential databases.* (submitted for publication).

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.D., Galland, N., & Peereman, R. (2004). Learning nonadjacent 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.

Redington, M., & Chater, N. (1998). Connectionist and statistical approaches to language acquisition: A distributional perspective. *Language & Cognitive Processes*, *12*, 129–191.

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

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

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.

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

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.

Wang, M., Koda, K., & Perfetti, C. A. (2004). Language and writing systems are both important in learning to read: A reply to Yamada. *Cognition*, *93*, 133–137.

Weber, A., & Cutler, A. (2006). First-language phonotactics in second-language listening. *The Journal of the Acoustical Society of America*, *119*, 597–607.

Weiss, D. J., Gerfen, C., & Mitchel, A.D. (2009). Speech segmentation in a simulated bilingual environment: A challenge for statistical learning? *Language Learning and Development*, *5*, 30–49.