Word segmentation: Trading the (new, but poor) concept of statistical computation for the (old, but richer) associative approach

Pierre Perruchet and Bénédicte Poulin-Charronnat

1. Introduction: The state of the affairs

Oft-cited studies have shown that infants, children, and adults can extract word-like units (hereafter: words) from an artificial language in which these units have been concatenated without any phonological or prosodic markers (e.g., Saffran, Newport, and Aslin 1996). This attests to the fact that listeners are able to exploit the statistical information available in language. More precisely, boundaries between words could be found through the computation of transitional probabilities (hereafter: TPs; Aslin, Saffran, and Newport 1998). Participants would exploit the fact that, on average, the TPs between word internal syllables are stronger than the TPs between syllables spanning word boundaries[1]. The idea that word segmentation of artificial languages attests to the computation of TPs is taken as a foundational principle in most papers in the domain.

Although statistical structure is the only source of information made available in the experimental studies mentioned above, it is commonly thought that word segmentation of natural languages also relies on other cues. The role of phonological and prosodic features, such as lexical stress placement, on word discovery is the best documented (e.g., Creel, Tanenhaus, and Aslin 2006; Curtin, Mintz, and Christiansen 2005; Thiessen and Saffran 2007). Some of these cues (e.g., the pauses) are certainly universal, gestalt-like cues for the formation of perceptual units, and even cues that are seemingly language-specific could be more universal than once thought (Berent and Lennertz 2010; Endress and Hauser 2010, Yoshida, Iversen, Patel, Mazuka, Nito, Gervain, and Werker 2010). In addition to acoustical cues, a second

[1] In keeping with the prevalent view, only the syllabic level will be considered. However, it is worth stressing that statistics computed at the level of phonemes could be also, and even more relevant to find word boundaries (e.g., Hockema 2006). All the proposals of this chapter could be applied at the phonemic level as well.

potential source of information (hereafter coined as "contextual information") is provided by the known words surrounding the to-be-discovered new words. To borrow an example given by Dahan and Brent (1999): "If *look* is recognized as a familiar unit in the utterance *Lookhere!* then *look* will tend to be segmented out and the remaining contiguous stretch, *here*, will be inferred as a new unit" (p. 165). Various experimental evidence of this phenomenon has been reported (Bortfeld, Morgan, Golinkoff, and Rathbun 2005; Cunillera, Camara, Laine, and Rodriguez-Fornells 2010; Dahan and Brent 1999; Perruchet and Tillmann 2010).

The view that statistical computations are complemented by the exploitation of acoustical cues and contextual information is quite consensual (e.g., Aslin et al. 1998; Christiansen, Allen, and Seidenberg 1998; Gomez 2007; Seidenberg and MacDonald 1999; Thiessen and Saffran 2003). There are some disagreements, however, regarding how statistical computations combine with other cues, and notably with prosodic or phonological information[2]. The action of different cues can be thought of as being mediated by independent processes, which would operate in parallel. Statistical computations would be blind to the perceptual properties of the material. This is the view advocated by Shukla, Nespor, and Mehler (2007), at least with regard to prosody. The authors suggest that TP computations (or other forms of statistical computations) over syllabic representations of speech rely on encapsulated, automatic processes, which proceed irrespective of prosodic break-points. Prosody would act subsequently as a filter, suppressing possible word-like units that straddle two prosodic constituents. Another possibility is that statistical learning is "guided" by perceptual factors. It has long been claimed that the exploitation of statistical regularities needs to be constrained by external factors. The acoustical properties of the speech flow could serve as such constraints (e.g., Gomez 2007; Onnis, Monaghan, Chater, and Richmond 2005; Saffran 2002; Seidenberg and MacDonald 1999). Still another view is that statistical computations would be performed on representations that embed prosodic or phonological

[2] To the best of our knowledge, the role of contextual information in the discovery of new words has never been considered jointly with the role of statistical computation. A TP, for instance, is a value inherent to a pair of syllables, which does not depend on whether the local context in which this pair of syllables appears is known by the learner. As asserted by Dahan and Brent (1999), "transitional-probability computations do not take into account the segmentation points in previous utterances; in other words, having isolated some words does not help in isolating other words or even the same words later on" (p. 166).

information. Curtin et al. (2005), for instance, suggested that stressed and unstressed syllables with the same segmental content could be considered as different primitives for the computation of TPs.

Our starting point is that the lack of any principled constraints regarding the interplay of statistical computations with the other factors that have been shown to influence word extraction in natural languages comes from the conceptual indeterminacy of the notion of statistical computation itself. To begin with, what is meant exactly by *computation* in this context is not clearly defined. Some authors may have in mind the type of formal computations that a statistician would do in the same situation (the main difference being that human learners would perform them unconsciously), whereas others may prefer to think that learners approximate statistics through the progressive tuning of associative links on the model of neural networks. However, whatever the preferred option, the same caveat remains: The notion of statistical computations remains underdetermined, as a superimposed piece in the architecture of the mind.

2. Our thesis

The statement above could prompt us to delve into the notion of statistical computation, with the elaboration of a new integrative framework as an ultimate objective. But do we actually need a *new* framework? Is it plausible that word segmentation would require a mode of learning that, despite its claimed importance and pervasiveness, would have been undetected after the overwhelming amount of research devoted to learning processes during so many decades?

In a nutshell, our thesis is that the phenomena currently encompassed under the label of "statistical learning" are nothing else that the end-product of ubiquitous associative learning processes and moreover, that the associative tradition in fact provides the basic architecture for a much more powerful integrative framework than provided by the recent literature on word segmentation and language acquisition.

Such a thesis may look paradoxical. Associative learning may sound like an outdated, old-fashioned concept, which would have lost much of its relevance for language since Chomsky's (1959) famous commentary on Skinner's Verbal Behavior. Statistical learning, by contrast, seemingly provides a flourishing framework, and there is some irony to replace a new and promising concept by an older, and apparently worn-out one.

An immediate objection to our proposal could be that an associative framework is *a priori* ill-suited because it would be unable to account for the learning of TPs in the first place. Going against this objection, we will show in the next section (Section 3) that it *has been* demonstrated for 40 years or so that conventional processes of associative learning make learners sensitive to TPs and even to more complex measures of statistical association. Of major interest is that researchers, far from referring to some unspecified computational abilities to account for this sensitivity, have provided interpretations relying on simple psychological processes. Section 4 will show that endorsing an associative view does not change only our understanding of the way human learners exploit the statistical structure of the input to extract the words: An associative view naturally accounts for the action of the other factors that have been shown to be influential on word segmentation, namely the presence of acoustical cues and contextual information. Moreover, as we will argue in Section 5, the mechanisms involved in the exploitation of statistical regularities, when considered in a dynamic perspective, give to the other sources of information an influence that would be far weaker if they were considered in isolation. To summarize, our claim is that trading the recent statistical learning view for the older associative tradition allows a dynamic integration of phenomena that otherwise would require an array of limited-scope and *ad-hoc* processes.

3. Accounting for the sensitivity to statistical regularities

3.1. Which statistics?

Before examining how associative principles account for the human sensitivity to statistical regularities, one needs to make clear the type of statistics that is involved here, and notably, to assess whether the current focus on TPs is actually warranted. For the sake of simplicity, let us consider the case of two successive events, A and X. Table 1 displays a standard 2 × 2 contingency matrix between the two events.

A first index of relationship is given by *a*, which represents the number of AX pairs. For somewhat obvious reasons, the pure co-occurrence frequency is quite limited as an indicator of the strength of the relationships between two events. The TP, which is the probability that A be followed by X and can be computed as $\mathrm{TP(X|A)} = a/(a+b)$, indisputably provides a more relevant measure. Considering co-occurrence frequency and TP may lead to opposite conclusions about the strength of an association. For instance, "ED" is a more frequent bigram than "QU" in written English,

Table 1. A contingency matrix: "*a*" stands for the number of AX sequences, "*b*" for the number of occurrences of A not followed by X, "*c*" for the number of occurrences of X not preceded by A, and "*d*" for the number of events comprising neither A nor X. $\bar{X}$ can be read as any contexts from which X is absent (this is the usual case in the conditioning domain), or alternatively, as an identified event or class of events (this is the usual case in the word segmentation literature), and likewise for $\bar{A}$.

		e2	
		X	$\bar{X}$
el	A	*a*	*b*
	$\bar{A}$	*c*	*d*

hence suggesting that "E" is more strongly associated with "D" than "Q" with "U"; However, "Q" is much more predictive of "U" than "E" is predictive of "D" ("E" is more often followed with "R" or "S" than with "D"; Source: Wikipedia, *letter frequency*). Aslin et al. (1998) provided the first demonstration in the word segmentation domain that humans, and more precisely 8-month-old infants, are sensitive to TPs when the raw co-occurrence frequency has been controlled.

However, TP as such provides only a part of the information about relationships between A and X. To assess whether A is a useful cue for the occurrence of X, the probability of X in the presence of A should be compared with the probability of X in the absence of A. If, for instance, there are better or more salient predictors of X, it may be adaptive to ignore A and to focus on more relevant events, irrespective of whether A carries some predictive information on the occurrence of X when it is considered in isolation. The resulting statistic is Delta P, which stands as: Delta P $= a/(a+b) - c/(c+d)$. In a classical paper, Rescorla (1968) demonstrated that rats were not only sensitive to TPs (rather than raw frequencies), but also to Delta P, and this result has been confirmed and generalized to many other species in subsequent studies.

It is worthwhile to note that neither TP nor Delta P are complete measures of associations, because both are limited to assess the forward relations between A and X (i.e., the probability that X follows A). The strength of an association may also be related to the backward relationships (i.e., the probability that A precedes X). Backward TP, denoted as TP$'$, can be computed as: TP$'$(A|X) $= a/(a+c)$, and likewise, Delta P$'$ can be computed as: Delta P$' = a/(a+c) - b/(b+d)$. Relying on forward or backward relations may, again, lead to opposite conclusions. For instance, in the bigram "QU", the forward TP is nearly 1, whereas the backward TP is very low, because "U" can be preceded by most other letters of the alphabet. Perruchet and Desaulty (2008) showed that adult participants were able to learn the words from an artificial language when the only available cues were the backward TPs. This ability was confirmed by Pelluchi, Hay, and Saffran (2009) in 8-month-old infants.

It makes sense to conceive that the tightness of the association between A and X depends on both forward and backward relationships. Interestingly, Pearson r, commonly called "r_Φ" with dichotomous data, can be expressed as the geometric mean of forward and backward Delta Ps (Perruchet and Peereman 2004), and written as:

$$r_\Phi = \sqrt{a/(a+b) - c/(c+d)).(a/(a+c) - b/(b+d))}$$

Note that alternative measures of association (such as χ^2 and mutual information) also assess bidirectional relations. A few studies suggest that learners could be sensitive to bidirectional measures of association, such as Pearson r (Perruchet and Peereman 2004) and mutual information (Swingley 2005)[3].

3.2. Associative interpretations

Up to now, whereas our initial objective was to account for the behavioral sensitivity to forward TPs through simple associative mechanisms, we

[3] As an aside, this discovery has implications in the debate surrounding the question of the relative weight different factors may play in language acquisition. Indeed, Yang (2004) has questioned that statistics could play a substantial role, based on analyses of child-directed corpuses showing that even an optimal exploitation of (forward) TPs has nearly negligible usefulness in extracting the words from natural language. His estimations may have been underestimated, however. Swingley (1999) considered bidirectional relations, and when his analyses were restricted to the words that occurred five or more times in the corpus, more than 60% of the extracted units were words (i.e., accuracy score), and less than 40% of actual words were not extracted (i.e., completeness score). These values clearly undermine the *a priori* argument that statistical information would be too impoverished to be useful in word learning, and, although this does not prove that infants actually use this information, it would seem somewhat ill-adaptive that a source of information that is both useful (Swingley 1999) and easily exploitable by learners including infants (e.g., Saffran et al. 1996), would be neglected.

have seemingly moved in the wrong direction. Indeed, the statistics to which human learners are sensitive are in fact much more complex than forward TPs[4]. From a computational standpoint, the increase in complexity is indeed unquestionable: As it can be seen above, the formula for (forward) TPs is included as a small initial component in the formula for r_Φ. We intend to show now that sensitivity to complex statistical measures such as r_Φ (and *a fortiori* to simpler measures) can in fact be accounted for by very simple mechanisms.

Let us consider again the 2 × 2 contingency table above. To recap, claiming that performance does not only depend on the frequency of co-occurrence means that with *a* fixed, performance also depends on *b* (increasing *b* decreases the forward TP), on *c* (increasing *c* decreases the backward TP), or on both *b* and *c* (increasing *b* and *c* decreases the correlation). In other terms, for a fixed number of AX pairs, the probability of creating an association between A and X depends on the number of A and/or X events that are perceived in other contexts. Overall, the larger the number of A and/or X events perceived in other contexts, the more difficult is the formation of the AX chunk. We propose below two complementary accounts for this outcome, an attention-based account and an interference-based account, both relying on associative learning principles.

3.3. An attention-based account

A problem in referring to the associative learning tradition is that everyone having heard about the effects of repetition and extinction, and knowing concepts such as memory strength, decay and interference, may believe to master the basic tenets of this approach. These notions are indeed important. However, there are also other essential notions that are often neglected, with the role played by attentional factors certainly being the main one. Indeed, a fundamental principle is that the formation of any associative links between two elements depends on the learner's joint attention to those elements. This principle obviously holds for complex and supervised forms of learning (any teacher presumably attempts to capture students' attention), but for the most simple and implicit forms of

[4] Note that learners' ability to exploit backward TPs as well as forward TPs is no longer compatible with the prediction-based logic of Simple Recurrent Networks (SRNs), which are the most used computational models to simulate sequential learning. Because relying on SRNs for forward TPs and other mechanisms for backward TPs would lack parsimony, this pattern of results suggests that the SRNs are not the most appropriate models in this context.

learning as well (e.g., Hsiao and Reber 1998; Logan 1988; Mackintosh 1975; Pacton and Perruchet 2008). Unsurprisingly, the word segmentation domain does not make exception (Toro, Sinnett, and Soto-Faraco 2005).

Note that conceiving attention as a condition for the formation of an association between two events is not a late and surreptitious addition from cognitive researchers to the conventional associative view. A study reported in a book published in 1932 by Thorndike (who may hardly be suspected of cognitive penchant) illustrates the point. We report this study with some details[5], because it also enlightens what is meant exactly by "attention" in this context.

Thorndike gave learners 254 word-number pairs. Half of the participants were asked to hear the words and numbers without paying special attention to them, while the other half were asked to pay as close attention as they could do to the materials. In a first test, participants were presented with a word and asked to supply its associated number in the training list. Results revealed that subjects who actively attended to the list outperformed those who were told to remain passive, but the recall score of passive subjects was largely above chance, nevertheless. More importantly, in a second test, the participants were presented with a *number* and asked to supply the *word* that began the next pair in the training list. In fact, unbeknown to the participants, some word-number pairs were so placed in the training series that a particular word always followed a particular number. Recall scores were at chance on the second test, even for the attentive group (for recent, conceptually similar findings, see Pacton and Perruchet 2008).

For Thorndike, these results exemplify a condition for associative learning that he coined as *the principle of belonging*. The property of belongingness refers to as whether events are perceived as going together, without a logical or causal link between the events of concern being required in any way. The intended meaning is that incoming information is naturally divided into a succession of units, and that a necessary condition for the creation of an associative link is that the to-be-associated events are perceived as belonging to a same unit. If a list is perceived as a succession of word-number pairs, associations between the (final) number of a pair and the (initial) word of the next pair are not learned, whatever the "objective" contiguity of the two events, and whatever the amount of attention devoted to the task. The principle of belonging usefully specifies the claim

[5] Based on the report given by Postman (1962). We thank David Shanks for having drawn our attention to this work and made it available on his website.

that attention is necessary for learning. Certainly, some quantitative level of orientation towards the to-be-learned materials is needed, but the amount of attention learners naturally pay to their environment, without special intentional effort and concentration, seems to be sufficient, as exemplified in the "passive" group of Thorndike (1932). The critical issue is the qualitative content of the fleeting attentional focus. A condition for the establishment of stable associations between two events is the perception of these events within a single attentional chunk.

As an aside, learning dependency on attention is essential to the explanatory power of an associative view[6]. Indeed, one of the recurrent objections against the relevance of associative processes for language acquisition is the idea that the number of possible associations between the elements displayed in the input is so large that an explanation based on the exploitation of statistical regularities would be doomed to failure due to combinatorial explosion (e.g., Pinker 1984). This objection essentially reflects a misrepresentation of the laws of associative learning, because it is known for long that the possibility of formation of new associations depends on a number of constraints. However, the most pervasive of these constraints is certainly those stemming from the need of attention: Attention serves as a natural filter to avoid combinatorial explosion. Interestingly, this filter does not act as a blind mechanism that would operate a random selection among the possible candidates for the creation of new associations. Indeed, attention is naturally oriented towards events that have high chance of being relevant in the current context, due to the intrinsic properties of these events (e.g., attention is captured by novelty, and novel events are presumably those that need to be integrated in new associative networks) and to the efficiency of social cues, even in infants (e.g., Wu and Kirkham 2010).

Coming back to the concern of this chapter, how may attention account for the behavioral sensitivity to statistical structure? The classical literature on animal conditioning has provided some responses (e.g., Mackintosh 1975; Pearce and Hall 1980). We do not intend to enter into the intricacies of the debates surrounding this issue. Rather, we present here a simplistic, general sketch, in the hope of making clear the gist of attention-based

[6] Paradoxically, considerable effort has been devoted to demonstrate the possibility of learning without attention or awareness of the to-be-learned relationships, presumably driven by the idea that relaxing learning from the constraints inherent to limited attentional resources seemingly extends its power and field of application. Arguably, the exact opposite is true.

theories. Referring again to the terminology used in the 2 × 2 contingency table above, it is obvious that the attention devoted to A does not only depend on the number of AX pairs (a), but more generally on the overall number of A events ($a + b$). Indeed, the amount of attention devoted to a given event is known to be inversely related to its frequency, because repetitions induce habituation. As a consequence, increasing b (assuming a fixed) necessarily decreases the amount of attention devoted to A, and hence the probability for A and X to be perceived in a single attentional chunk. This accounts for the sensitivity to forward TPs. It suffices to switch round A and X (and, as a way of consequence, b and c) in the reasoning above to account for the sensitivity to backward TPs. Those permutations are perfectly consistent with our knowledge about the role of attention, given that in this account A and X play a symmetrical role (i.e., the formation of an associative link depends on the joint attention to *both* events). Of course, accounting for bidirectional measures of association such as a Pearson r naturally follows.

3.4. An interference-based account

In the reasoning above, the only envisioned consequence of increasing the number of b (and/or c) in Table 1 bears on the raw frequency of A (and/or X): Increasing frequency of an event entails some attentional deficit for this event. However, the matter is a little more complex. In fact, increasing the number of b (and/or c) in Table 1 also increases the number of potential associations of A (and/or X) with one or several other event(s). For example, if (a) is the sequence *gati* and (b) *gafo* and *gamu*, the presence of the latter events make that *ga* is now associated with three different syllables, *ti*, *fo*, and *mu*. This consideration does not invalidate the reasoning above and hence, an explanation based on attention still holds: the syllable *ga* may receive a larger amount of attention if *ga* is only played in *gati* than if *ga* also appears in *gafo* or *gamu*, all simply because *ga* turns out to be less frequent in the former than in the latter case. However, another phenomenon may combine with the modulation of attention, namely the generation of interference.

We refer here to nothing else than the classical paradigm of interference, such as described in any psychology textbooks: People learn a first list of pairs (usually coined as AB) and then a second list of pairs that bears relation to the initial target pairs (AC). Learning AC has a more detrimental influence on the retrieval of AB than learning a list of unrelated items (e.g., DE). In keeping with this phenomenon, the memory for *gati* will be impaired by the presentation of *gafo* or *gamu*, with regard to a situation where *ga* would be always followed by *ti*. This effect directly accounts for the behavioral sensitivity to forward TPs. Because interference also occurs backwards (the memory for *gati* would be impaired by the presentation of *foti* or *muti*), the processes of interference also account for the behavioral sensitivity to backward TPs and, by way of generalization, to bidirectional measures of association.

Any explanations relying on simple and general principles to account for complex phenomena are often questioned for their power by those who find *ad-hoc* modules or mechanisms more attractive. Fortunately, implementing these simple accounts into computational models allows to address directly this concern. PARSER (Perruchet and Vinter 1998) is a computational model that is devised to discover words from a non-segmented speech flow without involving any other principles or processes than those belonging to the associative tradition. Based on the phenomenon that, in humans, attentional coding of the ingoing information naturally segments the material into disjunctive parts, the model is provided online with a succession of candidate units, some of them relevant to the structure of the language and others irrelevant. The relevant units emerge through a selection process based on forgetting. Crucially, forgetting is the end-product of both decay and interference.

Decay is implemented as a linear decrement of the weight of the candidate units across the training session. If forgetting was only due to decay, PARSER would be only sensitive to the raw frequency of co-occurrences (i.e., *a* in Table 1): The candidate units resisting to forgetting would be all simply those that occur the more frequently in the speech flow. Interference allows the model to be sensitive to much more sophisticated measures of cohesiveness. Let us consider two artificial languages, one (L1) in which *gati* would be a word, and a second language (L2) in which *gati* would occur as a part-word (i.e., at least one word would end by *ga* and at least one other word would begin by *ti*). It is of course possible that *gati* occurs as frequently in L1 as in L2 (as in the Aslin et al.'s 1998 design), hence making a decay process inefficient to discover that *gati* is a word in L1 but not in L2. The point is that in L2, *ga* will be necessarily followed by other syllables, hence making possible the creation of other candidate units, such as *gafo* or *gamu*, which generate interference with *gati*. As a consequence of interference, *gati* should normally disappear from the lexicon of a learner trained with L2, by contrast with a learner trained with L1, for whom interference would be null or reduced. These processes should account for the model's sensitivity to forward TPs. As for the attention-based model described above, the role of the two to-be-associated events can be switched, hence accounting for backward TPs and bidirectional measures of associations.

Data modelization fully confirmed these predictions (e.g., Perruchet and Peereman 2004). Whereas some critics of PARSER doubted that such a simple model would be sensitive to more than raw co-occurrence frequencies (e.g., Hunt and Aslin 2001), this achievement demonstrates that very elementary associative principles are powerful enough to account for the behavioral sensitivity to sophisticated statistics.

To resume our main point up to now: The concepts evolved in the context of the literature on associative learning and memory are sufficient to account for the behavioral sensitivity to measures of associations that include the standard TPs, but also much more sophisticated and powerful statistical measures. Of course, this does not rule out the idea that human learners compute TPs or other statistics, as the prevalent view contends. However, there is no need for such a postulate: Behavioral sensitivity to these measures can be understood alternatively as a by-product of a few ubiquitous processes, which are at the core of the associative tradition.

4. Integrating other sources of information

As recalled in the introductory section, acoustical cues in the speech flow and contextual information are also exploited in the word segmentation of natural languages. Insofar as the exploitation of statistical structure is attributed to *ad-hoc* computational abilities, as in the mainstream tradition, the role played by these factors appears as superimposed to statistical computation, without principled constraints to predict whether and how the multiple factors could interact. Our thesis is that once the vague notion of "statistical computation" has been traded for well-known, specified mechanisms of associative learning and memory, a unified, integrative interpretation arises naturally.

Let us consider again the contingency matrix in Table 1. The core condition for the formation of an association between A and X, or in other words the formation of the chunk AX, is the learner's joint attention to A and X. In the prior section, we have examined how pure distributional factors (i.e., the occurrence of A and/or X in other contexts) can modulate the amount of attention devoted to these events and the pattern of interference, hence accounting for the sensitivity to the statistical structure. However, it is obvious that the chance for A and X to belong to a same attentional chunk at the outset of training also depends on a number of factors, and notably on prosodic and phonological factors.

Let us assume that the sequence AX appears in one of the three following conditions (these conditions are freely inspired from Shukla et al. 2007):

(1) [S1-S2-S3-A-X-S6-S7-S8-S9-S10] [S1-S2-S3-A- . . .

(2) [S1-S2-S3-S4-S5-S6-S7-S8-A-X] [S1-S2-S3-S4- . . .

(3) [X-S2-S3-S4-S5-S6-S7-S8-S9-A] [X-S2-S3-S4- . . .

where Sx are different syllables (their number conveys no other information than the fact they are different), and square brackets mark intonational phrases, that is, a set of syllables bounded by natural break-points in speech. For instance, "] [" may be a short pause in the speech flow. From a purely statistical standpoint, the relation between A and X is the same in all three cases. If one endorses the view that statistical computations are performed by the listeners, there is no reason to think that these computations would differ between the three conditions, and hence, any difference between conditions will be taken as attesting to the action of other processes. By contrast, if the apparent results of statistical computations are taken into account by the action of ubiquitous processes of associative learning and memory as proposed above, the predictions that can be put forth are straightforward. In (1), AX is included within an intonational phrase, so it is possible for AX to be perceived within an attentional chunk. This is not necessary however: Ten syllables surely exceed the range of an attentional percept, and it is possible that spontaneous processes of chunking lead to introduce a subjective boundary between A and X. The probability for A and X to be separated under condition (2) is certainly far smaller, due to the fact that the pair AX is limited on the right by a natural break-point. By contrast, the probability for A and X to be perceived in a single attentional chunk is nearly null under condition (3), given that they are separated by a natural break-point. Therefore an associative view predicts that discovering AX will be possible in (1), optimal in (2) because the edge effect maximizes the probability for AX to be perceived within a single percept, and nearly impossible in (3), because there is no chance for AX to be perceived as an attentional unit. These predictions are direct consequences of the Thorndike's principle of belonging. Interestingly, these predictions are clearly confirmed by the results provided by Shukla et al. (2007; see also Seidl and Johnson 2006, for related data on infants).

The effect of contextual information of the speech flow on the discovery of new words can be easily accounted for within the very same framework. Indeed, attentional chunks do not overlap. In the Thorndike's (1932) experiment outlined above, if word1-number1 is a unit and word2-number2 is the following unit, number1-word2 cannot be also perceived as a unit. As a consequence, the element following a known chunk is naturally perceived as the beginning of a new chunk. Let us consider the two following conditions:

(1) S1-S2-S3-A-X-S6-S7- . . .

(2) S1S2S3-A-X-S6S7- . . .

where the deletion of hyphens in (2) is intended to mean that "S1S2S3" and "S6S7" are known words, instead of being perceived as sequences of independent syllables. The probability of perceiving AX as a single percept is obviously stronger under condition (2) than condition (1), and hence again, although a computational view would consider the two situations as similar, the predictions of an associative framework are straightforward: AX should be easier to learn under condition (2) than under condition (1), as found in empirical studies (e.g., Bortfeld et al. 2005; Cunillera et al. 2010).

To resume, in an associative view, the effects of statistical structure, acoustical cues, and contextual information, cannot be separated. All of them are aimed at modifying the probability that the to-be-related events (A and X in Table 1, which may be, for instance, the syllables composing a word in real world settings) are perceived within a single attentional unit, to serve as raw material to the action of associative processes.

5. Towards a dynamic view

5.1. A general outline

Up to now, only general principles from the associative tradition have been involved. To the best of our knowledge, their application to the word segmentation issue is novel, as well as the demonstration that simple interference processes may account for the sensitivity to TPs and more complex measures of association. However, we did not introduce any new postulates or concepts. By contrast, the proposal that follows relies on a principle that, although in no way contradictory to the laws of associative learning and memory, has not been exploited yet in this approach. This principle posits that among the conditions susceptible to focus attention on a set of components – or in other words, to fulfill the Thorndike's principle of belonging – is the fact that these components have been perceived as a chunk in earlier processing episodes.

If the earlier processing episodes have been so frequent that a long-lasting representational unit has been created, the above principle is trivial: Everyone would agree that a familiar word or object is perceived as a whole in adults. The new point here is that the phenomenon occurs at a much smaller scale, for instance during the few seconds or minutes following a single episode. Although the literature on priming could provide at least indirect evidence supporting this idea, more direct demonstration have been brought out in the domain of object perception (e.g., Scholl 2001). Objects are often defined as the product of gestalt-like grouping principles, presumably innate, such as the principles of continuation or common fate. It is also largely acknowledged that long-term familiarity with objects leads to process their features as a whole. However, it has been shown also that even a recent and sporadic experience with a novel shape is sufficient to facilitate its processing as a unitary whole upon subsequent occurrences (e.g., Zemel, Behrmann, Mozer, and Bavelier 2002). When applied to the language domain, this kind of phenomena lies at the root of a dynamic approach to word segmentation.

How does this principle work? Let us assume that a given sequence, say *gati*, has been perceived for the first time within a single attentional chunk due to its acoustical properties, for instance because it was displayed at the end of an intonational phrase. It is highly unlikely that this single presentation would be sufficient to create *gati* as a definitive lexical unit, and it is not ascertain that *gati* will appear in subsequent occurrences in so favorable conditions. Let us suppose that the subsequent occurrence of *gati* occurs within the speech flow without any acoustical markers. If "statistical learning" is conceived of as statistical computations, there is no reason for *gati* to be processed differently from any comparable sequence of two syllables. In our framework, by contrast, *gati* will be specifically strengthened, because provisional internal representations now guide learner's attention as a substitute to external cues.

The very same reasoning holds for the action of contextual information. We referred in the introduction to the Dahan and Brent's (1999) example: If *look* is a word of the language, then *here* will be perceived as a new unit when hearing "*Lookhere!*". Our claim is that the same is true if *look* is a provisional, short-lived chunk. Suppose, for instance, that a child knowing neither *look* nor *here* is told "*Look! Lookhere!*". Because the first occurrence of *look* will be presumably processed as a (provisional) unit due to the prosodic markers, "*lookhere*", instead of being possibly perceived as a single unit, will be correctly segmented, with the beneficial consequences of both strengthening *look* and creating *here*.

What the prior examples illustrate is that a given unit does not need to be perceived in acoustically or contextually favorable conditions on each of its occurrences to acquire a stable internal representation: Once this unit has been perceived as such, it will be naturally strengthened on the near subsequent occurrences whatever its perceptual salience, due to the fact that even short-lived representational units capture attention. Word segmentation, in this framework, appears as the end-result of a dynamical organization. Initially, words (or parts of words) may be perceived as subjective units, due to their acoustical properties or to contextual information, which have been shown to be specially salient in child-directed speech; Then these units are strengthened by the action of associative processes, because the nascent and short-lived representational units generated by earlier experiences create themselves the condition for their own strengthening, namely their processing as a single attentional percept. This phenomenon is all the more adaptive given that analyses of natural language corpuses have shown that the probability of encountering a word that has just been met is strong in the near future, then decreases when time elapses (e.g., Anderson and Schooler 1991).

5.2. Empirical and computational evidence

This framework allows to draw original predictions. If the processing of acoustical and contextual information is independent from statistical computations, the effect of statistical, acoustical, and contextual information should be roughly additive (assuming no ceiling effects). In the framework outlined above, the various sources of information have clearly interactive effects. For instance, minor and sporadic acoustical factors providing a positive cue for word discovery may have a strong effect. Also, if prosodic factors prevent any possibility for two syllables of being perceived in a same attentional focus (as in Condition (3) of the above example, where constituents are on each side of a prosodic boundary), no association will be created, irrespective of the strength of the relation that a statistician could calculate.

These predictions have been tested in a recent study (Perruchet and Tillmann 2010). The experimental situation was quite similar to the situation introduced by Saffran and collaborators (e.g., Saffran et al. 1996).

Participants had to listen to an artificial language composed of six trisyllabic artificial words, randomly concatenated without any pauses. The main difference with regard to the standard situation was that the Initial Word Likeness (IWL) of three of the six artificial words (hereafter, the biased words) was manipulated. IWL refers to the probability of a new sound sequence to be considered as a word, due to its acoustical properties (Bailey and Hahn 2001). For one group of participants (IWL+), the biased words, when heard in a continuous speech stream, were spontaneously perceived as words more often than part-words, whereas the relation was reversed for the second group of participants (IWL−).

Participants from both groups were presented with two successive two-alternative forced choice tests (a word and a part-word) and had to select the syllable set forming a word in the previously heard syllable stream. The first test occurred after a very limited exposure to the language. Unsurprisingly, performance for the biased words was better in the Group IWL+ than in the Group IWL−, due to the effects of acoustical factors (Figure 1, Initial conditions, Black bars). The second test occurred at the end of the experiment, as usual. Two opposite predictions were possible. If one posits that statistical computations are independent from the effect of acoustical factors, then the improvement in performance should be the same in both groups, hence generating additive effects. By contrast, if one posits that the initial biases in favor of some chunks are exploited by associative mechanisms following the dynamic organization outlined above, then there should be a positive interaction. Positively biased words should be *learned* better than negatively biased words. Results clearly supported the second hypothesis (Figure 1, the four Black bars in the top panel).

The other three words of the language were unbiased, which means that, on average, the IWL of these words and the IWL of the part-words that are generated by the concatenation of these words did not differ. The unbiased words, although identical with regard to their IWL and their statistical properties were learned more quickly in the Group IWL+ than in the Group IWL− (Figure 1, four Grey bars in the top panel), attesting that participants were able to exploit their growing knowledge of the biased words to guide the discovery of unbiased words.

To resume, Perruchet and Tillmann's (2010) paradigm allowed to investigate the joint influences of three factors on the discovery of new word-like units in a continuous artificial speech stream: the statistical structure of the ongoing input, the initial word likeness of parts of the speech flow, and the contextual information provided by the earlier emergence of other word-like units. Results showed that these sources of information have

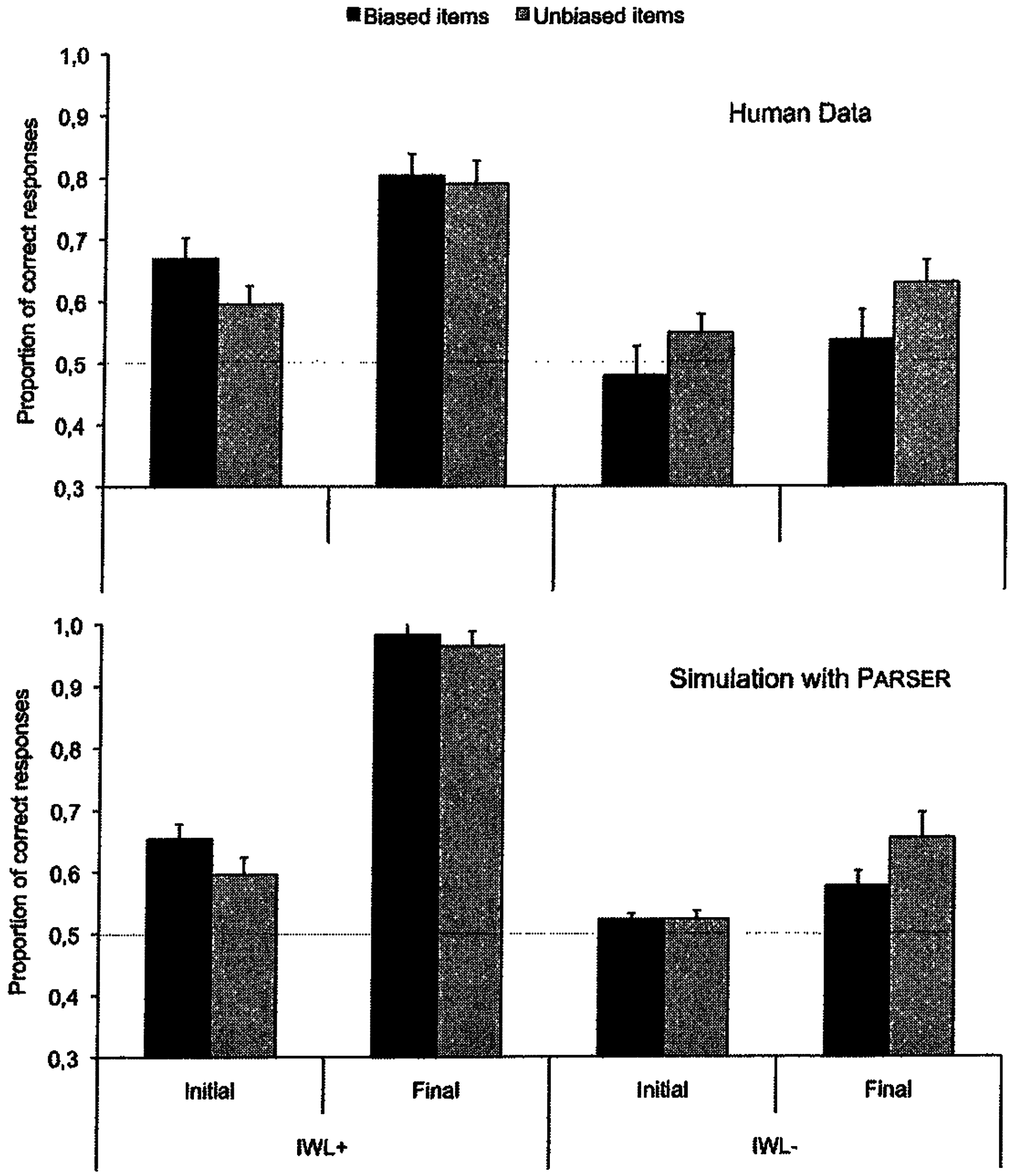

Figure 1. Proportion of correct responses for the initial and final tests, as a function of Groups, for biased and unbiased items. IWL stands for Initial Word Likeness, and designates the probability for a new sound sequence to be considered as a word of the language due to its intrinsic properties, before any training. The biased items were positively biased in the Group IWL+, and negatively biased in the Group IWL−. Unbiased items were identical for the two groups. The top panel shows the data collected on adult humans, and the bottom panel shows the simulations with PARSER (adapted from Perruchet and Tillmann 2010).

strong and interactive influences on word discovery, as clearly anticipated in our framework. To assess the quality of fit in a quantitative way, the data were simulated with PARSER (Perruchet and Vinter 1998). PARSER has been introduced above to illustrate how interference makes learners sensitive to statistics, but one crucial aspect was passed over in silence. Indeed, the model implements the principle that perception is dynamically guided by the emerging representational units.

Although the role of acoustical factors was not implemented in the original version, doing so is straightforward. In the simulations, the initial selection of the candidate units was biased in such a way that instead of being randomly drawn within a given length range, the candidate units were selected (within the same range) as a function of their relative IWL, which was previously assessed in human participants. When the standard parameters used in Perruchet and Vinter (1998) were applied, ceiling effects were observed. When forgetting was progressively increased until ceiling effects disappeared in all conditions, the pattern of results obtained by PARSER reproduced the main effects observed for human participants, as reported in Figure 1 bottom panel. Learning for biased items was better for the Group IWL+ than for the Group IWL–, even though taking performance in the first test as a baseline controlled for the direct effect of IWL on word/part-word selection, and this difference transferred to the unbiased items. The fact that PARSER was successful in generating the pattern observed in human participants without implementing any *ad-hoc* algorithmic changes is worthy of note.

6. Discussion

The starting point of this chapter was the conceptual indeterminacy of the notion of "statistical computation", which is usually invoked to account for the human sensitivity to the statistical structure of the environment, including the language. We showed that behavioral sensitivity to statistical structures has been evidenced long ago in the context of research on conditioning in animals, and has been interpreted as the by-product of basic associative learning principles. At the core of these interpretations is the ubiquitous necessity for two events to be created as a long-lasting representational unit to be perceived within a single attentional chunk in the first place, a phenomenon already described as the "principle of belonging" by Thorndike (1932). We argued that putting associative learning principles as the primary cause for the behavioral sensitivity to statistical structures has the unique advantage of also accounting for the action of the other factors that have been shown to be influential on word segmentation, especially the presence of acoustical cues and contextual information. Finally, we have shown that an associative framework, when complemented with the idea that even a fleeting, short-lived representation of a set of features as a unit shapes further perception, opens to a dynamic interpretation of word segmentation, in which the contributing factors exhibit overadditive interactions. This view has received some experimental and computational support (Perruchet and Tillmann 2010).

Although our approach is critical towards the prevalent account of statistical learning, it is worth noting that we have no reservation with regards to the current empirical studies in the domain and moreover, we fully endorse the view that statistical learning plays a much more substantial role in language acquisition than once thought. In this regard, our approach stands at the exact opposite of another critical vision, emphasizing the need for completing statistical learning mechanisms with a symbolic machinery (e.g., Endress, Dehaene-Lambert, and Mehler 2007; Endress and Mehler 2009; Endress, Nespor, and Mehler 2009). In this view, the action of any perceptual factor modulating the end-result of statistical computations is referred to the action of specialized detectors. For instance, observing that components located at the boundaries of a sequence are processed more efficiently than components located in a middle position, the authors infer the action of an "edge detector" coding this information in a symbolic format. This symbolic information is then sent to other systems, including mechanisms devoted to statistical computations, and acts as a constraint: Any event situated at a boundary is taken as potentially more significant than other events. To take an analogy in physics, it is like if to account for the fact that ice melts at about 0 degrees Celsius, a "head detector" continuously monitors the temperature of the water, stores this information in a propositional format, and when the temperature starts to warm above the target value sends to another system in charge of the melting task a message such as "The melting point has been reached". This account is obviously nonsensical in the physical domain, but half of a century of prevalence of the cognitivist, information-processing view makes that what we construe as its counterpart in cognitive sciences is not always perceived as such.

Instead of adding a (symbolic) layer in the explanatory sketch, our approach amounts to a withdrawal of a notion that has been recently proposed, namely that idea that the mind performs statistical computations. When the principle that the formation of a cognitive unit needs the initial processing of its components into a single attentional percept has been laid down, the action of distributional factors as well as the action of other variables guiding attention – such as edges – directly follows through direct functional couplings.

The question of whether the principles exploited here for the word segmentation issue can be extended to account for other aspects of language acquisition stands beyond the scope of this chapter. This extension to language could meet other approaches relying on similar principles such as the emergentist theory developed by MacWhinney (e.g., MacWhinney 2010). Such an extension has also been proposed as a part of a general model of the mind framed around the concept of self-organizing consciousness (e.g., Perruchet 2005; Perruchet and Vinter 2002).

References

Anderson, John R. and Lael J. Schooler
1991 Reflections of the environment in memory. *Psychological Science*, 2: 396–408.

Aslin, Richard N., Jenny R. Saffran and Elissa L. Newport
1998 Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9: 321–324.

Bailey, Todd M. and Ulrike Hahn
2001 Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44: 568–591.

Baker, Christopher I., Carl R. Olson and Marlene Behrmann
2004 Role of attention and perceptual grouping in visual statistical learning. *Psychological Science*, 15: 460–466.

Berent, Iris and Tracy Lennertz
2010 Universal constraints on the sound structure of language: Phonological or acoustic. *Journal of Experimental Psychology: Human Perception and Performance*, 36: 212–223.

Bortfeld, Heather, James L. Morgan, Roberta M. Golinkoff and Karen Rathbun
2005 *Mommy* and Me. *Psychological Science*, 16: 298–304.

Noam, Chomsky
1959 A Review of B. F. Skinner's Verbal Behavior. *Language*, 35: 26–58.

Christiansen, Morten H., Joseph Allen and Mark S. Seidenberg
1998 Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13: 221–268.

Creel, Sarah C., Michael K. Tanenhaus and Richard N. Aslin
2006 Consequences of lexical stress on learning an artificial lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32: 15–32.

Cunillera, Toni, Estela Camara, Matti Laine and Antoni Rodriguez-Fornells
2010 Words as anchors: Known words facilitate statistical learning. *Experimental Psychology*, 57: 134–141.

Curtin, Suzanne, Toben H. Mintz and Morten H. Christiansen
2005 Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, 96: 233–262.

Dahan, Delphine and Michael R. Brent
1999 On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General*, 128: 165–185.

Endress, Ansgar D., Ghislaine Dehaene-Lambertz and Jacques Mehler
2007 Perceptual constraints and the learnability of simple grammars. *Cognition*, 105(3): 577–614.

Endress, Ansgar D. and Mark D. Hauser
2010 Word segmentation with a universal prosodic mechanism. *Cognitive Psychology*, 61: 177–199.

Endress, Ansgar D. and Jacques Mehler
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.

Endress, Ansgar D., Marina Nespor and Jacques Mehler
2009 Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, 13: 348–353.

Gomez, Rebecca
2007 Statistical learning in infant language development. In M. G. Gaskel (Ed.). *The Oxford handbook of psycholinguistics*, New York: Oxford University Press.

Hockema, Steve A.
2006 Finding words in speech: An investigation of American English. *Language Learning and Development*, 2: 119–146.

Hsiao, Andrew T. and Arthur S. Reber
1998 The role of attention in implicit sequence learning: Exploring the limits of the cognitive unconscious. In Michael. Stadler and Peter. Frensch (Eds.), *Handbook of implicit learning* (pp. 471–494). Thousand Oaks, CA: Sage Publications.

Hunt, Ruskin H. and Richard N. Aslin
2001 Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General*, 130: 658–680.

Logan, Gordon D.
1988 Toward an instance theory of automatization. *Psychological Review*, 95: 492–527.

Mackintosh, Nicolas J.
1975 A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82: 276–298.

MacWhinney, Brian
2010 *Language development.* In M. Bornstein and M. Lamb (Eds.), Developmental science: An advanced textbook (pp. 467–508). New York: Psychology Press.

Onnis, Luca, Padraick Monaghan, Nick Chater and Korin Richmond
2005 Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, 53: 225–237.

Pacton, Sébastien and Pierre Perruchet
2008 An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34: 80–96.

Pearce, John M. and Geoffrey Hall
1980 A model for Pavlovian learning: Variations in the effectiveness of conditioned but not unconditioned stimuli. *Psychological Review*, 87: 532–552.

Pelucchi, Bruna, Jessica F. Hay and Jenny R. Saffran
2009 Learning in reverse: Eight-month-old infants track backwards transitional probabilities. *Cognition*, 113: 244–247.

Perruchet, Pierre and Stéphane Desaulty
2008 A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, 36: 1299–1305.

Perruchet, Pierre
2005 Statistical approaches to language acquisition and the self-organizing consciousness: A reversal of perspective. *Psychological Research*, 69: 316–329.

Perruchet, Pierre and Ronald Peereman
2004 The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics*, 17: 97–119.

Perruchet, Pierre and Annie Vinter
1998 PARSER: A model for word segmentation. *Journal of Memory and Language*, 39: 246–263.

Perruchet, Pierre and Annie Vinter
2002 The self-organized consciousness. *Behavioral and Brain Sciences*, 25: 297–388.

Perruchet, Pierre and Barbara Tillmann
2010 Exploiting multiple sources of information in learning an artificial language: Human data and modeling. *Cognitive Science*, 34: 255–285.

Pinker, Steven
1984 *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Postman, Leo
1962 Rewards and punishments in human learning. In L. Postman (Ed.), *Psychology in the making: Histories of selected research problems* (pp. 331–401). New York: Knopf.

Rescorla, Robert A.
1968 Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, 66: 1–5.

Saffran, Jenny R.
2002 Constraints on statistical language learning. *Journal of Memory and Language*, 47: 172–196.

Saffran, Jenny R., Elissa L. Newport and Richard N. Aslin
1996 Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35: 606–621.

Scholl, Brian J.
2001 Objects and attention: The state of the art. *Cognition*, 80: 1–46.

Seidenberg, Mark S. and Maryellen C. MacDonald
1999 A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, 23: 569–588.

Seidl, Amanda and Elisabeth Johnson
2006 Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, 9: 565–573.

Shukla, Mohinish, Marina Nespor and Jacques Mehler
2007 An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, 54: 1–32.

Swingley, Daniel
1999 Conditional probability and word discovery: A corpus analysis of speech to infants. In M. Hahn and S. C. Stoness (Eds.), *Proceedings of the 21st annual conference of the cognitive science society* (pp. 724–729). Mahwah, NJ: LEA.

Swingley, Daniel
2005 Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50: 86–132.

Thiessen, Erik D and Jenny R. Saffran
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, Erik D. and Jenny R. Saffran
2007 Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, 3: 73–100.

Thorndike, Edward L.
1932 *The fundamentals of learning.* New York: Teachers College, Columbia University.

Toro, Juan M., Scott Sinnett and Salvador Soto-Faraco
2005 Speech segmentation by statistical learning depends on attention. *Cognition*, 97: B25–B34.

Wu, Rachel and Natasha Z. Kirkham
2010 No two cues are alike: Depth of learning during infancy is dependent on what orients attention. *Journal of Experimental Child Psychology*, 107: 118–136.

Yang, Charles D.
2004 Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, 8: 451–456.

Yoshida, Kazuhiro, John R. Iversen, Aniruddh D. Patel, Reiko Mazuka, Hiromi Nito, Judith Gervain and Janet Werker
2010 The development of perceptual grouping biases in infancy: A Japanese-English cross-linguistic study. *Cognition*, 115: 356–361.

Zemel, Richard S., Marlene Behrmann, Michael C. Mozer and Daphne Bevalier
2002 Experience-dependent perceptual grouping and object-based attention. *Journal of Experimental Psychology: Human Perception and Performance*, 28: 202–217.