TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science (2018) 1–16

ISSN:1756-8765 online
DOI: 10.1111/tops.12403

This article is part of the topic "Two Approaches, One Phenomenon: Aligning Implicit Learning and Statistical Learning," Padraic Monaghan and Patrick Rebuschat (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview

# What Mechanisms Underlie Implicit Statistical Learning? Transitional Probabilities Versus Chunks in Language Learning

Pierre Perruchet[a,b]

[a]*Department of Psychology, University of Bourgogne Franche-Comté*
[b]*LEAD-CNRS, UMR 5022*

Received 17 December 2016; received in revised form 13 November 2018; accepted 13 November 2018

**Abstract**

In a prior review, Perrruchet and Pacton (2006) noted that the literature on implicit learning and the more recent studies on statistical learning focused on the same phenomena, namely the domain-general learning mechanisms acting in incidental, unsupervised learning situations. However, they also noted that implicit learning and statistical learning research favored different interpretations, focusing on the selection of chunks and the computation of transitional probabilities aimed at discovering chunk boundaries, respectively. This paper examines the state of the debate 12 years later. The link between contrasting theories and their historical roots has disappeared, but a number of studies were aimed at contrasting the predictions of these two approaches. Overall, these studies strongly question the still prevalent account based on the statistical computation of pairwise associations. Various chunk-based models provide much better predictions in a number of experimental situations. However, these models rely on very different conceptual frameworks, as illustrated by a comparison between Bayesian models of word segmentation, PARSER, and a connectionist model (TRACX).

*Keywords:* Implicit learning; Statistical learning; Computation; Transitional probability; Chunk; Computational modeling; Segmentation; Bayesian inference

Correspondence should be sent to Pierre Perruchet, LEAD, Pôle AAFE, 11 Esplanade Erasme, 21000 DIJON, France. E-mail: pierre.perruchet@u-bourgogne.fr

## 1. Introduction

Perruchet and Pacton (2006) reviewed the arguments supporting the joint consideration of two areas of research, the older area of implicit learning and the more recent approach focusing on statistical learning. They noted that these two approaches were concerned by basically similar leaning situations, hence justifying their common consideration under the label of implicit statistical learning (ISL), as proposed by Conway and Christiansen (2006). However, Perruchet and Pacton also noted that studies on implicit learning and statistical learning favored different interpretations, focusing on chunk formation and statistical computations, respectively. Twelve years later, the historical coupling between research fields and interpretations has virtually disappeared. For instance, Jimenez (2008) and Jimenez, Méndez, Pasquali, Abrahamse, and Verwey (2011) suggest that learning in Serial Reaction Times paradigms, one of the prototypical situations of implicit learning, could not rely on chunks formation as initially believed, and, as will be described later, various models of chunking have been developed primarily for word segmentation, a widely used situation of statistical learning. However, the deep issue has received considerable attention, and the present paper is aimed at reviewing the recent studies that have moved the debate forward.

As in Perruchet and Pacton (2006), the main issue of concern will be the formation of elementary cognitive units. Most studies explored the extraction of artificial words from a continuous sequence of syllables, following the seminal studies of Saffran and coworkers (e.g., Saffran, Newport, & Aslin, 1996). However, the notion of cognitive units is much larger. For instance, a growing number of papers have investigated the formation of units comprising a few words (e.g., Arnon, McCauley, & Christiansen, 2017; Christiansen & Arnon, 2017), in keeping with the upsurge of "usage-based" models of language, which assume that the starting point of language acquisition is the storing of short multi-word utterances (see review in Tomasello, 2009). The formation of visual units from elementary shapes (e.g., Orbán, Fiser, Aslin, & Lengyel, 2008), or still the creation of word-referent pairings (e.g., Benitez, Yurovsky, & Smith, 2016) also illustrates the widespread extension of the notion of cognitive units. Concurrently, the simplistic artificial material initially exploited has been often replaced with more natural settings, such as child-directed language (e.g., Pelucchi, Hay, & Saffran, 2009a). Thus, the field extends its scope to a wider range of issues than the segmentation of a continuous corpus into a few artificial words. This scope could even spread to issues of grammatical categorization and syntax acquisition, as discussed in the final section.

## 2. Statistical computations and chunk formation: The issue

No one denies the role of cognitive units in cognition. For instance, a word–object association may only occur when the word, on the one hand, and the object, on the other hand, are processed as units. It would make no sense to look for a link between, say, a given phoneme and a fragment of an object, because the association exists only at a higher hierarchical level (but see Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011). There is now clear evidence that ISL is one of the processes that lead to the formation of exploitable units (e.g., Fernandes, Kolinsky, & Ventura, 2009; Graf Estes, Evans, Alibali, & Saffran, 2007; Hay, Pelucchi, Estes, & Saffran, 2011; Kibbe & Feigenson, 2016).

Disagreements arise when the mode of formation of these units is considered. The first and more common idea is that when the relevant units are not directly available in the sensory input, prior statistical computations are required. As claimed by Adini, Bonneh, Komm, Deutsch, and Israeli (2015): "Statistical learning and subsequence learning are two successive stages in implicit sequence learning, with chunks inferred from prior statistical computations." The computations mentioned in this context are in all cases measures of pairwise associations between elements of the input, and most often transitional probabilities (TPs; i.e., given AB, the probability for A to be followed by B). The logic is straightforward. In artificial or natural languages, for instance, TPs between syllables composing a word are higher, on the mean, than TPs between successive syllables overlapping two words. As a consequence, dips in the distribution of TPs mark word boundaries. Note that the task of computing TPs is elegantly approximated by Simple Recurrent Networks (SRNs, e.g., Mirman, Graf Estes, & Magnuson, 2010), which have proven their relevance and efficiency in many other domains.

However, the relevant units can also be discovered from the same input using a very different strategy. Because it turns out that even some experts in the field miss the point, a small illustration is worthwhile. Let us consider a text printed without space, beginning with "onceuponatime." The sequence may be segmented as "on/ceu/ponat/ime," "onceu/po/nati/me," "onc/eupo/natim/e," the correct segmentation, "once/upon/a/time," being one among an overwhelming number of possibilities. The key point of the so-called chunk-based models is to consider all or a subset of these possible segmentations, and to select the one that meets better certain criteria. In particular, the units created in the correct segmentation (once, upon...) are more likely to reoccur later in the text than the units created as a consequence of erroneous segmentation (for instance, "ponat" may only reoccur if the whole sequence "upon a time" happens again). As a consequence, selecting the partitioning of a corpus that requires the smallest number of different units (i.e., the shorter lexicon) generally leads to discovering the words. Obviously, this is an oversimplified sketch (a few other constraints are required), but it should suffice for giving an existence proof for a method of segmenting a continuous corpus that requires no preliminary detection of probabilistic dips between units, and more generally, no computation of pairwise statistics, whether unit-internal or unit-external. The sensory input is chunked from the outset, and the correct units emerge through selection.

Before turning to an assessment of these two general classes of models, a terminological clarification is wanted. Even within the cognitive science community (by contrast with the computer science approach), the term of statistical learning is ambiguous. A large proportion of introductory texts to the topic endorses the view initially proposed by Saffran et al. (e.g., 1996), in which statistical learning is "the psychological process by which the transitional probabilities from one syllable to another in the continuous speech streams could enable word segmentation" (Aslin & Newport, 2009). Endorsing this kind of definition unfortunately enshrines the confusion between a set of empirical phenomena and a specific interpretation. ''Statistical learning'' is used here as a theoretically neutral label designating any form of exploitation of the statistical structure of the input. A partially related problem is related to the concept of "statistical computations." Again for historical reasons, "statistical computation" is often identified with the computation of pairwise associations, mostly TPs. Now, it is obvious that some chunk-based models exploit statistical tools. The comparison below will oppose "TP-based" against "chunk-based "views, with the convention that statistical computations may be involved in both cases but for different objectives: identifying the frontiers between chunks as a probabilistic gap (and the chunks in a second step), or directly searching for the correct chunks among a set of candidates (a contrast sometimes framed in terms of *bracketing* vs. *clustering* strategies, e.g., Swingley, 2005).

## 3. TP-based versus chunk-based approaches to ISL: The data

This section reviews the recent studies that were aimed at testing the relative validity of the two approaches, or at least produced data potentially relevant for the debate. Most important, TP knowledge turns out to be neither necessary nor sufficient for extracting chunks.

### *3.1. TPs are not necessary*

If chunks are inferred from the prior computations of pairwise correlations between their components, pairwise correlations must be informative about the chunk structure to allow chunk extraction. Orbán et al. (2008, see also Fiser, 2009) were able to build visual scenes in which pairwise correlations between shapes were of no help in establishing the identity of chunks composed of three elements. Subjects showed a significant preference for these chunks when they were compared to test triplets that shared the same correlational structure, but were not displayed as chunks. Unsurprisingly, models computing conditional or transitional probabilities between two elements of the scene, or all the pairwise correlations, were unable to account for human performance.

The constraints for generating Orban et al.'s material, however, result in a very peculiar structure, possibly endowed with confounded properties (notably because chunks must have many components in common). Another strategy consists in using a standard arrangement, in which pairwise relationships are indicative of chunk structure, and to assess whether these relationships are actually learned whenever chunks are successfully extracted. Giroux and Rey (2009) used an artificial language comprising both trisyllabic words (e.g., ABC) and bisyllabic words (e.g., DE). Then subjects were exposed to a set of two bisyllabic items, and they had to choose the one that seemed more likely to belong to the language to which they had been exposed. For each test pair, one of the items was a bisyllabic part word (e.g., CD) while the other item was either a bisyllabic word (DE) or a bisyllabic component of a trisyllabic word (e.g., AB). Note that AB and DE had the same frequency and the same internal consistency (the between-syllable TPs were 1 in each case). Unsurprisingly, an SRN predicted no difference between the two kinds of pairs. This was indeed the result observed in a group exposed to the language during only 2 min. However, in another group trained during 10 min, performances were significantly better when the test pairs involved bisyllabic words than a bisyllabic pair embedded in a trisyllabic word. This pattern has been successfully simulated by several chunk-based models (French, Addyman, & Mareschal, 2011; Giroux & Rey, 2009; Robinet, Lemaire, & Gordon, 2011). Slone and Johnson (2018) recently provided successful replication in 8-month-olds, using visual shapes. Infants discriminated relevant pairs from pairs embedded in triplets, but, echoing Giroux and Rey, only when they had sufficient exposures to each pair and triplet (80 vs. 40 exposures), suggesting that the formation of larger units impedes the representation of their components.

Fiser and Aslin (2005) and Glicksohn and Cohen (2011) reported that their subjects learned three-element chunks without learning their embedded pairs. As Glicksohn and Cohen pointed out, this finding suggests that "learning cannot be based on conditional probability per se, because such computation should also discriminate embedded pairs from random pairs" (p. 709). In a related vein, although not directly referring to the literature on statistical learning, several studies on sequence learning by Perlman and collaborators (Perlman, Pothos, Edwards, & Tzelgov, 2010; Perlman, Hoffman, Tzelgov, Pothos, & Edwards, 2016; see also Hoffman et al., 2017) lead to the same conclusion. One of the questions raised in these studies is whether identical subsequences embedded into two different, longer chunks are processed identically. For instance, if one of the long chunks is more frequent than another, is the common part processed at the same speed, as it could be expected if performance were guided by the computation of pairwise correlations? The empirical response is clearly "no." The general conclusion of this first subsection is that knowledge of the pairwise relations between elements composing a chunk are not *necessary* for the creation and maintenance of chunk knowledge.

### *3.2. TPs are not sufficient*

While the preceding section is concerned with the issue of necessity, other studies have raised the question of *sufficiency*: Is the knowledge of the pairwise relations between components sufficient to build a chunk? In Endress and Mehler (2009), subjects were familiarized with a continuous language containing trisyllabic words that were generated from (unheard) prototypes. If a prototype is designated as ABC, the heard words were AB*X*, *Y*BC, and A*Z*C (with *X*, *Y*, and *Z* standing for invariant syllables). For instance, participants heard *tazepi*, *mizeRu*, and *tanoRu*, which were all derived from the prototype *tazeRu*. In this way, the (unheard) prototypes had exactly the same TPs between their constituent syllables (i.e., AB, BC, and A_C) than the trisyllabic words composing the language. If subjects have formed genuine chunks, that is, some acoustical word candidates that could be later associated as a whole to a meaning, they should select words over the prototypes when both are played in a subsequent forced-choice test. However, if they only learned pairwise relations, they should be unable to distinguish between the actual words and their prototypes. Endress and Mehler reported that participants failed to distinguish between words and prototypes, hence suggesting that statistical learning generates knowledge about TPs within each pair of syllables, which are common to words and their prototypes.

Null results, however, provide only weak evidence. Perruchet and Poulin-Charronnat (2012a) exploited the Endress and Mehler (2009) procedure, and they get very different results. Participants showed a significant preference for words over prototypes after only 5 min of exposure to the language. Overall, the effect appears remarkably stable, with the rate of correct responses on eight independent groups ranging from 59.52% to 66.25% (with chance set to 50%). Slone and Johnson (2015) exploited the very same paradigm, but replaced the auditory syllables with colored shapes. As in Perruchet and Poulin-Charronnat, adult participants chose the chunks as more familiar than their statistically matched prototypes significantly more often than chance. Slone and Johnson (2018) used a similar procedure, but introduced several simplifications to make the task manageable by 8-month-old children. There was only one prototype and three chunks, and the non-adjacent pair of the prototype (A_C) was no longer present in the chunks. Nevertheless, pairwise relationships between adjacent shapes (AB and BC) were still matched between the chunks and their prototype. Confirming results in adults, infants looked significantly longer during chunk test trials, compared to the prototype test trials. To quote the authors (p. 96), "In contrast to the predictions of statistical models, infants did not appear to represent the familiarization sequence primarily in terms of TPs between adjacent items. [...] Infants represented the familiarization sequence in terms of extracted units, not statistical relations."

To summarize the two prior points, a chunk may be built without knowledge of the pairwise relationships between its components, and knowing these relationships is not sufficient to build a chunk. Other data also run against the idea that chunks emerges from the prior computation of statistical co-occurrences, although more indirectly.

### *3.3. Backward TPs*

When proposing the computation of TPs as the basic mechanism underlying word segmentation, Saffran et al. (1996) and Aslin, Saffran, and Newport (1998) referred in fact to *forward* TPs, which designate how A predicts B in a sequence AB. Now, without specific constraints, a word comprises bidirectional relationships between its constituents (i.e., A predicts B, and B predicts A), and this is true for artificial languages as well as for natural languages (e.g., Swingley, 1999).[1] The focus on forward TPs is consistent with the enduring propensity of many researchers, especially neuroscientists and philosophers (e.g., Clark, 2013), to consider that the mind is specifically engineered to predict future events. However, as noted by Jones and Pashler (2007), "there is a remarkable absence of behavioral tests of this idea." Perruchet and Desaulty (2008) examined the role of forward and backward TPs in word segmentation from artificial languages in which words were based only on forward TPs or backward TPs. They showed that adult subjects were as good at discovering the two kinds of words. This result was replicated in infants with auditory stimuli (French et al., 2011; Hay et al., 2011; Pelucchi, Hay, & Saffran, 2009b) and visual stimuli (Tummeltshammer, Amso, French, & Kirkham, 2017).

Certainly, these data may be encompassed in introducing some modifications into the standard TP-based view. Backward TPs may be used instead of forward TPs (e.g., McCauley & Christiansen, 2014). More generally, the detection of probabilistic dips may rely on the computation of correlation coefficients or Mutual Information, which give equal weight to forward and backward relationships. However, these changes appear to add ad hoc complexity to the initial model, which, as a consequence, loses the support provided by the achievement of an SRN in simulating the postulated processes. Indeed, given that the backpropagation algorithm exploits the error between the predicted and the actual next event in a sequence, an SRN, as a matter of principle, is unable to learn backward TPs or any statistics, including backward-directed information.

By contrast with a TPs-based approach, taking into account both forward and backward relationships is a natural by-product of a chunking process. It is inherent to the nature of a chunk that all its components are mutually linked, without a privileged direction (obviously, this does not mean that the order of its components is irrelevant: "baby" is not "byba"). As will be seen in the next sections, all chunk-based models are sensitive to both forward and backward relationships between chunk components without introducing any ad-hoc machinery (rather, the limitation of these models to an exclusive sensitivity to forward relationships would need ad-hoc algorithmic modifications).

### *3.4. Zipfian distribution and other variables*

Several recent studies have manipulated variables generating different, and generally opposite, predictions from TPs-based and chunk-based models of ISL. Let us consider a representative example (Kurumada, Meylan, & Frank, 2013). By contrast with most artificial languages where words have a uniform frequency distribution, words in natural languages have a frequency distribution that approximately follows the Zipf's law (e.g., Piantadosi, 2014). There are few very high–frequency words, and many low–frequency words. The consequences for a model postulating that word boundaries are identified by the presence of low TPs (e.g., Saffran et al., 1996) are straightforward. Such a model works well with artificial languages because with the standard uniform frequency distribution, virtually all between-word TPs are lower than within-word TPs. However, this is no longer true with a skewed distribution. Indeed, the within-word TPs for low-frequency words can be lower than the between-word TPs for high-frequency words, hence both leading to introduce extra boundaries into low–frequency words, and to pool together high–frequency words. As a consequence, subjects' segmentation should be better with a uniform than a zipfian distribution. A chunk-based model makes the opposite prediction. Such a model should predict better performance with a skewed distribution than with a uniform distribution, notably because high–frequency words should be easily discovered, providing information about the beginning and the end of the surrounding words (this is because successive chunks are generally assumed to be disjunctive: they do not overlap). As a consequence, the whole process of segmentation should be speeded up.

Kurumada et al. (2013) reported simulations confirming that a TP-based model (Saffran et al., 1996) segmented a uniform language better than a Zipfian language, while different chunking models segmented a Zipfian language better than a uniform language. Crucially, they also collected experimental data, and it turned out that adult learners performed better with a Zipfian language, as do the chunking models (see also Frost, Monaghan, & Christiansen, 2016).

Similar studies involving other variables have been published, all of them using artificial languages. A non-exhaustive list includes:

1. The number of different words (Frank, Goldwater, Griffiths, & Tenenbaum, 2010).
2. The effect of the amount of exposure and the time course of performance. Some units are quickly discovered (Trueswell, Medina, Hafri, & Gleitman, 2013), but very difficult to change if statistical regularities are subsequently altered (Zellin, von Mühlenen, Müller, & Conci, 2014).
3. The length of the sentences. Introducing pauses in the speech stream as in natural language has a very beneficial effect on segmentation (Johnson & Tyler, 2010; Sohail & Johnson, 2016).
4. The effect of the prior exposure to a language involving the same syllables. Prior exposure to a language in which words are subsequently located at word transition has a detrimental effect (Franco & Destrebecqz, 2012; Perruchet, Poulin-Charronnat, Tillmann, & Peereman, 2014; Poulin-Charronnat, Perruchet, Tillmann, & Peereman, 2017; see also Mersad & Nazzi, 2012).
5. The beneficial effect of finding the first words for the discovery of the other ones (e.g., Perruchet & Tillmann, 2010).

Space is lacking for analyzing why TP-based and chunk-based models predict different outcomes in each case, but overall, studies involving model comparisons have shown that TP-based models (and more generally models relying on statistical co-occurrences) fail to simulate human data (predictions are often in the opposite direction), while various chunking models are successful (Frank et al., 2010; Kurumada et al., 2013; Meylan, Kurumada, Börschinger, Johnson, & Frank,2012; Perruchet & Tillmann, 2010; Poulin-Charronnat et al., 2017).

When the four points above are considered jointly, it may hardly be denied that the weight of evidence supports chunk-based models against TP-based models. This is a rather startling conclusion, given that a TP-based interpretation was initially put forward as the only possible account for the extraction of words from a continuous artificial language (e.g., Saffran et al., 1996) and, maybe due to this historical prominence, is often still identified with the notion of statistical learning.

## 4. Models of chunking

Up to now, chunk-based models have been taken as a whole, which makes sense because they share a same general strategy, outlined above. However, they differ substantially. The objective of this section is to give insight into the plurality of mechanisms that may account for chunking without involving the prior computation of pairwise statistics. As in Kurumada et al. (2013), three models or classes of models will be considered: the probabilistic or Bayesian models, PARSER (Perruchet & Vinter, 1998), and TRACX (French et al., 2011). Among other major models, the CBL (Chunk-Based Learning; e.g., McCauley & Christiansen, 2014) will be not considered, because despite its name, it is in fact a hybrid model: The initial formation of chunks relies on computations aimed at locating boundaries in the dips of the TP distribution (as earlier TP-based models, except that the CBL model exploits backward TPs instead of forward TPs).

Let us start from the observation above that selecting among all the possible partitioning of a linguistic corpus, the one fulfilling certain criteria results in identifying the words. We suggested as a first approximation that a relevant criterion could be the partitioning that requires the smallest number of different units. However, using this criterion alone would generally lead to the primitives (e.g., the letters in our example above, or the phonemes, the syllables, etc.), resulting in chunking failure. It has been often posited that what needs to be minimized is the sum of the codelength for the chunks (the lexicon), and the codelength for the input data when written using the lexicon. A statistical tool for assessing the best coding is the Minimum Description length (MDL) method. One of the first models of segmentation was based on the MDL criterion (Brent & Cartwright, 1996). The problem of the optimal segmentation may be formulated in a Bayesian framework. Starting from an hypothesis space that consists of all possible segmentation of the data, the Bayesian model of Goldwater, Griffiths, and Johnson (2009) is aimed at finding the segmentation with the highest probability, a higher probability being assigned to segmentations that contain relatively few word types, with each of which occurring frequently and comprising only a few primitives.

These models could not serve as plausible psychological models, let alone because the whole corpus needs to be stored and processed before extracting a single word. However, the MDL principles have been exploited in more realistic models in which learners pass through the data sequentially, without storing the sentences in memory. In the MDL-Chunker of Robinet et al. (2011), for instance, a new chunk is created on-line whenever the overall representation of the data when this chunk is used as a coding unit becomes simpler than before chunk creation. Likewise, the Bayesian Goldwater et al.'s model has been modified to process utterances one at a time (e.g., Frank et al., 2010; Meylan et al., 2012; Pearl, Goldwater, & Steyvers, 2011). Along the same lines, these studies have addressed the question of how learners may approximate Bayesian inference by using algorithms implementing known limitations of the cognitive apparatus. For instance, in one of the algorithms explored by Pearl et al. (2011), greater prominence is given to the data encountered more recently, in order to simulate the human memory decay process. In most studies, implementing processing and memory limitations did not substantially impair the performance of a model simulating an ideal learner.

By contrast, the primary motivation of PARSER (Perruchet & Vinter, 1998) is to account for optimal segmentation in terms of simple and ubiquitous psychological processes. Starting from the observation that, in humans, attentional coding naturally segments the ingoing information into small and disjunctive parts of variable length, the model encodes the input as a succession of provisional units comprising a random number of components (typically, between 1 and 3). These units are stored in a lexicon and their future depends on ubiquitous laws of memory: They are strengthened whenever they reoccur in the input, and otherwise, their strength vanishes as a consequence of spontaneous decay and/or interference with the processing of similar material. The selection of the units that are relevant to the structure of the language among all the (irrelevant) others operates as a natural consequence of memory laws. Decay eliminates the units that do not occur often enough, while interference makes the model sensitive to statistics such as bi-directional transitional probabilities between the unit components (see Perruchet & Poulin-Charronnat, 2012b, for a justification). Because perception is guided by internal representations, the learned chunks become new primitives, making the system able to build chunks whose components were not initially perceived in a single attentional focus.

Finally, TRACX (French et al., 2011) exploits a connectionist architecture which basically works as an autoassociator network, with one hidden layer comprising half of the units of the input and output layers. The to-be-learned sequence is displayed as successive pairs of elements, as in a moving window. As a consequence of learning through a back-propagation algorithm, the output error (i.e., the difference between input and output) becomes smaller whenever the same pair of elements reoccurs. If this error is below a preset threshold value, this is taken as evidence that the network "recognized" the pair as having previously occurred. In this case, the weights of the hidden unit are copied in the input layer on the next processing step. The end result is that by iterative accretion, TRACX learn to form chunks of a variable number of elements that are "recognized" as co-occurring. Mareschal and French (2017) present a version, TRACX 2, in which "recognition" is not an all-or-none process: The contribution of the hidden layer to the input is graded as a function of the output error.

Certainly one of the most surprising outcome of studies comparing different chunk-based models, and especially the Bayesian models to either PARSER or TRACX, is that overall, they turn out to make very similar predictions (Frank et al., 2011; Kurumada et al., 2013; Meylan et al., 2012; Robinet et al., 2011). This may be not as surprising as one might think, however. For instance, as mentioned above, the Bayesian models exhaustively examine all possible partitionings of a corpus or a given utterance, while PARSER relies on the variety generated by successive random drawings to provide provisional chunks. PARSER's algorithm certainly does not explore the entire hypothesis space, but approximates this objective. In both models, the segmentation problem is solved by some direct competition between different possible chunks, instead of being an inference on a continuous distribution of probabilities over syllables as in TP-based models. Moreover, ending up with the same chunks after selection is not unexpected. Indeed, for a given corpus and other things being equal, minimizing the number of different words (the primary target of Bayesian models) also maximizes the number of repetitions and the cohesiveness of each word. PARSER and TRACX exploit this logical corollary.

This analysis suggests that the differences between models could be more related to the choice of technical options than to their theoretical underpinnings. Several contributors to Bayesian literature (e.g., Goldwater et al., 2009; Pearl & Goldwater, 2016; Qian, Jaeger, & Aslin, 2016), referring to Marr (1982)'s framework, have emphasized that Bayesian models provide a *computational*-level account for learners' behavior, while most other models, either symbolic or connectionist, would be aimed at simulating the same behavior at the *algorithmic* level. Marr's framework provides an attractive solution to the heterogeneity of models. PARSER could be thought of as an algorithmic implementation of some Bayesian principles, and likewise, at a more fine-grained level, TRACX could be thought of as an algorithmic implementation of PARSER on some aspects. For instance, in PARSER, interference is assessed as the decrement of a numerical value that is assumed to measure the memory strength of a stored lexical unit. Decrement is proportional to the number of primitives shared with the currently process unit. Much more realistically, in TRACX, interference is a function of a similarity gradient consecutive to distributed representations.

Although the reference to Marr (1982)'s level of analysis certainly accounts for a part of the differences between chunk-based models of segmentation, whether this reference exhausts the issue is questionable, however. Let us consider another difference between Bayesian approaches and PARSER for illustration. Bayesian models examine what an ideal learner can learn from a given corpus, then they introduce what they construe as attentional and memory limitations, in order to process information in a more cognitively plausible way. By contrast, those alleged limitations are what allows PARSER to work. Assuming that the human mind would be endowed with boundless abilities would lead to behavior improvement in a Bayesian framework (at least theoretically), whereas this would make PARSER unable to learn anything. For instance, decay and interference are the processes that lead the learner to be sensitive to frequency and contingency and as a way of consequence, to discover the words. It appears quite difficult to account for so deep differences as a simple shift from a computational level to an algorithmic level. In fact, in the conceptual framework behind PARSER and TRACX, the very notion of *ideal learner,* essential for Bayesian models, makes no sense, because an optimal learning strategy may be defined only in regards of the actual learner's abilities.

## 5. Discussion and research agenda

Converging empirical data suggest that forming chunks when chunk boundaries are not given in the input does not proceed by computing the pairwise statistics allowing the discovery of those boundaries, as initially claimed in the word segmentation literature (e.g., Aslin & Newport, 2009; Saffran et al., 1996). Instead, most studies are consistent with the idea that a particular segmentation is selected among a number of other possibilities in a way that differs as a function of models, but in all cases relies on the intrinsic properties of the chunks.

As outlined in the Introduction, the notion of chunk is broader than that of word. Nevertheless, the explanatory power of chunking may be perceived as restricted to elementary matters. Regarding language, chunking processes seemingly leave aside lexical categorization and syntax. More generally, chunks appear to be far from exhausting all the distributional regularities in the world, and therefore they appear as only one among the multiple targets of statistical learning and may be the simplest one. This statement has the potential of challenging the conclusion above, because a class of processes that would be able to account for both chunking and the learning of higher structures appears to be more parsimonious than postulating different processes for each subdomain. Beyond the parsimony argument, the fact that individual differences in chunk extraction are predictive of children's comprehension of syntactic structures (e.g., Kidd & Arciuli, 2016) also prompts us to consider that lexical formation and syntax acquisition could rely, at least partially, on the same processes. Considering the models of chunking through this lens, connectionist and Bayesian approaches appear more promising than Parser. Connectionist models have been used in the service of syntax acquisition (e.g., Reali, Christiansen, & Monaghan, 2003; Williams, 2010) and, likewise, Bayesian modeling has been applied to discover abstract syntactic structures (e.g., Perfors, Tenenbaum, & Regier, 2011). By contrast, Parser, as parsimonious it may be for extracting chunks, is seemingly unable to go beyond this primary objective.

This view is not the only option, however. Instead of favoring models of chunk extraction that are ready-made for more complex processing, the objective of overall parsimony may be fulfilled in following the other way round, namely in extending the explanatory power of chunking processes beyond their conventional frontiers. Regarding language, for instance, a model of chunking coding the frequency of words does more than building a lexicon, because highly frequent words are helpful for the formation of grammatical categories (e.g., Frost et al., 2016). This alternative approach becomes still more persuasive whenever the conventional definition of chunks is broadened. To illustrate, let us consider the notion of frequent frames, as introduced by Mintz (2003). In this context, a frame is an ordered pair of words with any word intervening. Frequent frames form the basis of an early categorization strategy, because it has been shown that in many languages (for French, see Chemla, Mintz, Bernal, & Christophe, 2009), a given frame tends to surround words from a particular syntactic category. Admittedly, Parser is unable to account for the extraction of frequent frames, because it is devised to extract blocks of *consecutive* elements. However, this is only a superficial handicap: In both Parser and the frequent frames framework, the relevant elements, whether consecutive or not, are linked due to their joint attentional processing, and the selection of the most frequent units is due to the limited processing resources of the learner. What is illustrated through this example is that the principles underlying Parser can be extended to syntactical issues, provided that chunks comprising non-adjacent components are allowed. Another, still more crucial extension to the conventional notion of chunks would be the use of abstract components. Hamrick (2014) shows that when both an SRN and PARSER receive abstract categories as input, PARSER better captures the learning of syntactic sequences of a semi-artificial language than the SRN. Going a step further in the extension of chunks to complex cognitive units, Perruchet and Vinter (2002) proposed a general model of learning and development consisting in the progressive shaping of such units, which relies on the generalization of the processes involved in PARSER.

An objective for further research could be to assess the validity of thinking about the chunks as the simplest type of very general representational units, potentially able to capture an essential part of cognitive functioning. Another objective immediately follows, because it is obvious that large and composite units would need to be subsequently broken down on some occasions. In fact, this is also true for the conventional chunks. The multiword blocks need to be decomposed into words, words need to be analyzed into components in order to learn to read and write, and so on. Now, this process of disassembly is currently neglected. It should be noted that, although chunk formation is usually thought of as an implicit, non-intentional, and unsupervised process, the need for decomposition generally occurs in the context of explicit, often academic learning. Studying together chunk construction and chunk deconstruction appears as especially promising, and it should provide a gateway to a more general issue that has been inexplicably neglected from the early studies on implicit learning, namely the interactions between implicit and explicit forms of learning (see Batterink, Reber, & Paller, this issue).

## Note

1. Note that Aslin et al. (1998, note 3) acknowledged that backward TPs provided information about word boundaries in their languages.

## References


Adini, Y., Bonneh, Y. S., Komm, S., Deutsch, L., & Israeli, D. (2015). The time course and characteristics of procedural learning in schizophrenia patients and healthy individuals. *Frontiers in Human Neuroscience*, *9*, 475.

Arnon, I., McCauley, S. M., & Christiansen, M. H. (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. *Journal of Memory and Language*, *92*, 265–280.

Aslin, R. N., & Newport, E. L. (2009). What statistical learning can and can't tell us about language acquisition. In J. Colombo, P. McCardle, & L. Freund (Eds.), *Infant pathways to language: Methods, models, and research disorders* (pp. 15–29). New York: Psychology Press.

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. https://doi.org/10.1111/1467-9280.00063.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481.

Batterink, L., Reber, P. J., & Paller, K. A. (this issue). A memory-systems perspective on implicit and statistical learning. *Topics in Cognitive Science*.

Benitez, V. L., Yurovsky, D., & Smith, L. B. (2016). Competition between multiple words for a referent in cross-situational word learning. *Journal of Memory and Language*, *90*, 31–48.

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

Chemla, E., Mintz, T., Bernal, S., & Christophe, A. (2009). Categorizing words using "frequent frames": What cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science*, *12* (3), 396–406. https://doi.org/10.1111/j.1467-7687.2009.00825.x.

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, *9*, 542–551.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences.*, *36*, 181–204.

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities: Pitting abstract against stimulus specific representations. *Psychological Science*, *17*, 905–912.

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

Fernandes, T., Kolinsky, R., & Ventura, P. (2009). The metamorphosis of the statistical segmentation output: Lexicalization during artificial language learning. *Cognition*, *112*, 349–366.

Fiser, J. (2009). Perceptual learning and representational learning in humans and animals. *Learning & Behavior*, *2009*, 141–153.

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General*, *134*(4), 521–537.

Franco, A., & Destrebecqz, A. (2012). Chunking or not chunking? How do we find words in artificial language learning? *Advances in Cognitive Psychology*, *8*, 144–154.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107–125. https://doi.org/10.1016/j.cogni tion.2010.07.005.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*, 614–636. https://doi.org/10.1037/a0025255.

Frost, R. L. A., Monaghan, P., & Christiansen, M.H. (2016). Using statistics to learn words and grammatical categories: How high frequency words assist language acquisition. In A. Papafragou, D. Mirman & J. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (CogSci 2016) (pp. 81–86). Austin, TX: Cognitive Science Society.

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, *33*, 260–272. https://doi.org/10.1111/j.1551- 6709.2009.01012.x.

Glicksohn, A., & Cohen, A. (2011). The role of Gestalt grouping principles in visual statistical learning. *Attention, Perception, & Psychophysics*, *73*(3), 708–713.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

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

Hamrick, P. (2014). A role for chunk formation in statistical learning of second language syntax. *Language Learning*, *64*(2), 247–278.

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106.

Hoffman, Y., Perlman, A., Urtreger, B. O., Tzelgov, J., Pothos, E. M., & Edwards, D. J. (2017). Unitization of route knowledge. *Psychological Research*, *81*, 1241–1254. https://doi.org/10.1007/s00426-016-0811-0.

Jimenez, L. (2008). Taking patterns for chunks: Is there any evidence of chunk learning in continuous serial reaction-time tasks? *Psychological Research*, *72*(4), 387–396.

Jimenez, L., Méndez, A., Pasquali, A., Abrahamse, E., & Verwey, W. (2011). Chunking by colors: Assessing discrete learning in a continuous serial reaction-time task. *Acta Psychologica*, *137*(3), 318–329.

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning of word segmentation. *Developmental Science*, *13*(2), 339–345. https://doi.org/10.1111/j.1467-7687.2009.00886.x.

Jones, J., & Pashler, H. (2007). Is the mind inherently forward-looking? Comparing prediction and retrodiction. *Psychonomic Bulletin & Review*, *14*, 295–300.

Kibbe, M., & Feigenson, L. (2016). Infants use temporal regularities to chunk objects in memory. *Cognition*, *146*, 251–263.

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development*, *87*(1), 184–193. https://doi.org/10.1111/cdev.12461.

Kurumada, C., Meylan, S., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*, 439–453.

Mareschal, D., & French, R. M. (2017). TRACX2: A connectionist autoencoder using graded chunks to model infant visual statistical learning. *Philosophical Transactions of the Royal Society, B*, *372*, (1711) pii: 20160057.

Marr, D. (1982). *Vision*. Cambridge, MA: W. H. Freeman and Co.

McCauley, S. M., & Christiansen, M. H. (2014). Acquiring formulaic language: A computational model. *The Mental Lexicon*, *9*, 419–436.

Mersad, K., & Nazzi, T. (2012). When mommy comes to the rescue of statistics: Infants combine top-down and bottom-up cues to segment speech. *Language Learning and Development*, *8*, 303–315.

Meylan, S. C., Kurumada, C., Börschinger, B., Johnson, M., & Frank, M. C. (2012). Modeling online word segmentation performance in structured artificial languages. In N. Miyake, D. Peebles & R. P. Cooper (Eds.), *Building bridges across cognitive sciences around the world: Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (pp. 1–6). Austin, TX: Cognitive Science Society.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.

Mirman, D., Graf Estes, K., & Magnuson, J. S. (2010). Computational modeling of statistical learning: Effects of transitional probability versus frequency and links to word learning. *Infancy*, *15*(5), 471–486.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 2745–2750. https://doi.org/10.1073/pnas.0708424105.

Pearl, L., & Goldwater, S. (2016). Statistical learning, inductive bias, and Bayesian inference in language acquisition. In J. Lidz, W. Snyder, & J. Pater (Eds.), *Oxford handbook of developmental linguistics* (pp. 664–695). Oxford, UK: Oxford University Press.

Pearl, L., Goldwater, S., & Steyvers, M. (2011). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, *8*(2), 107–132.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009a). Statistical learning in a natural language by 8-month-old infants, *Child Development*, *80*, 674–685. https://doi.org/10.1111/j.1467-8624.2009.01290.

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

Perfors, A., Tenenbaum, J., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

Perlman, A., Hoffman, Y., Tzelgov, J., Pothos, E. M., & Edwards, D. J. (2016). The notion of contextual locking: Previously learnt items are not accessible as such when appearing in a less common context. *The Quarterly Journal of Experimental Psychology*, *69*(3), 410–431.

Perlman, A., Pothos, E. M., Edwards, D., & Tzelgov, J. (2010). Task relevant chunking in sequence learning. *Journal of Experimental Psychology: Human Perception & Performance*, *36*, 649–661.

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

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

Perruchet, P., & Poulin-Charronnat, B. (2012a). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*, 807–818. https://doi.org/10.1016/j.jml.2012.02.010.

Perruchet, P., & Poulin-Charronnat, B. (2012b). Word segmentation: Trading the (new, but poor) concept of statistical computation for the (old, but richer) associative approach. In P. Rebuschat, & J. N.

Williams (Eds.), *Statistical learning and language acquisition* (pp. 119–143). Berlin, Germany: De Gruyter Mouton.

Perruchet, P., Poulin-Charronnat, B., Tillmann, B., & Peereman, R. (2014). New evidence for chunk-based models in word segmentation. *Acta Psychologica*, *149*, 1–8. https://doi.org/10.1016/j.actpsy.2014. 01.015.

Perruchet, P., & Tillmann, B. (2010). Exploiting multiple sources of information in learning an artificial language: Human data and modeling. *Cognitive Science*, *34*, 255–285. https://doi.org/10.1111/j.1551-6709.2009.01074.x.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*, 246–263. https://doi.org/10.1016/jmla.1998.2576.

Perruchet, P., & Vinter, A. (2002). The self-organizing consciousness. *Behavioral and Brain Sciences*, *25*, 297–388.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*, 1112–1130.

Poulin-Charronnat, B., Perruchet, P., Tillmann, B., & Peereman, R. (2017). Familiar units prevail over statistical cues in word segmentation. *Psychological Research*, *5*, 990–1003.

Qian, T., Jaeger, T. F., & Aslin, R. N. (2016). Incremental implicit learning of bundles of statistical patterns. *Cognition*, *157*, 156–173.

Reali, F., Christiansen, M. H., & Monaghan, P. (2003). Phonological and distributional cues in syntax acquisition: Scaling up the connectionist approach to multiple-cue integration. In R. Alterman and D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 970–975). Mahwah, NJ: Lawrence Erlbaum.

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDLChunker: A MDL-based cognitive model of inductive learning. *Cognitive Science*, *35*, 1352–1389. https://doi.org/10.1111/j.1551-6709.2011.01188.x.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621. https://doi.org/10.1006/jmla.1996.0032.

Slone, L. K., & Johnson, S. P. (2015). Statistical and chunking processes in adults' visual sequence learning. In R. P. Cooper (Ed.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Pasadena, CA: Cognitive Science Society.

Slone, L. K., & Johnson, S. P. (2018). When learning goes beyond statistics: Infants represent visual sequences in terms of chunks. *Cognition*, *178*, 92–102.

Sohail, J., & Johnson, E. K. (2016). How transitional probabilities and the edge effect contribute to listeners' phonological bootstrapping success. *Language Learning and Development*, *12*(2), 105–115.

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.

Tomasello, M. (2009). The usage-based theory of language acquisition. In E. Bavin (Ed.), *Handbook of child language*. Cambridge, UK: Cambridge University Press.

Trueswell, J. C., Medina, T., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156.

Tummeltshammer, K. S., Amso, D., French, R. M., & Kirkham, N. Z. (2017). Across space and time: Infants learn from backward and forward visual statistics. *Developmental Science*, *20*(5), 1–9.

Williams, J. N. (2010). Initial incidental acquisition of word order regularities: Is it just sequence learning? *Language Learning*, *60*, 221–244.

Zellin, M., von Mühlenen, A., Müller, H., & Conci, M. (2014). Long-term adaptation to change in implicit contextual learning. *Psychonomic Bulletin & Review*, *21*(4), 1073–1079.