

## A role for backward transitional probabilities in word segmentation?

PIERRE PERRUCHET AND STÉPHANE DESAULTY  
LEAD/CNRS, University of Bourgogne, Dijon, France

A number of studies have shown that people exploit transitional probabilities between successive syllables to segment a stream of artificial continuous speech into words. It is often assumed that what is actually exploited are the forward transitional probabilities (given  $XY$ , the probability that  $X$  will be followed by  $Y$ ), even though the backward transitional probabilities (the probability that  $Y$  has been preceded by  $X$ ) were equally informative about word structure in the languages involved in those studies. In two experiments, we showed that participants were able to learn the words from an artificial speech stream when the only available cues were the backward transitional probabilities. Learning is as good under those conditions as when the only available cues are the forward transitional probabilities. Implications for some current models of word segmentation, particularly the simple recurrent networks and PARSER models, are discussed.

A question of major importance for language acquisition is, How do infants discover the words of their language, given that linguistic utterances usually consist of sequences of words without clear physical boundaries? The seminal studies by Saffran and collaborators (e.g., Saffran, Newport, & Aslin, 1996) have allowed a major advance with respect to this issue, by showing that infants, children, and adults may extract words from an artificial language in which the words have been concatenated without any phonological or prosodic markers. The authors argued that learners exploit the transitional probabilities (hereafter, TPs) between successive syllables. For instance, participants are able to extract the word *befoki* from a sequence such as . . . *dubefokita* . . . because, when computed on a sizable corpus, the TPs between word-internal syllables [ $p(\text{fo}|\text{be})$  and  $p(\text{ki}|\text{fo})$ ] are higher than the TPs between syllables spanning word boundaries [ $p(\text{be}|\text{du})$  and  $p(\text{ta}|\text{ki})$ ].

This hypothesis received strong support from a follow-up study of infants by Aslin, Saffran, and Newport (1998). In Aslin et al.'s ingenious design, some of the three-syllable words in the language were played more frequently than the others in the familiarization speech flow, in such a way that the three-syllable units (called *part-words*) straddling the boundaries between the more frequently occurring words occurred as often as the less frequent words. In a subsequent, frequency-balanced test, infants showed reliable discrimination between these two categories of items. Since words and part-words differed with regard to the TPs between their constituent syllables, Aslin et al. concluded that infants relied on TPs for discrimination. A number of subsequent studies have used the same design with infants (Thiessen & Saffran, 2007) and adults (e.g., Hunt & Aslin,

2001; Peña, Bonatti, Nespor, & Mehler, 2002), with similar results.

### Forward and Backward Transitional Probabilities

The transitional probability between two successive syllables  $X$  and  $Y$ , such as that instantiated above, is defined as

$$\text{probability of } Y|X = \frac{(\text{frequency of } XY)}{(\text{frequency of } X)}.$$

This is the probability that  $X$  will be followed by  $Y$ , and, therefore, this coefficient measures the strength of the forward relations within the  $XY$  pair. The forward transitional probability will be noted  $TP_{for}$ . However, the relations within the  $XY$  pair may also be assessed by the backward relationship between  $X$  and  $Y$ , which is the probability that  $Y$  was preceded by  $X$ . The backward transitional probability (hereafter,  $TP_{back}$ ) is defined as

$$\text{probability of } X|Y = \frac{(\text{frequency of } XY)}{(\text{frequency of } Y)}.$$

The two values can differ substantially, as illustrated by the relationships between  $q$  and  $u$  in written English:  $TP_{for}$  is 1 ( $q$  is always followed by  $u$ ), whereas  $TP_{back}$  is much lower ( $u$  may be preceded by many letters other than  $q$ ).<sup>1</sup>

It is worth noting that  $TP_{back}$  conveyed some information about word structure in the languages used in the above-mentioned studies. Considering, for instance, the material used in Saffran et al. (1996), the  $TP_{for}$  mean was .675 for word-internal pairs of syllables and .137 for word-external pairs,<sup>2</sup> which means that this information may have been used for the discovery of words, as Saffran

et al. claimed. However, the *TPback* means were .655 and .146, respectively, which means that this measure was almost as predictive of word boundaries as *TPfor*. More important, the same conclusion holds for Aslin et al.'s (1998) study, which is often considered as providing a compelling case for the role of *TPfor* despite the authors' warning about the potential influences of other possible measures of between-syllable co-occurrences, including *TPback*.<sup>3</sup> In Aslin et al., *TPfor* and *TPback* were identical (namely, 1 for word-internal pairs and .5 for word-external pairs), which means that learners may have exploited either one or the other measure, or a combination of both.

### Are Backward Transitional Probabilities (*TPback*) Learnable?

Talking about *backward* relations suggests that these statistics are aimed at predicting a past event—hence, measuring information devoid of any adaptive function. Indeed, this may be right on some occasions. For instance, in classical conditioning settings, it is unclear what can be gained from predicting the occurrence of the conditioned stimulus given the unconditioned stimulus; accordingly, the statistics used to describe the between-stimuli contingency are forward oriented (for instance, Delta P, which is the most common measure of contingency in this domain [see Shanks, 1995, for a review], assesses forward relationships). However, it makes sense that the tightness of the association between two events also depends on backward relationships in other situations, and there is at least some preliminary evidence that people may be sensitive to those relationships.

One of these situations is the acquisition of the syntax of artificial languages. Saffran (2001, 2002) examined the ability of adult learners and children to use predictive dependencies as cues to phrase structure. Importantly, the only perfect predictions between word categories involved backward relations, whereas, according to the subsequent analyses of this material by Thompson and Newport (2007, p. 5), the *TPfors* were relatively low. The participants' performance was better than that in a control situation in which *TPbacks* were not predictive, hence suggesting that the participants had been successful in exploiting this source of information. This evidence is somewhat indirect, however, given that these studies were not intended to investigate the role of the directionality of the relationships. Instead, *TPbacks* were chosen in order to implement a statistical pattern that was the opposite of that of the native language of the participants, in such a way that they could not project English structure onto the artificial language.

To the best of our knowledge, Perruchet and Peereeman (2004) is the only study in which the roles of *TPfor* and *TPback* were directly compared. The authors reported experimental data whose aim was to explore participants' sensitivity to a range of measures of association in the specific case of intrasyllabic relations. Specifically, they focused on the relationship between vowel (V) and consonant (C) in the terminal syllables of words. Focusing on VC terminal endings is justified by the fact that the analysis of linguistic corpora suggests that there are

strong probabilistic constraints on the VC combinations in rimes, at least in English (Kessler & Treiman, 1997) and in French (Peereeman, Dubois-Dunilac, Perruchet, & Content, 2004). Perruchet and Peereeman's experiment revealed that the various statistical measures of association were not equally good predictors of how much children and adults felt that different VC terminal endings sounded like French words. Surprisingly, *TPfor* [ $p(C|V)$ ] for a VC rime was the worse predictor among their sample of measures, whereas *TPback* [ $p(V|C)$ ] was a good predictor. To sum up, there is some evidence that *TPback* can be exploited at both supralexic (syntactic) and infralexic (rime) levels, hence suggesting that *TPback* could also be involved in lexicon formation. Perruchet and Peereeman's study even raises the intriguing possibility that participants may find it easier to exploit *TPback* than *TPfor*.

### The Present Study

The aim of the present study was to disentangle the role played by *TPfor* and *TPback* in word segmentation. In the two experiments reported below, participants were exposed to artificial languages in which the words could be discovered by exploiting transitional probabilities between syllables. Crucially, the relevant statistics were *TPfors* for one group of participants and *TPbacks* for another group. Our main questions were the following: Do participants learn to segment a continuous speech stream when the only available cues for segmentation are backward transitional probabilities, and if they do, what is the level of performance, in comparison with a situation in which the only available cues for segmentation are forward transitional probabilities?

## EXPERIMENT 1

Experiment 1 was intended to compare performances between two groups of participants who were exposed to words that were defined either by their forward or by their backward transitional probabilities. To do this, we used a standard paradigm of word segmentation, except that, in contrast with most recent studies, certain syllables were common to several words.

### Method

**Participants.** A total of 42 undergraduate students from the University of Bourgogne participated in the experiment in partial fulfillment of a course requirement. All the participants were native French speakers. The participants were randomly assigned to one of the two experimental groups (forward or backward,  $n = 21$  in each group).

**Materials.** The languages were composed of nine bisyllabic words, which differed as a function of groups. Table 1 shows the structure of the words for the two groups. In this table, each letter ( $A, B, \dots, H, I, X, Y, Z$ ) stands for a syllable. In the forward group, each word began with one of nine different syllables ( $A, B, \dots, I$ ), but the second syllable of each word ( $X, Y, Z$ ) was common to three words. As a consequence, the word-internal *TPfor* was 1, whereas its between-word counterpart was .11 (i.e., 1/9, given that all successions are possible). Considering now *TPback*, Table 1 shows that they were equal to .33, regardless of whether they were computed within or between words. Clearly, the only type of TP that could be exploited to discover the words was *TPfor*. The situation was reversed for the language played to the participants in the

**Table 1**  
**Schematic Presentation of the Materials Used in Experiment 1**

Group	Words	Transitional Probability			
		Forward		Backward	
		Within Words	Between Words	Within Words	Between Words
Forward	A X	1	.11	.33	.33
	B X				
	C X	Examples:			
	D Y	D→Y	Y→A	D←Y	Y←A
	E Y				
	F Y				
	G Z				
	H Z				
	I Z				
Backward	X A	.33	.33	1	.11
	X B				
	X C	Examples:			
	Y D	Y→D	D→X	Y←D	A←Y
	Y E				
	Y F				
	Z G				
	Z H				
	Z I				

Note—Each letter stands for a syllable. For the forward group, the statistical consistency of the words comes from the fact that the forward transitional probabilities are higher when they are computed within words than when computed between successive words. For the backward group, the statistical consistency of the words comes from their backward transitional probabilities.

backward group. All the words ended with a different syllable, but the first syllable of each word was common to three words. As a consequence, the information conveyed by *TPfor* and *TPback* was reversed. The only type of TP that could be exploited to discover the words was *TPback*.

The words were composed by mapping 12 syllables (*mã, bi, gy, tu, ne, kē, dā, do, te, pō, vo, pa*) to the letters in Table 1. The mappings were arbitrary, except that any set of words in which a word or a part-word sounded like a French word was eliminated (this selection was performed “by hand”). Different mappings were used for different participants. This ensured that, on average, any discrimination between words and part-words during the test had to be attributed to prior exposure to the language, rather than to pre-experimental perceptual biases.

The speech was synthesized through the MBROLA (Multiband Resynthesis Overlap Add) speech synthesizer ([tcts.fpms.ac.be/synthesis](http://tcts.fpms.ac.be/synthesis); Dutoit, Pagel, Pierret, Bataille, & Van der Vrecken, 1996) with the FR2 database. The mean syllable duration was 232 msec. The resulting WAV file was modified using CoolEdit. Progressive fade-ins and fade-outs were applied to the first and last 5 sec of the familiarization stream to avoid word boundary cues. The speech stream was played through headphones connected to a personal computer using CoolEdit.

**Procedure.** The participants were told that they would listen to an imaginary language. They were asked to avoid engaging in analytic, problem-solving processes. The phase of familiarization to the language lasted about 8 min. Each of the nine words occurred 115 times. The words were pseudorandomly ordered for each participant. After this phase, the 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. There were 18 pairs of bisyllabic items, with the 2 items of each pair being separated by a 500-msec silent interval. The first half of the test was composed of the nine words, paired with nine part-words (i.e., bigrams spanning word boundaries). With reference to the letters used in Table 1, the part-words were *XD, XE, XF, YG, YH, YI, ZA, ZB, and ZC* for the forward group and *AY, BY, CY, DZ,*

*EZ, FZ, GX, HX, and IX* for the backward group. The second half of the test was composed of the same items, although the specific word/part-word pairings were different. The order of the items within a pair and the order of the pairs in the test sequence were randomized for each participant.

## Results and Discussion

The percentage of correct responses was numerically higher in the backward group ( $M = 67.20\%$ ,  $SE = 5.59$ ) than in the forward group ( $M = 60.32\%$ ,  $SE = 5.15$ ), although the difference did not reach significance [ $t(40) = 0.905$ ,  $p = .37$ ]. The score was statistically above chance (50%) in the backward group [ $t(20) = 3.074$ ,  $p = .006$ ], whereas the difference with respect to chance fell just above the conventional significance threshold for the forward group [ $t(20) = 2.004$ ,  $p = .059$ ].

These results demonstrate that word segmentation is possible even though *TPfor* did not contribute any information about word boundaries. Indeed, in the backward group, learning clearly occurred in conditions in which the *TPfors* between the syllables composing the words were the same as the *TPfors* between syllables spanning word boundaries. Such results suggest that participants can exploit *TPback* to guide word discovery. However, there is an alternative interpretation. In the materials used in Experiment 1, words and part-words were not matched in frequency. Words appeared three times more often than did part-words. Aslin et al. (1998) demonstrated that it was possible to discriminate words and part-words when their frequency was equalized, but they rightly noted that this does not entail that frequency cannot be exploited when it conveys some information. It is possible that the participants in the backward group based their choice on the frequency of co-occurrence of syllables, rather than on *TPback*.

## EXPERIMENT 2

Experiment 2 was intended to compare performances between two groups of participants who were exposed to words that were defined either by their *TPfor* or by their *TPback*, as in Experiment 1, but after controlling for frequency. We borrowed the strategy initiated by Aslin et al. (1998). Some words composing the language of familiarization were more frequent than the others, and the final forced choice test contrasted the less frequent words to the part-words generated by the succession of the more frequent words.

### Method

**Participants.** A total of 62 undergraduate students from the University of Bourgogne and high school students in Dijon, France, participated in the experiment in partial fulfillment of a course requirement. All the participants were native French speakers. The participants were randomly assigned to one of the two experimental groups (forward or backward,  $n = 31$  in each group).

**Materials.** The language was composed as in Experiment 1, except that three words out of nine were played three times more often than the six others. Table 2 shows the structure of the words for the two groups. In the pairs played in the forced choice test, the less frequent words were contrasted with the part-words generated by the succession of the more frequent words. This means that, for the two groups, words and part-words were equalized in frequency.

**Table 2**  
**Schematic Presentation of the Materials Used in Experiment 2**

Group	Words	Relative Frequency	Transitional Probability			
			Forward		Backward	
			Within Words	Between Words	Within Words	Between Words
Forward	A X	3	1	.20	.20	.33
	B X	1	Examples: B→X    X→D    B←X    X←D			
	C X	1				
	D Y	3				
	E Y	1				
	F Y	1				
	G Z	3				
	H Z	1				
	I Z	1				
Backward	X A	3		.20	.33	1
	X B	1	Examples: X→B    A→Y    X←B    A←Y			
	X C	1				
	Y D	3				
	Y E	1				
	Y F	1				
	Z G	3				
	Z H	1				
	Z I	1				

Note—Each letter stands for a syllable, for the forward and backward groups. For each group, three words were three times more frequent than the six others. The between-word units that are considered in the table (and used in the forced choice test; e.g., *XD*) are formed by the succession of the most frequent words. They appeared as often as the less frequent words (e.g., *BX*).

The TPs reported in Table 2 refer to the items that were involved in the forced choice test. For the forward group, word-internal *TPfors* (1) were notably higher than the *TPfors* (.20) computed between the words involved in the test (e.g., *XD*). Backward relationships cannot be exploited to discover the words, given that the word-internal *TPback* was even slightly lower than the *TPback* between successive words. The situation was exactly reversed for the backward group. The only source of information that the participants in this group could exploit was provided by the *TPback*.

The words were composed as in Experiment 1, and they were played by the same speech synthesizer, with the same parameters.

**Procedure.** The participants were told that they would listen to an imaginary language. They were asked to avoid engaging in analytic, problem-solving processes. The phase of familiarization to the language lasted about 8 min. Each of the three more frequent words occurred 210 times, and each of the six remaining words occurred 70 times. The words were pseudorandomly ordered for each participant. After this phase, the 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. There were 36 pairs of bisyllabic items, which were obtained by the exhaustive combinations of the less frequent six words, with six part-words fulfilling the criteria above. Referring to the letters used in Table 2, the six part-words were *XA*, *XD*, *YD*, *YG*, *ZA*, and *ZG* for the forward group and *AX*, *AY*, *DY*, *DZ*, *GX*, and *GZ* for the backward group. The order of the items within a pair and the order of the pairs in the test sequence were randomized for each participant.

## Results and Discussion

In contrast to Experiment 1, the percentage of correct response performances was numerically higher for the forward group ( $M = 66.7\%$ ,  $SE = 4.32$ ) than for the backward group ( $M = 61.1\%$ ,  $SE = 5.11$ ), but the difference did not reach significance [ $t(60) = 0.831$ ,  $p = .41$ ]. The score was statistically above chance (50%) in the forward group [ $t(30) = 3.863$ ,  $p = .0006$ ] and in the backward group [ $t(30) = 2.176$ ,  $p = .038$ ].

These results confirm the commonly accepted contention that word segmentation is possible when frequency has been controlled and when the only information useful for learning is the *TPfor* between syllables. However, crucially, our experiment also demonstrates that participants learned the words when the only available information was the *TPback* between syllables. For those participants, the *TPfors* were even (slightly) misleading about word boundaries.

## GENERAL DISCUSSION

We started from the observation that *TPback* has been as informative about word structure as *TPfor* in the studies that have commonly been taken as demonstrative of the exploitation of *TPfor* in the segmentation of artificial languages (e.g., Aslin et al., 1998; Saffran et al., 1996). Given that there was some preliminary evidence that *TPback* can be exploited to capture the structure of sound sequences at both syntactic (e.g., Saffran, 2001, 2002) and sublexical (Perruchet & Peereman, 2004) levels, the aim of the present study was to examine the possible contribution of *TPback* to the extraction of words. The participants were exposed to artificial languages in which the words could be discovered through either the *TPfor* or the *TPback* between adjacent syllables. In Experiment 1, the words and the part-words were not matched in frequency, whereas frequency was controlled in Experiment 2. Two main results emerged.

First, participants can learn the words from a continuous language in which only *TPbacks* are indicative of word structure. In Experiment 2, learning occurred even in conditions in which the *TPfors* between the word-internal



syllables were slightly *lower* than the *TPfors* between the syllables spanning word boundaries.

Second, there was no clear advantage for a specific directionality. The participants also learned in conditions in which only *TPfors* were indicative of word structure, and even in conditions in which the *TPbacks* between the word-internal syllables were slightly lower than the *TPbacks* between the syllables spanning word boundaries. At a descriptive level, a better efficiency of *TPback* was observed in Experiment 1, but the trend was inverted in Experiment 2, and in the two experiments, no significant difference emerged between the participants that learned from *TPfor* and those that learned from *TPback*. Therefore, the hypothesis suggested by the Perruchet and Peereeman (2004) results that *TPback* may be a better cue than *TPfor* for sublexical units was not confirmed at the word level. The results suggest, rather, that *TPfor* and *TPback* can play a roughly symmetrical role when both of them are available.

### Implications for the Models of Word Segmentation

The implications of our results for the current models of word segmentation are straightforward: The ability of those models to account for people's sensitivity to backward relationships ought to be considered as an additional criterion of validity. Interestingly, the extant models largely differ in this regard. In one of the prevalent conceptions, word boundaries are defined as the points at which the consistency between successive syllables or subsyllabic phonological units is the lowest. At the computational level, the location of those points is generally inferred from the predictions of connectionist models—most often, simple recurrent networks (SRNs; e.g., Christiansen, Allen, & Seidenberg, 1998; Elman, 1990). An SRN is typically trained to predict the next element of a sequence. To do so, the network is presented, at each time step  $t$ , with the current element of the sequence and with a copy of its own internal state (i.e., the vector of hidden unit activations) at time step  $t-1$ . On the basis of these inputs, the network predicts which element should appear at time step  $t+1$ . The network's predictions are compared with the actual successor in the sequence, and the resulting error signal is then used to modify its connection weights, using the back-propagation algorithm. This brief description is sufficient to make it clear that it is inherent to the structure of an SRN to approximate *TPfors*. In an attempt to substantiate this analysis, Perruchet and Peereeman (2004) tested the ability of an SRN to find units that were defined by either the *TPfors* or the *TPbacks* between their constituents. SRN's predictions matched nearly perfectly the *TPfor*, so that the correlation between the two measures reached  $r = .99$ , whereas they did not correlate with the *TPback*. This insensitiveness of SRNs to *TPback* strongly calls into question the relevance of SRNs for the simulation of word segmentation learning. It is a challenge for further research to conceive of connectionist models responsive to the bidirectional relationships between the successive elements of sequences, while keeping a level of complexity compatible with considerations of biological or psychological plausibility.

To borrow the terminology proposed by Goodsitt, Morgan, and Kuhl (1993), connectionist models of word segmentation rely on a *bracketing* strategy, in which the primary aim of computations is to insert word boundaries within a continuous sequence. Other models rely on a *clustering*, or chunking, strategy, in which computations are primarily devised to group together elementary components into units. It turns out that extant clustering models are seemingly better suited to account for participants' sensitivity to *TPbacks*. For instance, Swingley (2005) proposed a model in which the probability that a given sequence of syllables will compose a new word depends on how much these syllables tend to appear together in a given corpus. As a measure of the level of co-occurrence, he used a coefficient known as mutual information. Like most other conventional measures of association (e.g., Pearson correlation and  $\chi^2$ ), mutual information is a bidirectional measure, which gives equal weight to forward and backward relationships. In this regard, Swingley's (2005) model of word segmentation could normally encompass the results presented above.

PARSER (Perruchet & Vinter, 1998) is a model of word segmentation that is based on a very similar intuition, except that chunk selection is no longer guided by direct statistical computations but is, instead, the by-product of elementary processes of associative learning. In PARSER, provisional chunks are initially formed on a random basis, as a natural consequence of the capacity-limited attentional processing of the incoming information. The selection of the most cohesive chunks is due to the action of ubiquitous memory process, such as strengthening with repetition and forgetting. Repetition and forgetting lead to the progressive emergence of the more frequent units among the randomly generated candidates, because only the most frequent candidate units resist forgetting. However, chunk selection also depends on *TPfors* and *TPbacks*. As developed in Perruchet and Peereeman (2004; see also Perruchet & Pacton, 2006, box 3), this is because forgetting depends on both decay and interference. For the sake of illustration, let us assume three bisyllabic units: *AB*, *AC*, and *DE*. *TPfor* is higher between *D* and *E* [ $p(E|D) = 1$ ] than between *A* and *B* or *A* and *C* [ $p(B|A) = p(C|A) = .5$ ]. If only decay processes were at play in the selection process, whether a unit resists forgetting or not would depend only on its relative frequency. However, it is in keeping with classical principles of associative learning and memory to posit that *AB* will interfere with *AC*, whereas *DE* has no interfering counterpart. As a consequence, assuming that their frequency of occurrence is identical, *DE* has more chance to resist forgetting than does *AB* and *AC*. Because in PARSER, in accordance with the traditional literature, interference also works in a backward direction between the elements of a pair (e.g., the reasoning above can be transposed when *AB*, *CB*, and *DE* are the initial units), the model is also responsive to *TPbacks*. For instance, the formation of the chunk *tudo* is hindered to the same extent by the presence of either *tubi* or *pado*; interference with *tubi* makes PARSER sensitive to the *TPfor*, whereas interference with *pado* makes it sensitive to the *TPback*.

To sum up, PARSER is sensitive to bidirectional measures of contingency such as mutual information, correlation, or  $\chi^2$  (for an empirical demonstration involving Pearson correlation, see Perruchet & Peereman, 2004). However, in contrast with Swingley's (2005) model, these statistics are never computed: The sensitivity to bidirectional contingency emerges online as a natural by-product of basic associative learning principles.

### Are Transitional Probabilities Useful After All?

For the moment, we have dealt only with artificial, miniature languages. The utility of statistical learning for word discovery in natural languages has on occasion been challenged. 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. However, Yang, in keeping with most authors in the field, only considered TP<sub>for</sub>. To what extent would the consideration of TP<sub>back</sub> change this assessment?

To the best of our knowledge, the only relevant investigation on this issue has been reported by Swingley (1999). Swingley (1999) examined how much TP<sub>for</sub>, TP<sub>back</sub>, and mutual information were informative about the word structure of a 50,000-word child-directed corpus. Although detailed results were not reported for TP<sub>for</sub> and TP<sub>back</sub> separately, the predictive power of TP<sub>for</sub> was seemingly within the same range as that found by Yang (2004). However, crucially, Swingley (1999) noted that "similar results were obtained using predictive and reverse transitional probability" (p. 726). In addition, he showed that removing the less frequent words from his corpus substantially improved performance on the whole corpus, for both TP<sub>for</sub> and TP<sub>back</sub>. When the corpus was restricted to words that occurred five or more times, and the criterion for deciding whether a transition was word internal or word external was set to TP = 0.4, completeness and precision were within the 40%–50% range for both TP<sub>for</sub> and TP<sub>back</sub> (as estimated from his Figure 7). Overall, when mutual information was used, which may be thought of as reflecting the cumulative effect of TP<sub>for</sub> and TP<sub>back</sub>, both precision and completeness rose above 60%. Obviously, these values, as high as they may be, are not evidence that infants actually exploit statistical information to learn the words of their language. However, they clearly undermine the a priori argument that such statistical information is too impoverished to be useful in word learning.

Overall, this conclusion strengthens the power of the statistical learning view of language acquisition. This being said, we are not arguing for a view in which statistical learning would be the unique determinant of word segmentation. Analyses of different languages have shown that a number of phonological and prosodic features are potentially informative about word structure, and there is overwhelming evidence that at least some of these features are actually exploited by the learner (see Jusczyk, 1997, for a review). The question of how statistical and other available cues combine has been investigated in a

number of recent studies (e.g., Christiansen et al., 1998; Gómez, 2007; Johnson & Jusczyk, 2001; Onnis, Monaghan, Richmond, & Chater, 2005; Seidenberg & MacDonald, 1999; Shukla, Nespore, & Mehler, 2007; Thiessen & Saffran, 2003; Tyler, Perruchet, & Cutler, 2006). We do not challenge the general framework on which this literature is grounded. However, considering (1) that people are able to exploit TP<sub>back</sub> when TP<sub>backs</sub> are the only cues for word boundaries in an artificial language, as has been shown in this article, (2) that TP<sub>back</sub> can be exploited in parallel with TP<sub>for</sub> without committing the learner to an intractable computational burden, as is suggested in PARSER (Perruchet & Vinter, 1998), and (3) that TP<sub>backs</sub> are as informative about word boundaries as TP<sub>fors</sub> in a natural language (Swingley, 1999), our point is simply that TP<sub>back</sub> should be included as one of the potential cues to guide infants' discovery of the words of their language.

### AUTHOR NOTE

This work was supported by a grant from the Conseil Régional de Bourgogne (Contrat de Plan État-Région 2006 to Ronald Peereman), by the Université de Bourgogne, and by the Centre National de la Recherche Scientifique (CNRS, UMR 5022). The authors thank Bob French for his helpful comments on an earlier version of the manuscript. Correspondence concerning this article should be addressed to P. Perruchet, Université de Bourgogne, LEAD/CNRS, Pole AAFE, Esplanade Erasme, 21000 Dijon, France (e-mail: pierre.perruchet@u-bourgogne.fr).

### 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.
- 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.
- DUTOIT, T., PAGEL, V., PIERRET, N., BATAILLE, F., & VAN DER VRECKEN, O. (1996). The MBROLA project: Towards a set of high-quality speech synthesizers free of use for noncommercial purposes. In H. T. Bunell & W. Isardi (Eds.), *Proceedings of the Fourth International Conference on Spoken Language Processing* (pp. 1393-1396). Wilmington, DE: Alfred I. duPont Institute.
- ELMAN, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.
- GÓMEZ, R. L. (2007). Statistical learning in infant language development. In M. Gareth Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 601-615). New York: Oxford University Press.
- GOODSITT, J. V., MORGAN, J. L., & KUHL, P. K. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language*, *20*, 229-252.
- HUNT, R. H., & ASLIN, R. N. (2001). Statistical learning in a serial reaction time task: Simultaneous extraction of multiple statistics. *Journal of Experimental Psychology: General*, *130*, 658-680.
- JOHNSON, E., & JUSCZYK, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory & Language*, *44*, 548-567.
- JUSCZYK, P. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.
- KESSLER, B., & TREIMAN, R. (1997). Syllable structure and phoneme distribution. *Journal of Memory & Language*, *37*, 295-311.
- ONNIS, L., MONAGHAN, P., RICHMOND, K., & CHATER, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory & Language*, *53*, 225-237.
- PEEREMAN, R., DUBOIS-DUNILAC, N., PERRUCHET, P., & CONTENT, A. (2004). Distributional properties of language and sub-syllabic processing units. In P. Bonin (Ed.), *Mental lexicon* (pp. 215-235). New York: Nova.
- PEÑA, M., BONATTI, L. L., NESPOR, M., & MEHLER, J. (2002). Signal-driven computations in speech processing. *Science*, *298*, 604-607.

- 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**, 97-119.
- PERRUCHET, P., & VINTER, A. (1998). PARSER: A model for word segmentation. *Journal of Memory & Language*, **39**, 246-263.
- SAFFRAN, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory & Language*, **44**, 493-513.
- SAFFRAN, J. R. (2002). Constraints on statistical language learning. *Journal of Memory & Language*, **47**, 172-196.
- SAFFRAN, J. R., NEWPORT, E. L., & ASLIN, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory & Language*, **35**, 606-621.
- SEIDENBERG, M. S., & MACDONALD, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, **23**, 569-588.
- SHANKS, D. R. (1995). *The psychology of associative learning*. Cambridge: Cambridge University Press.
- 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. (1999). Conditional probability and word discovery: A corpus analysis of speech to infants. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 724-729). Mahwah, NJ: Erlbaum.
- 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: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning & Development*, **3**, 73-100.
- THOMPSON, S. P., & NEWPORT, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning & Development*, **3**, 1-42.
- TYLER, M. D., PERRUCHET, P., & CUTLER, A. (2006, November). *A cross-language comparison of the use of stress in word segmentation*. Poster presented at the 152nd Meeting of the Acoustical Society of America, Honolulu.
- YANG, C. D. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences*, **8**, 451-458.

#### NOTES

1. The difference between the two values depends on the probabilities of  $X$  and  $Y$ , as defined by the Bayes theorem:

$$\text{probability of } X|Y = \frac{(\text{probability of } Y|X)(\text{probability of } X)}{(\text{probability of } Y)}$$

2. Computations were performed after eliminating the two pairs of syllables that occur both in within-word and between-word positions (*bupa* and *babu*).

3. As was pointed out by Aslin et al. (1998) in their note 1: "There are other conditional probability statistics (e.g., conditional entropy, mutual information, correlation) that are functionally equivalent to transitional probabilities, in that they all normalize co-occurrence frequency by the overall frequency of individual events. Any of these conditional probability statistics, including backward transitional probability (the probability of  $X$  given  $Y$ ), provides information for word segmentation on the basis of low predictability at word boundaries."

(Manuscript received August 22, 2007;  
revision accepted for publication May 16, 2008.)