Transitional probabilities count more than frequency, but might not be used for memorization

Ansgar D. Endress
Department of Psychology, City, University of London, UK

Alan Langus
Cognitive Neuroscience Sector, International School for Advanced Studies, Trieste, Italy

Learners often need to extract recurring items from continuous sequences, in both vision and audition. The best-known example is probably found in word-learning, where listeners have to determine where words start and end in fluent speech. This could be achieved through universal and experience-independent statistical mechanisms, for example by relying on Transitional Probabilities (TPs). Further, these mechanisms might allow learners to store items in memory. However, previous investigations have yielded conflicting evidence as to whether a sensitivity to TPs is diagnostic of the memorization of recurring items. Here, we address this issue in the visual modality. Participants were familiarized with a continuous sequence of visual items (i.e., arbitrary or everyday symbols), and then had to choose between (i) high-TP items that appeared in the sequence, (ii) high-TP items that did *not* appear in the sequence, and (iii) low-TP items that appeared in the sequence. Items matched in TPs but differing in (chunk) frequency were much harder to discriminate than items differing in TPs (with no significant sensitivity to chunk frequency), and learners preferred unattested high-TP items over attested low-TP items. Contrary to previous claims, these results cannot be explained on the basis of the similarity of the test items. Learners thus weigh within-item TPs higher than the frequency of the chunks, even when the TP differences are relatively subtle. We argue that these results are problematic for distributional clustering mechanisms that analyze continuous sequences, and provide supporting computational results. We suggest that the role of TPs might not be to memorize items per se, but rather to prepare learners to memorize recurring items once they are presented in subsequent learning situations with richer cues.

Introduction

In many situations, we need to extract recurring units from continuous sequences. For example, we move continuously through space, but can separate the continuous motion into discrete actions (e.g., Newtson, 1973; Newtson, Engquist, & Bois, 1977; Zacks & Tversky, 2001; Zacks & Swallow, 2007). We can recognize motifs from hour-long symphonic works, and, while navigating, we experience a sequence of visual snapshots (e.g., of landmarks), that we can retrieve as a sequence from memory when we try to take the same trajectory again.

This class of problems has been studied most extensively in the context of language acquisition. To competent adult listeners, fluent speech seems to be composed of a discrete sequence of words, much as words are separated by white space in written text. However, fluent speech does not comprise the equivalent of white space. Hence, before infants can learn the meaning of any word, they first have to isolate them from fluent speech, and learn where words start and where they end. This problem is called the segmentation problem.

A mechanism that extracts recurring units must have a crucial property (Endress & Mehler, 2009b; Endress & Hauser, 2010): it must allow learners to store these units in memory. A particularly prominent class of cues that might have this property are distributional cues such as transitional probabilities (TPs). For example, with speech material, the transitional probability between two syllables is the conditional probability of encountering the second syllable given the first one. While

ADE was supported by grant PSI2012-32533 from Spanish Ministerio de Economía y Competitividad and Marie Curie Incoming Fellowship 303163-COMINTENT. We thank Fernando Genestar and Neus Colomer for assistance with data collection, and Antonia Stanojević for helpful comments on an earlier version of this manuscript. Contributions: ADE designed, implemented and analyzed the studies. Both authors contributed to the writing of the manuscript.

there is substantial evidence that human adults, infants and other animals are sensitive to TPs in a variety of modalities, including language and the visual modality (see, among many others, Aslin, Saffran, & Newport, 1998; Creel, Newport, & Aslin, 2004; Endress, 2010; Endress & Wood, 2011; Fiser & Aslin, 2002a, 2005; Glicksohn & Cohen, 2011; Hauser, Newport, & Aslin, 2001; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999; Saffran & Griepentrog, 2001; Toro & Trobalón, 2005; Turk-Browne, Jungé, & Scholl, 2005; Turk-Browne & Scholl, 2009), studies assessing whether the output of TP computations is stored in memory have yielded conflicting results (see below). Moreover, the degree to which TPs lead to memorization seems to depend on the participants' native language (see Endress & Mehler, 2009b; Langus & Endress, under review; Perruchet & Poulin-Charronnat, 2012, and below).

Here, we ask whether visual TPs provide segmented chunks that can be memorized. We use visual stimuli for two reasons. First, given that speakers of different languages behaved differently in previous experiments (Endress & Mehler, 2009b; Langus & Endress, under review; Perruchet & Poulin-Charronnat, 2012), we sought to investigate this issue in a domain that is unlikely to be influenced by prior knowledge of one's native language, and might thus give a relatively uncontaminated picture of how statistical learning operates in the absence of language-specific knowledge. Second, it is unknown whether statistical learning of visual sequences leads to memorization (though there is some evidence on this issue for spatial visual arrays; see below). This issue is important because statistical learning abilities seem relatively uncorrelated across domains (Frost, Armstrong, Siegelman, & Christiansen, 2015; Garcia, Hankins, & Rusiniak, 1974; Siegelman & Frost, 2015), and might, in some cases, have different properties in different domains.

The question of whether statistical learning leads to memorization of units also touches on a second important question: How do statistical learning mechanisms operate? In the word segmentation literature, two important classes of mechanisms have been proposed: bracketing and clustering (Goodsitt, Morgan, & Kuhl, 1993; see also Thiessen, Kronstein, & Hufnagle, 2013, for a review). Bracketing operates by inserting boundaries between words, and thus presumably requires additional mechanisms to place items in memory. Clustering mechanisms, in contrast, chunk syllables together, and thus create units that can be memorized directly. To use a visual analogy, letters that are part of the same word have two related properties: they have at least one spatially adjacent letter (unless they are single letter words), and they are adjacent to at most one white space or punctuation element. In principle, either property might be sufficient to find word boundaries: one can infer where words start or end by monitoring the statistical cohesiveness of the statistics within a chunk of letters (clustering), or one could posit the start of a new word once a boundary cue (such as white space) is encountered (bracketing). However, the distinction between bracketing and clustering has rarely been addressed in the visual modality.

Below, we will first ask whether statistical learning of visual sequences leads to memorization of recurring units. Based on these and earlier results, we will then provide general arguments and illustrative simulations to show that these results are problematic for distributional chunking mechanisms.

Statistical learning of recurring sequences

The most widely researched strategy to extract recurring units from sequences relies on distributional cues, notably on transitional probabilities (TPs). In the speech domain, TPs reflect the conditional probability of one syllable σ_1 following another syllable σ_2: $P(\sigma_2|\sigma_1) = P(\sigma_1\sigma_2)/P(\sigma_1)$, where $\sigma_1\sigma_2$ is a syllable string. For example, Saffran, Aslin, and Newport (1996) showed that seven-months old infants can track TPs across syllables. Specifically, infants were exposed to a concatenation of made-up words. As a result, syllables within words had high TPs, while syllables spanning a word-boundary had relatively lower TPs. Infants discriminated high-TP items from low-TP items. Following this seminal paper, there has been a wealth of demonstrations that infants and also non-human animals are sensitive to TPs not only for speech material, but also for a variety of other stimuli (Creel et al., 2004; Endress, 2010; Fiser & Aslin, 2001, 2002b; Hauser et al., 2001; Saffran, Aslin, & Newport, 1996; Saffran et al., 1999; Toro & Trobalón, 2005; Turk-Browne et al., 2005). Further, the idea that distributional cues in a variety of guises might be used for extracting recurring sequences has been implemented in numerous computational models (e.g., Batchelder, 2002; Brent & Cartwright, 1996; Christiansen, Allen, & Seidenberg, 1998; Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Orbán, Fiser, Aslin, & Lengyel, 2008; Perruchet & Vinter, 1998; Swingley, 2005).[1]

[1]This is not to say that statistical learning is the only language-universal mechanism that might support word segmentation. Rather, a number of other mechanisms have been proposed that might be effective in a variety of languages. The list includes mechanisms that segment words on the basis of known words (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005; Brent & Siskind, 2001; Mersad & Nazzi, 2012; see also

Statistical learning and memory

To learn recurring sequences, learners must have a way to place them in memory. While this condition has rarely been discussed in the literature (see Endress & Mehler, 2009b; Endress & Hauser, 2010), it is more constraining than it seems. For example, with speech items, participants are not only sensitive to forward TPs, but also to backward TPs (Perruchet & Desaulty, 2008; Pelucchi, Hay, & Saffran, 2009). That is, in a syllable sequence $\sigma_1\sigma_2$, participants are not only sensitive to TPs of the form $P(\sigma_2|\sigma_1)$, but also to TPs of the form $P(\sigma_1|\sigma_2)$. Moreover, at least in vision, if participants are familiarized with a sequence *ABC*, they are as good at recognizing *ABC* as *CBA* (Endress & Wood, 2011; Turk-Browne & Scholl, 2009). Unless one assumes that participants also memorize the *CBA* sequences although they never see them, it appears that successful discrimination of high-TP items from low-TP items does not imply that the high-TP items have been memorized. For example, a representation of a syllable or of a visual item might form associations with other representations of such items. However, in contrast to what is generally assumed in the segmentation literature, the representations that end up being associated with each other might not be integrated into a single memory representation.

At first sight, these conclusions seem to be contradicted by Graf-Estes, Evans, Alibali, and Saffran's (2007) results (see also Hay, Pelucchi, Graf Estes, & Saffran, 2011). These authors familiarized infants with a continuous speech stream that contained nonsense words defined by TPs. Following this familiarization, infants had to associate visual images with words (i.e., high-TP items), non-words or part-words (i.e., low-TP items). The results revealed better learning of image-sound associations when the sounds in the association were high-TP items than when they were low-TP items. These results seem to suggest that the output of the TP computations (i.e., high-TP items) has been stored in memory. However, there are two alternative possibilities. First, individual syllables might not only form associations with each other, but also with visual items; as a result, in high-TP items, the second order associations between individual syllables and visual items are stronger as well. That is, a syllable that strongly predicts another syllable also predicts the visual stimulus that is associated with the latter syllable, and these effects are stronger for high-TP items than for low-TP items.

The second alternative interpretation relies on the observation that the auditory items were presented in isolation during the sound-image association phase. Given that they were presented in isolation, they were presumably memorized as well. However, given the prior exposure to the speech stream, it might be easier to memorize high-TP items once they are presented in isolation or with other explicit boundary cues. This is because it is presumably easier to memorize sequences whose constituent syllables have stronger co-occurrence statistics.

This view is in line with results from visual memory, where memory for objects that are predictable in a scene (e.g., a pan in a kitchen) *appears* to better than for unpredictable objects because observers can reconstruct the objects from their world knowledge. In reality, however, memory is less precise for predictable objects because observers do not need to encode them precisely (e.g., Hollingworth & Henderson, 2000, 2003). For example, we don't need to remember a picture to know that a pan is likely to be in a kitchen, because it is associated with kitchens. If this view of the role of statistical computations in word learning is correct, the role of statistical computations in word segmentation would be to *prepare* learners to acquire words for when they encounter situations conducive to learning them (whatever these situations might be), but not to identify word candidates.

To our knowledge, there is no evidence that would allow us to choose between these possibilities. As such, it is currently unknown whether the output of distributional segmentation strategies can help listeners to form memory representations, such as those that will populate the mental lexicon. Alternatively, statistical computations might become useful only once word candidates (or their visual equivalents) have been identified by other means.

Can learners place the output of distributional mechanisms in memory?. Endress and Mehler (2009b) tested explicitly whether learners would place the output of distributional mechanisms into memory. As in other word segmentation experiments, (adult) participants were exposed to a random concatenation of made-up words. As a result, TPs between syllables within words were higher than TPs between syllables across words. Crucially, the words were constructed such that there were "phantom-words" that never occurred in the speech stream, but that had exactly the

---

Van de Weijer, 1999, but see Aslin, Woodward, LaMendola, & Bever, 1996), mechanisms that pay attention to the edges of utterances (i.e., their beginning and their end; see e.g., Seidl & Johnson, 2006, 2008; Shukla, Nespor, & Mehler, 2007), and universal aspects of prosody (e.g., Beckman & Pierrehumbert, 1986; Brentari, González, Seidl, & Wilbur, 2011; Endress & Hauser, 2010; Fenlon, Denmark, Campbell, & Woll, 2008; Nespor & Vogel, 1986; Pilon, 1981; Selkirk, 1984, 1986; for an overview see e.g., Cutler, Oahan, & van Donselaar, 1997; Langus, Marchetto, Bion, & Nespor, 2012; Shattuck-Hufnagel & Turk, 1996).

same TPs as the words that did occur in the speech stream (see Figure 1).

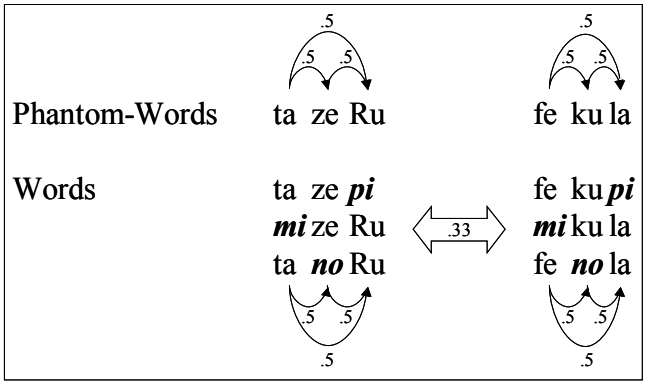

Figure 1. Design of the Endress and Mehler's (2009b) experiments. Participants were familiarized with continuous speech consisting of a concatenation of nonce words. These "words" were chosen such that TPs among syllables in words would be identical to TPs among syllables in "phantom-words", that is, in items that did not occur in the stream but had the same TPs as words. For each of the two phantom-words, there was a word sharing the first and the second syllable, a word sharing the second and third syllable, and a word sharing the first and third syllable. (The syllable that is *not* shared between a word and the corresponding phantom-word is printed in light gray characters in the figure.) In this way, TPs among adjacent and non-adjacent syllables within words and phantom-words were 0.5, and TPs among syllables across words 0.33. Reproduced from Endress and Mehler (2009b)

In different experiments, Endress and Mehler (2009b) asked participants to choose between words and phantom-words, words and part-words, and phantom-words and part-words. Part-words are low-TP items that straddle word boundaries but are attested in the speech stream.[2] Endress and Mehler (2009b) found three crucial results. First, participants preferred the phantom-words to part-words, even though they had heard part-words, but not phantom-words. Second, they had a much stronger preference for words over part-words than for words over phantom-words. Third, participants failed to choose words over phantom-words. These results show that participants are exquisitely sensitive to TPs. However, this sensitivity did not allow them to place words in memory (see also Aslin et al., 1998, for evidence that frequency of syllable groups is not necessary for preferring high-TP items to low-TP items). After all, if they had memorized the items, they should prefer words to phantom-words. Likewise, if they had memorized what they have heard, they should have preferred even part-words to phantom-words.

Interestingly, Ngon et al. (2013) found similar results in natural language acquisition. They showed that 11-months-old French learning infants prefer to listen to frequent syllable sequences compared to infrequent ones, even though neither of the sequences were French words. In contrast, they had no preference for real words compared to frequent syllable sequences, again suggesting that a preference for one item type over the other does not necessarily imply that the items have been placed in memory, though, as mentioned above, TPs might still help acquiring them.

Importantly, and as mentioned above, these results do not imply that TPs are not used for words learning. Rather, they do suggest a different role from what is generally considered. Specifically, if high-TP items are easier to learn than low-TP items, we suggest that TPs might, in line with our interpretation of Graf-Estes et al.'s (2007) data, prepare learners to acquire words once they are presented in more conducive situations for word learning.

However, it has been argued that participants' difficulty to reject phantom-words may be caused by non-statistical information. For example, it might be because words and phantom-words are similar. However, because words overlap with part-words to the same extent as with phantom-words, participants should be equally confused for either choice. Crucially, if participants were confused due to the overlap between words and phantom-words, they should be even more confused when familiarized with a speech stream comprising explicit boundary cues between words. Empirically, however, they readily prefer words to phantom-words under these conditions (Endress & Mehler, 2009b).

Relatedly, both Frank et al. (2010) and Perruchet and Poulin-Charronnat (2012) drew on research on categorization to suggest that phantom-words might be a "prototype," and that the actual words in the speech stream might be "distortions" of this prototype. If this view is correct, it is well known that learners readily recognize the prototype even if they have been trained only on distortions (e.g., Posner & Keele, 1968, 1970). As a result, they should recognize phantom-words as well.

However, this account somewhat misrepresents the literature. While participants clearly recognize the prototypes (e.g., Posner & Keele, 1968, 1970), they recognize exemplars they have seen better than the prototype, at least in an immediate test (e.g., Posner &

[2]They were constructed by either taking the last syllable of one word and concatenating it with the first syllable of the next word, or by concatenating the last two syllables of a word with the first syllable of the next word. As a result, these items had occurred in the speech stream, but had low TPs and straddled a (statistically defined) word boundary.

Keele, 1970). A similar conclusion follows from the false memory literature (Deese, 1959; Roediger & McDermott, 1995). If presented with a list of words that are semantically related to a prototype (e.g., words related to "sweet"), participants tend to think that they have encountered the prototype as well. Crucially, however, when prototypes are pitted against actual exemplars, participants readily prefer the actual exemplars over the prototype (Weinstein, McDermott, & Chan, 2010), contrary to what Frank et al.'s (2010) and Perruchet and Poulin-Charronnat's (2012) analogy suggests. While phantom-words might be prototypes in that they are similar to the actual words, the considerations above thus suggest that neither a prototype account nor a pure similarity account provide an adequate explanation of Endress and Mehler's (2009b) data.

However, the relative preferences for words, phantom-words and part-words depend to some extent on the participants' native language. Specifically, Italian and French speakers showed a higher preference for words over part-words than for words over phantom-words, suggesting that TPs carry more weight than frequency information (Endress & Mehler, 2009b; Langus & Endress, under review; Perruchet & Poulin-Charronnat, 2012). However, in contrast to Italian-speaking adults who failed to discriminate words from phantom-words, French-speaking participants preferred words to phantom-words (Perruchet & Poulin-Charronnat, 2012), and Spanish/Catalan bilinguals showed yet another pattern of results (Langus & Endress, under review).[3]

As a result, it is unclear under which conditions a sensitivity to item frequency is observed. We start investigating this issue in a language-neutral modality: vision.

Does distributional learning lead to memorization in vision? . The results discussed so far suggest that, in the verbal modality, statistical learning does not necessarily result in the memorization of syllable strings. In the visual modality, the situation is similarly mixed. As mentioned above, for sequences of visual objects, observers are as good at recognizing items played in forward-order as items played in backward-order, suggesting that success in a recognition test is not diagnostic of memorization. In contrast, for arrays of simultaneously presented visual objects (as opposed to object sequences), the evidence is mixed. It is well known that both adults and infants can extract co-occurrence statistics of simultaneously presented objects (e.g., Fiser & Aslin, 2001, 2002b, 2005). Fiser and Aslin (2005) and Orbán et al. (2008) also provided some evidence that shapes that are associated with each other through simultaneous presentations are also integrated into groups that might be stored in memory. Specifically, they proposed that upon learning a group, observers should become less sensitive to sub-groups, similar to how it is difficult to recognize the word "ham" in the group of syllables "hamster."[4] That is, it should be difficult to perceive the visual analogue of the sub-group "ham" once the visual analogue of the group "hamster" has been learned. However, the evidence for this proposal is mixed. For example, in Fiser and Aslin's (2005) Experiments 1 and 4, the prediction was supported, but not in their Experiment 5, nor in Slone and Johnson's (2015) Experiment 2 (though this experiment used sequential rather than simultaneous presentation). It is thus unclear to what extent distributional learning allows for the memorization of groups of shapes in vision.

Likewise, in an experiment with visual shapes, Slone and Johnson (2015) found that participants preferred the visual equivalent of words to the visual equivalent of phantom-words. However, in their experiments, shapes were presented one-by-one on a three-by-three grid, and each shape was associated with a unique position on the grid. Each shape loomed in for 750 ms, before the next shape (at a different grid location) loomed in. As a

[3]Perruchet and Poulin-Charronnat (2012) suggested that this difference between Italian- and French-speaking adults might be due to the intelligibility of the stimuli — speakers of both languages were tested with synthetized stimuli using a French voice. However, if the difference in performance between French speakers and Italian speakers were due to speech intelligibility, one would expect better performance in French speakers across the board. In contrast to this prediction, French-speaking participants showed an increase in performance only when discriminating words from phantom-words, but not when discriminating words from part-words. It is therefore possible that the difference between these results could also have emerged due to differences in how Italian and French-speaking adults process statistical regularities, especially if statistical computations over the speech signal depends on participants' previous experience with the statistical distribution of syllables in their native language. In line with this view, Langus and Endress (under review) successfully replicated Endress and Mehler's (2009b) finding with Italian speakers, but also found that Spanish/Catalan speakers preferred words to similar extents over part-words and phantom-words, showing a different pattern of results from both Italian speakers and French speakers (see Finn & Hudson Kam, 2008; Onnis, Monaghan, Richmond, & Chater, 2005; Mersad & Nazzi, 2011, for more predictable language-specific effects on statistical word segmentation).

[4]Such effects are found in the word recognition literature. However, they seem to be carried mainly by phonetic differences between syllables that are monosyllabic words and syllables that are parts of longer words, and might be due to the embedded word being suppressed (e.g., van Alphen & van Berkum, 2010; Salverda, Dahan, & McQueen, 2003; Shatzman & McQueen, 2006a, 2006b).

result, participants could not only rely on the sequential statistics as in auditory speech experiments (Endress & Mehler, 2009b; Perruchet & Poulin-Charronnat, 2012), but the sequencing of shapes also generated a pattern of motion. For example, if ABC was a word in the language, and if shape A appeared in the upper left corner, shape B in the middle square, and shape C in the upper right corner, the word ABC would also be associated with a V-shaped pattern of motion. These motion patterns might have helped participants discriminate words from phantom-words, because they were different between words and phantom-words. If so, they would be consistent with Endress and Mehler's (2009b) finding that words are preferred to phantom-words when additional cues are given.

The current experiments

In the experiments presented below, we provide a critical test of whether statistical learning in the visual modality leads to the memorization of recurring chunks. Specifically, we extend Endress and Mehler's (2009b) results to the visual modality using two stimulus sets: the non-sense shapes used by Fiser and Aslin (2002a), and the real-world objects used by Brady, Konkle, Alvarez, and Oliva (2008) (see Figure 2; while our stimuli are visual shapes, we refer to shape triplets as "words" in line with the auditory literature.)

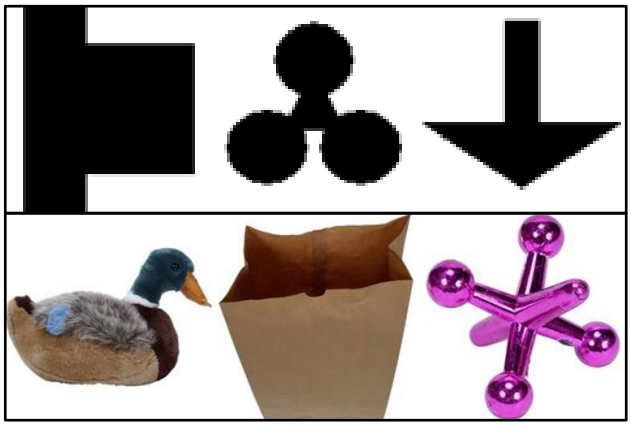

Figure 2. Example objects used in the current experiments. Top: objects from Fiser and Aslin (2002a). Bottom: objects from Brady et al. (2008)

We explore this issue in the visual modality for two reasons. First, we seek to circumvent the native language effects that appeared in this literature by using a modality that is plausibly less affected by the participants' native language. While visual statistical learning does not seem to correlate with auditory statistical learning (Siegelman & Frost, 2015), the underlying principles might be similar nonetheless (Frost et al., 2015). As a result, visual statistical learning is our best guess as to how statistical learning might operate in the auditory domain when people are not affected by prior experience with a language.

Second, given that visual statistical learning does not correlate with auditory statistical learning (Siegelman & Frost, 2015), it is important to find out how statistical learning "works" in the visual modality with respect to memorization.

The experiments are summarized in Table 1. To foreshadow our results, in Experiments 1 and 2, we found that words are preferred to part-words, words are not preferred to phantom-words, and phantom-words are preferred to part-words. However, we also found two results that were unpredicted (albeit with very small effect sizes): phantom-words were preferred to words, and the phantom-word vs. part-words discrimination was easier than the word vs. part-word discrimination.

In Experiment 3 and 4, we asked whether increased exposure would affect the results. However, contrary to expectation, performance did not even improve on the word vs. part-word discrimination, and became numerically worse.

In Experiments 5 and 6, participants were familiarized with visual streams where "words" were separated by blank screens. In the auditory domain, Endress and Mehler (2009b) showed that additional cues (e.g., from prosody) established a preference for words over phantom-words, and suggested that such cues might help participants memorizing actual items. Here, we found that performance improved, though we also found differences from the equivalent auditory experiments that most likely reflect true modality differences.

Based on these results, we will then address a critical question: *How* are recurring units placed in memory? The most relevant distinction comes from Goodsitt et al.'s (1993) classification of segmentation mechanisms into *bracketing* algorithms and *clustering* algorithms. Bracketing algorithms use cues (e.g., dips in TPs) to insert word boundaries between words (or their visual equivalent). Clustering algorithms use cues to join speech units (e.g., syllables) together, leading to their memorization. After presenting the experiments, we will show that TPs cannot carry more weight than frequency information for any clustering mechanism purely based on distributional information that places items in memory.

Table 1

Summary of the current experiments. In the "outcome" column, W stands for words, PW for part-words (i.e., low-TP items), and PhW for phantom-words. The "greater than", "smaller than" and "equality" signs stand for the presence or absence of statistically significant preferences or differences according to the main analyses. For example, (W > PW) > (W = PhW) *means that participants had a preference for words over part-words, had no significant preference for words over phantom-words, and that the preference for words over part-words was significantly greater than the preference for words over phantom-words.*

Exp.	# participants	Stimuli[a]	# Repetitions/ word	Break betw. words	Outcome[b]
1a,b	63	F&A, Brady	50	no	(W > PW) > (W < PhW)
2a	20	F&A	50	no	(W > PW) = (PhW > PW)
2b,c	62	F&A, Brady	50	no	(W > PW) = (PhW > PW)
2a,b,c	82	F&A, Brady	50	no	(W > PW) < (PhW > PW)
3a,b	41	F&A, Brady	100	no	(W = PW) = (W = PhW)
4	20	F&A	100	no	(W = PW) = (PhW = PW)
5a,b	55	F&A, Brady	50	yes	(W > PW) > (W > PhW)
6	13	Brady	50	yes	(W > PW) < (PhW > PW)

[a]F&A: Fiser and Aslin (2002a); Brady: Brady et al. (2008)

[b]W: Word; PW: Part-Word; PhW: Phantom-Word

Chunks vs. TPs with visual stimuli

Experiment 1: Are participants sensitive to frequency information?

In Experiment 1, participants were familiarized with a sequence of visual stimuli that conformed to the same statistical constraints as in Endress and Mehler's (2009b) experiments. Following this, they had to choose between words and part-words, and between words and phantom-words.

Materials and methods.

A note on sample sizes and the statistical approach. As shown in Table 1, the experiments differ in sample size. These differences arose for "historic" reasons. Specifically, in Experiments 1 to 5, we aimed for at least 20 or 30 participants per condition (depending on participant availability), and included the number of participants that could be recruited in a given period of time. Experiments 6 was aborted after a dozen participants after it became clear that the effects of interest were extremely unlikely to change with larger sample sizes.

For convenience, we will use null-hypothesis significance testing, despite its well-known drawbacks, including that some type I and II errors are expected when enough experiments are run. However, data are presented as scatter plots. Readers can thus perform their own intuitive Bayesian analysis by weighing their conclusion by their prior beliefs in them.

To provide evidence *for* null-hypotheses, we will use likelihood ratio analyses inspired by Glover and Dixon (2004). For example, to establish whether a condition differs from the chance level of 50%, we fit two models, assuming that the data is normally distributed. The *alternative* model estimates the variance and the mean from the data. In contrast, the null model estimates only the variance from the data, and sets the mean to 50%. As a result, the alternative model has one more parameter than the null model (i.e., the mean). We then use the fitted models to calculate the likelihood of the data, given the model. To account for the different numbers of parameters, we then use the likelihoods to calculate the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). We use these information criteria as corrected likelihoods to calculate likelihood ratios in favor of or against the null hypothesis. Our likelihood ratios are thus really ratios of AIC's and BIC's.

Participants. 31 Catalan/Spanish bilinguals (25 females, 6 males, mean age 19.9, 18-24) took part in Experiment 1a. 32 Catalan/Spanish bilinguals (22 females, 10 males, mean age 20.9, 18-31) took part in Experiment 1b. In these and all other experiments, partici-

pants took part in only one experiment from the current series.

Apparatus. Stimuli were presented on a Philips 109B CRT monitor at a resolution of 1280 × 960 pixels and a refresh rate of 60 Hz. The experiment was administered in a soundproofed booth and was run using the Matlab psychophysics toolbox (Brainard, 1997; Pelli, 1997).

Materials. The stimuli in Experiment 1a were the visual shapes used by Fiser and Aslin (2002a); the stimuli in Experiment 1b were pictures of real-world objects used by Brady et al. (2008). The specific stimuli were randomly selected for each participant.

For the familiarization sequence, visual "words" were constructed such that there would be items that would not appear in the sequences, but that would have exactly the same transitional probabilities (TPs) between the first and the second, the second and the third, and the first and the third shape as the words; for mnemonic purposes, we call these items phantom-words.

For constructing the words, we first selected two phantom words, and then chose the actually occurring words accordingly (see Figure 1). For the phantom-word ABC, we included in the stream the three words ABG, HBC and AJC, where each letter represents a shape. For the phantom-word DEF, the stream contained the three words DEG, HEF and DJF. In this way, TPs between adjacent or non-adjacent shapes within words (and phantom-words) were .5, and TPs across word boundaries were .33 on average (range: .28 — .39).

The familiarization sequence was created by randomly concatenating placeholders for the six words so that the set of images that formed words for an individual participant varied across participants without changing the statistical structure of the familiarization. Each word appeared 50 times in this sequence. This random sequence was the same for all participants.[5] However, the correspondence between placeholders and images was randomized for each participant. Each image was shown for 1 s with no blank between images, leading to a familiarization duration of 15 min. Stimuli were presented on a gray background (RGB code B4B4B4).

The Fiser and Aslin (2002a) stimuli were presented at a size of 136 × 136 pixels (approximately 4.1 cm × 4.1 cm), subtending a visual angle of $1.96^o \times 1.96^o$ at a typical viewing distance of 60 cm. The Brady et al. (2008) stimuli were presented at a size of 256 × 256 pixels (approximately 7.7 × 7.7 cm), subtending a visual angle of $3.67^o \times 3.67^o$ at a typical viewing distance of 60 cm.

Procedure. Participants were informed that they would see a sequence of visual shapes. Following this, they saw the stimuli one after the other.

Before the test phase, participants were informed that they would see pairs of sequences of shapes, and that they had to indicate which of these shape sequences was more likely to come from the familiarization sequence. The test triplets were presented one after the other, with a blank screen of 1 s between the triplets.

Pairs of test triplets could be of two types. In the first one, participants had to choose between words and phantom-words. Words and phantom-words could overlap either in their first and second, their first and third, or their second and their shape; each overlap type was represented equally in the test pairs.

In the remaining trials, participants had to choose between words and part-words. Part-words could be of two types that were equally represented in the test pairs; they could either comprise one shape of the first word and two shapes of the next word (type CAB, if the stream is represented as a shape sequence ABCABCABC...), or of two shapes of the first word and one shape of the second word (type BCA). Most part-words shared two shapes with the word they were presented with.

There were 12 Word vs. Phantom-Word test pairs, half with the word presented first, as well as 24 Word vs. Part-Word test pairs, half with the word presented first. In addition to a different random assignment between shapes and placeholders, test pairs were presented in a different random order for each participant, with the constraint of not having more than three trials in a row with the word as the first or the second item, and not having more than three trials in a row with the same type of comparison.

Responses were collected from pre-marked keys on the keyboard.

Results and discussion. As shown in Figure 3, participants preferred words over part-words significantly more than expected by chance ($M = 58.27\%$, $SD = 12.75\%$), $t(62) = 5.14$, $p < .0001$, Cohen's $d = .65$, $CI_{.95} = 55.05\%, 61.48\%$. A likelihood ratio analysis suggested that the alternative hypothesis was 70,403 times more likely than the null hypothesis.

As in Endress and Mehler (2009b), participants did not prefer words to phantom-words. Here, however, they preferred words significantly *less* than expected by chance ($M = 45.5\%$, $SD = 15.25\%$), $t(62) = 2.34$, $p = .023$, Cohen's $d = .29$, $CI_{.95} = 41.66\%, 49.34\%$. However, the likelihood ratio in favor of the alternative hypothesis was just 1.95.

This result is unexpected by any account; after all, there is no reason to prefer phantom-words to words.

[5]The number of repetitions per shape is less than in EM. However, Experiment 3 yielded numerically worse performance when each word was presented 100 times.

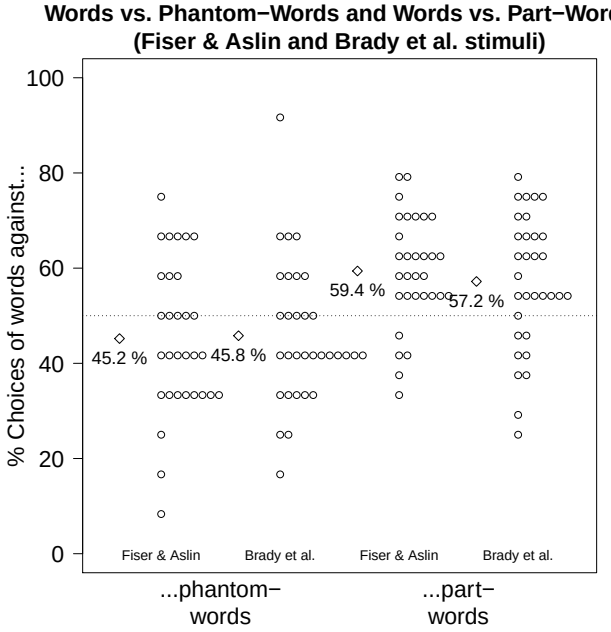

Figure 3. Results of Experiment 1. Circles represent individual participants and diamonds sample averages. Participants preferred words to part-words, but failed to prefer words to phantom-words.

Given that the stimuli were randomly selected for each participant, these results cannot be due to a preference for specific phantom-words. As a result, there remain two possibilities. First, these results might reflect a type I error. After all, the effect size of .29 is small, and visual inspection of Figure 3 reveals that the data with just 12 trials is rather discontinuous. In line with this interpretation, a multinomial test did not reach significance, $p = .0825$.[6] Alternatively, these results might be due to the single order of placeholders that was used to create the familiarization stream for all participants. Be that as it might, the current results replicate the finding that Italian participants do not prefer words to phantom-words. If the significant preference for phantom-words is due to the specific ordering of the placeholders and not a type I error, they would also show that statistical learning is sensitive to the specific order of items, which might create important problems if it were used in natural language acquisition. We will come back to this issue in the discussion of Experiment 2, where the concern of using only a single randomization of placeholders will be addressed.

To compare the test trial types, we performed an ANOVA with the within-participant predictor Trial Type (word vs. part-word or word vs. phantom-word) and the between-participant predictor Stimulus Set (Fiser and Aslin (2002a) vs. Brady et al. (2008)) as well as their interaction. The main effect of Trial Type reached significance, $F(1,61) = 29.51$, $p < .0001$, $\eta_p^2 = .325$, suggesting that it was significantly harder to reject phantom-words compared to words than part-words compared to words. This result replicates previous findings with both Italian and French speakers, and suggests that TP computations do not necessarily lead to the extraction of perceptual units. Finally, neither the main effect of Stimulus Set, $F(1,61) = .09$, $p = .77$, $\eta_p^2 = .001$, nor the interaction between these factors, $F(1,61) = .39$, $p = .537$, $\eta_p^2 = .004$, reached significance.[7]

Experiment 2: Do participants prefer unseen high-TP items to familiar low-TP items?

Experiment 2 was similar to Experiment 1, except that, during test, participants had to choose between words and part-words, and between phantom-words and part-words.

Materials and methods. As in Experiment 1, Experiment 2 used the stimuli from Fiser and Aslin (2002a) (Experiment 2b) and Brady et al. (2008) (Experiment 2c). In Experiment 2b and 2c, as in Experiment 1, the correspondence between image placeholders and images was randomized for each participant.

In Experiment 2a, the stimuli were again those used by Fiser and Aslin (2002a), but they were presented at a faster rate on a smaller screen. (This was not due to a design decision, but rather because Experiment 2a was originally a pilot experiment.) Crucially, the randomization across participants in Experiment 2a differed from that in Experiments 2b and 2c. Experiment 2a comprised two "languages" such that the words and phantom-words in Language 1 were part-words in Language 2 and vice versa. For these reasons, we will analyze Experiment 2a separately. The design of Experiment 2a is shown in Table 2.

Participants. 20 Catalan/Spanish bilinguals (17 females, 3 males, mean age 22.4, 19-29) took part in Experiment 2a, 30 Catalan/Spanish bilinguals (19 females, 11 males, mean age 22.5, 19-31) took part in

[6] In this analysis, we tabulated how many participants achieved different counts of correct responses, and compared this distribution to a binomial distribution with 12 trials and a success probability of .5, using Pearson's χ^2 to calculate the distance between the expected and actual distributions. This difference was estimated in 10^6 Monte Carlo trials using the EMT *R* package, version 1.1 (https://cran.r-project.org/web/packages/EMT/).

[7] We also analyzed word vs. part-word trials in an ANOVA with the between-participant predictor Stimulus Set and the within-participant predictor Part-word Type (BCA vs. CAB), but did not observe any significant main effect or interaction.

Table 2
Design of Experiment 2a

	L1		L2	
Phantom-words	ABC	DEF	CDE	FAB
Words	ABG	DEG	CDJ	FAJ
	HBC	HEF	GDE	GAB
	AJC	DJF	CHE	FHB

Experiment 2b, and 32 Catalan/Spanish bilinguals (21 females, 11 males, mean age 20.8, 18-26) took part in Experiment 2c.

In Experimentr 2b, two additional participants were excluded from analysis. One asked for instructions during the test phase, and the first few trials were answered by the experimenter, and thus randomly. The second participant was excluded for walking out of the test booth during the experiment or to software crashes, but we did not record which of these two reasons applied to the participant.

Apparatus. Experiment 2a was administered on a Macbook Pro in a soundproofed booth using Psyscope X (http://psy.ck.sissa.it). The apparatus for Experiments 2b and 2c was the same as in Experiment 1.

Materials. In Experiment 2a, each participant was familiarized with one of two familiarization streams, corresponding to the two languages described above. Shapes were taken from Fiser and Aslin (2002a). They were concatenated using the catmovie utility from the QTCoffee package (http://www.3am.pair.com/QTCoffee.html). The concatenation was saved using the H.264 codec and the mov container format with a frame rate of 4 frames/s. Shapes were presented at a rate of 750 ms per shape and had a size of 68 × 68 pixels. However, using Psyscope, the presentation rate was slowed down to about 1 s per shape (940 ms), and the image size was scaled to 200 × 200 pixels.

The materials for Experiments 2b and 2c were the same as for Experiment 1, except that a different random ordering of placeholders for words during familiarization was used. This ordering was the same for all participants, but, as in Experiment 1, the correspondence between placeholders and actual images was randomized for each participant.

Procedure. The procedure was identical to that in Experiment 1, except that participants had to choose between words and part-words (24 trials) and between phantom-words and part-words (8 trials). As in Experiment 1, each test item type occurred as the first item on half of the trials. Trials were randomized for each participant with the same constraint as in Experiment 1.

Results and discussion.

Experiment 2a. As shown in Figure 4, participants in Experiment 2a preferred words to part-words (M = 58.12%, SD = 12.28%), $t(19) = 2.96$, $p = .008$, Cohen's $d = .66$, $CI_{.95}$ = 52.38%, 63.87%. The likelihood ratio in favor of the alternative hypothesis was 17.9. Participants also preferred phantom-words to part-words (M = 64.38%, SD = 18.26%), $t(19) = 3.52$, $p = .0023$, Cohen's $d = .79$, $CI_{.95}$ = 55.83%, 72.92%. The likelihood ratio in favor of the alternative hypothesis was 109.8.

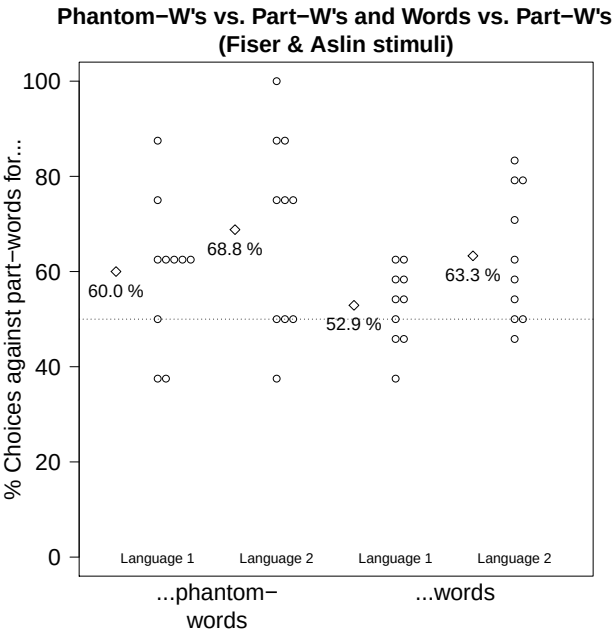

Figure 4. Results of Experiment 2a. Circles represent individual participants and diamonds sample averages. Participants preferred words to part-words and also phantom-words to part-words.

An ANOVA with Trial Type (word vs. part-word or phantom-word vs. part-word) as within-participant predictor and Language as between-participant predictor yielded no main effect of Language, $F(1,18) = 2.72$, $p = .117$, $\eta_p^2 = .131$, nor of Trial Type, $F(1,18) = 3.21$, $p = .090$, $\eta_p^2 = .151$, nor an interaction between these factors, $F(1,18) = .06$, $p = .814$, $\eta_p^2 = .003$.

Experiments 2b and 2c. As shown in Figure 5, the performance in word vs. part-word trials was relatively poor (M = 53.36%, SD = 11.44%), yet better than expected by chance, $t(61) = 2.31$, $p = .024$, Cohen's $d = .29$, $CI_{.95}$ = 50.46%, 56.26%. However, the likelihood ratio for the non-null hypothesis was just 1.84.

In phantom-word vs. part-word trials, participants preferred phantom-words over part-words (M = 57.86%, SD = 21.38%), $t(61) = 2.9$, $p = .005$, Cohen's $d = .37$, $CI_{.95}$ = 52.43%, 63.29%. The likelihood ratio for the

non-null hypothesis was 8.41.

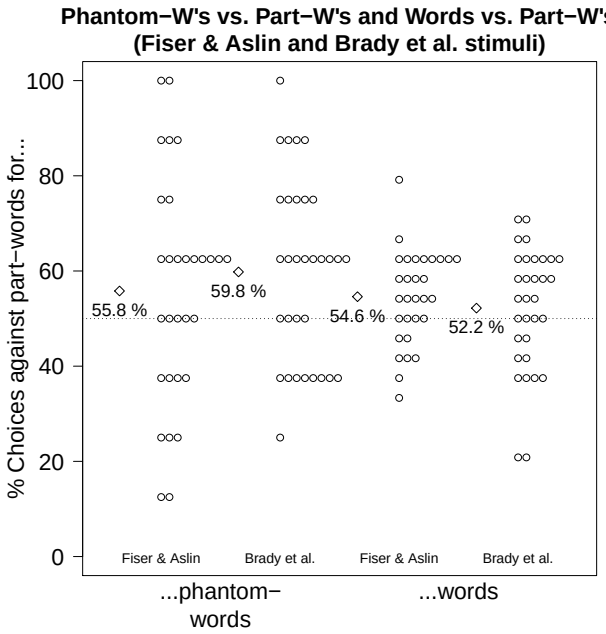

Figure 5. Results of Experiments 2b and c. Circles represent individual participants and diamonds sample averages. Participants preferred words to part-words and also phantom-words to part-words.

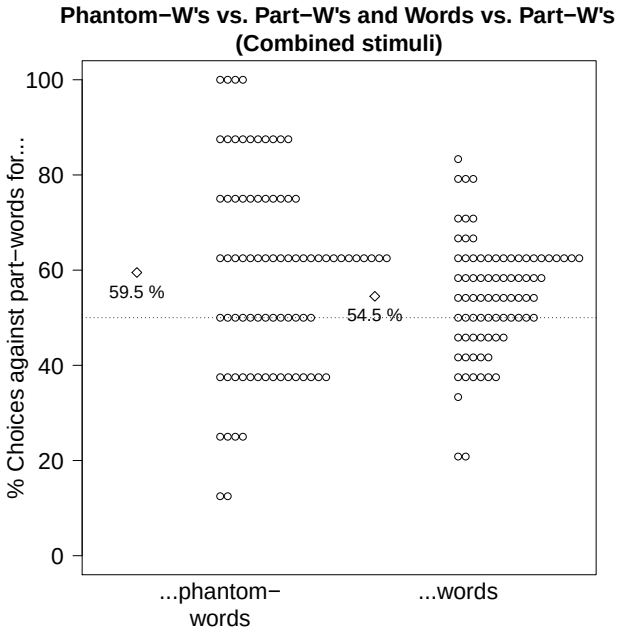

Figure 6. Results of Experiment 2a through c. Circles represent individual participants and diamonds sample averages. Participants preferred words to part-words and also phantom-words to part-words.

In an ANOVA with the within-participant predictor Trial Type (word vs. part-word or phantom-word vs. part-word) and the between-participant predictor Stimulus Set (Fiser and Aslin (2002a) vs. Brady et al. (2008)) as well as their interaction, neither the main effect of Stimulus Set, $F(1,60) = .05$, $p = .824$, $\eta_p^2 < .001$, nor the interaction, $F(1,60) = 1.45$, $p = .233$, $\eta_p^2 = .023$, reached significance. However, the main effect of Trial Type approached significance, $F(1,60) = 2.97$, $p = .09$, $\eta_p^2 = .046$. We will discuss this result below.

Combined analysis. In a combined analysis of Experiments 2a through 2c, participants preferred words to part-words ($M = 54.52\%$, $SD = 11.75\%$), $t(81) = 3.48$, $p = .0008$, Cohen's $d = .38$, $CI_{95} = 51.94\%, 57.10\%$ (see Figure 6). While the likelihood ratio for the non-null hypothesis was 47.8, performance was still relatively poor.

Participants also preferred phantom-words over part-words ($M = 59.45\%$, $SD = 20.75\%$), $t(81) = 4.13$, $p < .0001$, Cohen's $d = .46$, $CI_{95} = 54.89\%, 64.01\%$. The likelihood ratio in favor of the non-null hypothesis was 547.[8]

In an ANOVA with the within-participant predictor Trial Type (word vs. part-word or phantom-word vs. part-word) and the between-participant predictor Stimulus Set (Experiments 2a and 2b vs. Experiment 2c) as well as the interaction, neither the main effect of Stimulus Set, $F(2,79) = 1.30$, $p = .279$, $\eta_p^2 = .032$, nor the interaction, $F(2,79) = .9$, $p = .420$, $\eta_p^2 = .020$, reached significance. Surprisingly, however, the main effect of Trial Type reached significance, $F(1,79) = 5.28$, $p = .024$, $\eta_p^2 = .061$. According to this result, the preference for phantom-words over part-words was greater than the preference for words over part-words.

To assess the reliability of the surprising effect of Trial Type, we performed a number of follow-up analyses (see Appendix A). These analyses suggest that the effect of Trial Type is small, and carried by a subset of the participants. However, before dismissing this small effect as a type I error, it should be noted that about a third of participants generally perform around the chance level in typical statistical learning experiments (Frost et al., 2015). The current results thus reveal a small effect that is relatively consistent across experiments, but has no obvious explanation.

Discussion. Experiments 1 and 2 replicate the crucial results previously obtained with Italian speakers: the word vs. phantom-word discrimination is harder than the word vs. part-word one, phantom-words are

[8]We did not analyze word vs. part-word trials as a function of the part-word type because we did not record the part-word types.

preferred to part-words, and words are not preferred to phantom-words (Endress & Mehler, 2009b).

However, Experiments 1 and 2 also reveal two unexpected results. In Experiment 1, participants significantly prefer phantom-words to words, and in Experiment 2, the phantom-word vs. part-word discrimination is easier than the word vs. phantom-word discrimination, although the effect size was rather small in both cases.

There are three possible explanations of these results. First, they might be type I errors. After all, the effect sizes are rather small, and the bootstrap analysis of Experiment 2 (see Appendix A) shows that they are so small that they would probably not be detected in the smaller sample sizes used in typical statistical learning experiments. Further, the effect sizes observed here are somewhat smaller than in comparable earlier experiments, perhaps because the differences in TPs between words and part-words were relatively subtle. For example, Fiser and Aslin (2002a) observed an effect size (Cohen's d) on the word vs. part-word discrimination with the same material as used here of 1.7 when words were pitted against part-words of type *CAB*, and of .4 when words were pitted against part-words of type *BCA* (though both effects relied on only 8 participants). However, in the current experiments (that used both *CAB* and *BCA* part-words), the effect sizes reached .65 in Experiment 1, and .66 in Experiment 2a, and .29 in Experiment 2b and c. Performance thus seems to be somewhat worse than in previous studies. However, the fact that the surprising results were relatively systematic across Experiments 1 and 2 makes a type I error somewhat less likely.

Second, in each experiment, only a single randomization of the sequence of image placeholders was used (while the correspondence between the placeholders and the actual images were randomized for each participant). Possibly, these randomizations made it particularly easy to recognize phantom-words even though they never appeared in the familiarization streams, though it is entirely unclear which features of the randomizations might have had such an effect. However, Experiment 1, Experiments 2b and c and Language 1 of Experiment 2a each used different randomizations. As a result, it would be somewhat surprising if each of these randomizations had independently made phantom-words easier to recognize.

Third, the effect might have arisen due the use of test lists with two different trial types (words vs. part-words and words vs. phantom-words in Experiment 1, and words vs. part-words and phantom-words vs. part-words in Experiment 2). However, this explanation is rather unlikely as well, and requires additional assumptions.[9] Be that is it might, in studies where French speakers were tested with auditory stimuli, experiments with just one test trial type and experiments with two test trial types yielded undistinguishable results (Perruchet & Poulin-Charronnat, 2012). Because none of the different possible explanations of these surprising results seems particularly convincing, we tentatively conclude that they probably represent type I errors.

Be that as it might, the crucial results of Experiments 1 and 2 are that the word vs. phantom-word discrimination is harder than the word vs. part-word one (with an effect size of $\eta_p^2 = .325$), that phantom-words are preferred to part-words (with an effect size of Cohen's $d = .46$), and that words are not preferred to phantom-words. These results thus support the conclusions that a sensitivity to TPs does not imply that TPs can be used to store words in memory (Endress & Mehler, 2009b). Further, these results suggest that even relatively weak differences in TPs count more than differences in chunk frequency.

Experiment 3: Are participants sensitive to frequency information with twofold exposure?

In Experiments 1 and 2, the participants' performance was relatively poor even on word vs. part-words trials. In Experiments 3 and 4, we attempted to improve performance by playing the familiarization streams twice. In Experiment 3, participants were tested on word vs. part-words trials and word vs. phantom-word trials. In Experiment 4, they were tested on word vs. part-words trials and phantom-word vs. part-word trials.

Materials and methods. Experiment 3 was identical to Experiment 1, except that the familiarization stream was played twice, resulting in 100 repetitions per word. The stimuli in Experiments 3a and 3b were taken

[9] For example, if we assume that, in Experiment 1, the randomization of a participant's test list started with a relatively high proportion of word vs. part-word trials, participants might have used the following reasoning upon encountering a word vs. phantom-word trial. Given that they already accepted the word on previous trials, they should not accept it in the current (word vs. phantom-word) trial, and choose phantom-words instead. However, the same kind of reasoning leads to the opposite pattern in Experiment 2: if they have rejected part-words in word vs. part-word trials, they should *accept* them in phantom-word vs. part-word trials, thus lowering performance on phantom-word vs. part-word trials. To the extent that participants can entertain such strategies, it is thus questionable whether a strategy exists that could explain the results of both Experiments 1 and 2.

from Fiser and Aslin (2002a) and Brady et al. (2008), respectively.

21 Catalan/Spanish bilinguals (14 females, 7 males, mean age 22.0, 17-35) took part in Experiment 3a. 20 Catalan/Spanish bilinguals (14 females, 6 males, mean age 19.9, 18-26) took part in Experiment 3b.

Results and discussion. As shown in Figure 7, participants showed no preference for words over part-words (M = 52.24%, SD = 11.3%), $t(40) = 1.27$, $p = .213$, Cohen's $d = .2$, $CI_{.95}$ = 48.67%, 55.8%. The likelihood ratio in favor of the null hypothesis was 2.87.

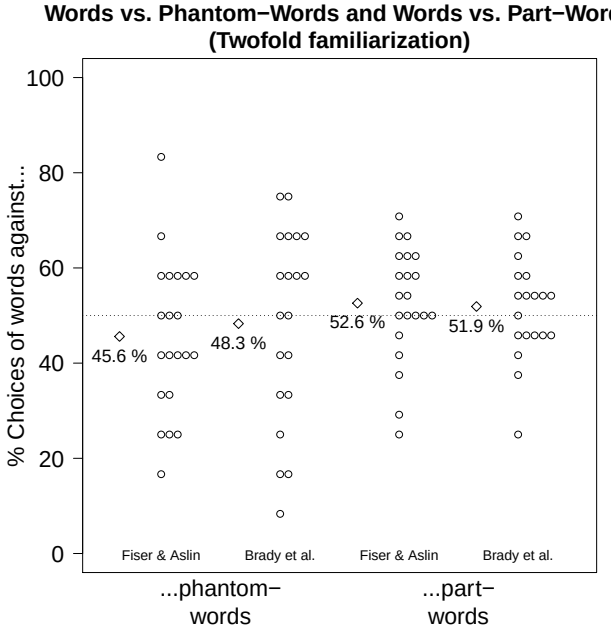

Figure 7. Results of Experiment 3. Circles represent individual participants and diamonds sample averages. Compared to Experiment 1, the exposure to the familiarization stream was doubled.

Participants showed no preference for words over phantom-words either (M = 46.95%, SD = 18.14%), $t(40) = 1.08$, $p = .288$, Cohen's $d = .17$, $CI_{.95}$ = 41.23%, 52.68%. The likelihood ratio in favor of the null hypothesis was 3.59. An ANOVA with the within-participant predictor Trial Type (word vs. part-word or word vs. phantom-word) and the between-participant predictor Stimulus Set (Fiser and Aslin (2002a) vs. Brady et al. (2008)) as well as their interaction yielded neither a main effect of Trial Type, $F(1,39) = 2.61$, $p = .114$, $\eta_p^2 = .0623$, nor of Stimulus Set, $F(1,39) = .08$, $p = .776$, $\eta_p^2 = .002$, nor an interaction between these factors, $F(1,39) = .27$, $p = .606$, $\eta_p^2 = .006$.[10]

Comparing Experiment 1 and 3, an ANOVA with the within-participant predictor Trial Type (word vs. part-word or word vs. phantom-word) and the between-participant predictors Stimulus Set (Fiser and Aslin (2002a) vs. Brady et al. (2008)) and Familiarization Duration (50 vs. 100 repetitions per word) as well as all interactions yielded only a main effect of Trial Type, $F(1,100) = 26.15$, $p < .0001$, $\eta_p^2 = .2$, and a marginal Familiarization Duration by Trial Type interaction, $F(1,100) = 3.63$, $p = .060$, $\eta_p^2 = .028$. A two-fold familiarization therefore failed to strengthen statistical computations over visual stimuli.

Experiment 4: Do participants prefer unseen high-TP items to familiar low-TP items with twofold exposure?

Materials and method. Experiment 4 was identical to Experiment 2a, with three exceptions. First, and crucially, the familiarization stream was played twice, yielding 100 repetitions per word. Second, shapes were presented at a rate of about 716 ms per shape (as opposed to 1 s/shape), and had a size of 68 × 68 pixels (as opposed to 136 × 136 pixels). This was not a design decision based on the results of Experiments 1 and 2 but rather a consequence of the fact that this experiment was an initial pilot experiment.

20 Catalan/Spanish bilinguals (11 females, 9 males, mean age 24.6, 18-44) took part in Experiment 4. Two additional participants were excluded from analysis. One walked out of the test booth to ask for instructions during the familiarization, and missed at least one minute of the familiarization stream. For the other participant, the computer crashed before reaching the test phase.

Results. As shown in Figure 8, participants had no preference for words over part-words (M = 54.79%, SD = 15.9%), $t(19) = 1.35$, $p = .193$, Cohen's $d = .3$, $CI_{.95}$ = 47.35%, 62.23%. The likelihood ratio in favor of the null hypothesis was 1.80. Participants had no preference for phantom-words over part-words either (M = 49.38%, SD = 21.64%), $t(19) = .13$, $p = .899$, Cohen's $d = .029$, $CI_{.95}$ = 39.25%, 59.5%. The likelihood ratio in favor of the null hypothesis was 4.43.

An ANOVA with the within-participant factor Trial Type (word vs. part-word or word vs. phantom-word) and the between-participant factor Language yielded no main effect of Trial Type, $F(1,18) = 2.24$, $p = .152$, $\eta_p^2 = .11$, nor of Language, $F(1,18) = .01$, $p = .918$, $\eta_p^2 = .0006$, nor an interaction, $F(1,18) = .01$, $p = .91$,

[10] We also analyzed word vs. part-word trials in an ANOVA with the between-participant predictor Stimulus Set and the within-participant predictor Part-word Type (BCA vs. CAB), but did not observe any significant main effect or interaction.

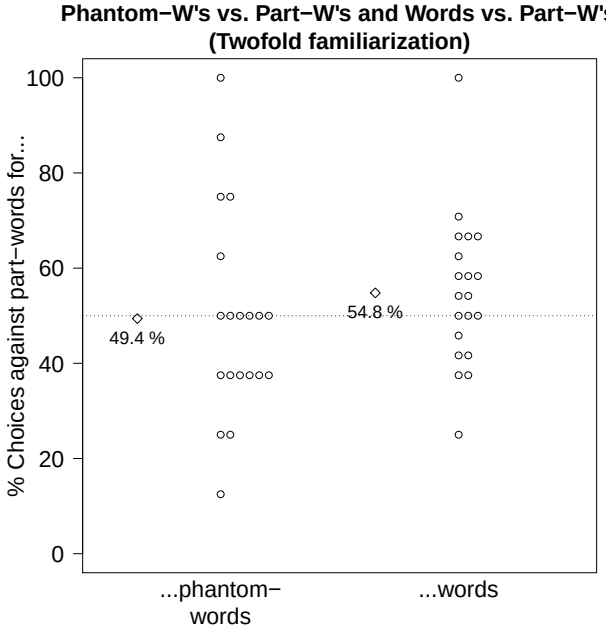

Figure 8. Results of Experiment 4. Circles represent individual participants and diamonds sample averages. Compared to Experiment 2, the exposure to the familiarization stream was doubled.

$\eta_p^2 = .0007$.[11]

Comparing Experiments 2a and 4, an ANOVA with the within-participant factor Trial Type (word vs. part-word or word vs. phantom-word) and the between-participant factors Language and Familiarization Duration yielded a significant Test Type by Familiarization Duration interaction, $F(1,36) = 5.38$, $p = .026$, $\eta_p^2 = .13$, as well as a marginal main effect of Familiarization Duration, $F(1,36) = 3.47$, $p = .071$, $\eta_p^2 = .084$.

Discussion. The results of Experiment 4 are unexpected. While increasing the exposure improved sensitivity to statistical cues in other experiments (e.g., Endress & Bonatti, 2007; Peña, Bonatti, Nespor, & Mehler, 2002), doubling the familiarization duration in Experiment 4 worsened participants' performance considerably. There is no immediate explanation for these results, except that the TP differences between words and part-words were relatively subtle in the current experiments. As a result, participants might learn the associations so well that they can longer discriminate items on the basis of the strength of TPs if they have been familiarized for too long with these items. Alternatively, and as discussed in the General Discussion, statistical computations might not be as robust as they are often believed to be.

Experiment 5: Are participants sensitive to frequency information after segmented familiarizations?

In the auditory modality, (Italian) participants preferred words to phantom-words when the familiarization streams contained explicit segmentation marks, such as short silences between words or lengthened word-final syllables (Endress & Mehler, 2009b). Endress and Mehler (2009b) concluded that such additional cues (that might be provided by prosody in real speech) help learners memorize actual items. In Experiments 5 and 6, we ask whether these results transfer to the visual modality. We thus assess whether the word vs. phantom-word discrimination would improve when words are separated by blank screens during familiarization.

Materials and methods. Experiment 5 was identical to Experiment 1, except that words in the familiarization stream were separated by a blank screen of 1 s. The stimuli in Experiment 5a were taken from Fiser and Aslin (2002a), and those in Experiment 5b from Brady et al. (2008).

30 Catalan/Spanish bilinguals (19 females, 11 males, mean age 20.7, 17-23) took part in Experiment 5a. 25 Catalan/Spanish bilinguals (20 females, 5 males, mean age 20.3, 18-25) took part in Experiment 5b.

Results. As shown in Figure 9, participants preferred words over part-words ($M = 90.76\%$, $SD = 12.8\%$), $t(54) = 23.62$, $p < .0001$, Cohen's $d = 3.2$, $CI_{.95} = 87.3\%, 94.22\%$. The likelihood ratio in favor of the alternative hypothesis was 1.88×10^{120}. They also had a marginal preference for words over phantom-words ($M = 56.36\%$, $SD = 24.11\%$), $t(54) = 1.96$, $p = .055$, Cohen's $d = .26$, $CI_{.95} = 49.85\%, 62.88\%$. However, the likelihood ratio in favor of the null hypothesis was 1.09.

An ANOVA with the within-participant predictor Trial Type (word vs. part-word or word vs. phantom-word) and the between-participant predictor Stimulus Set (Fiser and Aslin (2002a) vs. Brady et al. (2008)) as well as their interaction yielded a main effect of Test Type, $F(1,53) = 109.2$, $p < .0001$, $\eta_p^2 = .673$, but not of Stimulus Set, $F(1,53) = 2.28$, $p = .137$, $\eta_p^2 = .041$, nor an interaction, $F(1,53) = .159$, $p = .691$, $\eta_p^2 = .001$.[12]

[11] We did not analyze word vs. part-word trials as a function of the part-word type because we did not record the part-word types.

[12] We also analyzed word vs. part-word trials in an ANOVA with the between-participant predictor Stimulus Set and the within-participant predictor Part-word Type (BCA vs. CAB), but did not observe any significant main effect or interaction.

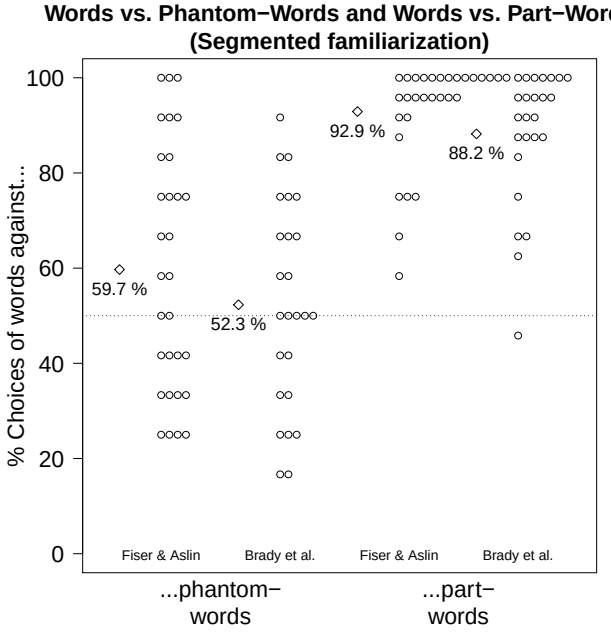

Figure 9. Results of Experiment 5. Circles represent individual participants and diamonds sample averages. In contrast to Experiment 1, the familiarization stream contained blank screens of 1 s after each word.

Comparing Experiment 1 and 5, an ANOVA with the within-participant predictor Trial Type (word vs. part-word or word vs. phantom-word) and the between-participant predictors Stimulus Set (Fiser and Aslin (2002a) vs. Brady et al. (2008)) and Familiarization Type (continuous vs. segmented) as well as all interactions yielded main effects of Test Type, $F(1,114) = 133$, $p < .0001$, $\eta_p^2 = .48$, and Familiarization Type, $F(1,114) = 84.56$, $p < .0001$, $\eta_p^2 = .419$, as well as an interaction between these factors, $F(1,114) = 29.66$, $p < .0001$, $\eta_p^2 = .107$. Crucially, however, performance improved between Experiments 1 and 5 both for word vs. part-word tests, $F(1,114) = 191$, $p < .0001$, $\eta_p^2 = 0.621$, and for word vs. phantom-word tests, $F(1,114) = 8.8$, $p = 0.004$, $\eta_p^2 = 0.07$.

Discussion. In the familiarization streams of Experiment 5, words were separated from one another by blank screens of 1 s. Compared to a continuous familiarization, these segmentation cues boosted performance in the word vs. part-word trials. In contrast, the preference for words over phantom-words was significantly improved as well, although it remained marginal (though visual inspection of Figure 9 revealed that the poor performance was driven by the condition with the Brady et al. (2008) stimuli).

These results are consistent with those of Endress and Mehler's (2009b) Experiments 3 and 4 in that including explicit segmentation cues led to a preference for words over phantom-words. However, they differ from these results in that performance on word vs. part-word trials improved much more than performance on word vs. phantom-word trials. In Endress and Mehler's (2009b) experiments, the improvement was similar in both trial types. However, visual inspection of Figure 9 suggests that those participants exposed to the stimulus set taken from Fiser and Aslin (2002a) performed rather similarly to Italian participants in Endress and Mehler (2009b) Experiments 3 and 4.

To the extent that the pattern of results in Experiment 5 is different from that in Endress and Mehler's (2009b) Experiments 3 and 4, the difference probably reflects a true modality difference. Specifically, in both the auditory and the visual modality in humans (Endress & Bonatti, 2007; Endress & Mehler, 2009a; Endress & Wood, 2011), and in non-human primates (Endress, Carden, Versace, & Hauser, 2010), a familiarization with streams containing explicit segmentation cues can lead participants to accept items that have onsets and offsets that have occurred in these positions during familiarization, even if the test item has not been encountered during familiarization. Given that phantom-words have correct onsets and offsets, they might be relatively hard to reject for this reason after a segmented familiarization.

Crucially, at least in the auditory modality, humans readily prefer actual words to items like phantom-words that have only correct edge syllables (see Endress and Bonatti's (2007) Experiment 8 and Endress and Mehler's (2009b) Experiments 3 and 4). To the extent that the results of Experiment 5 differ from Endress and Mehler's (2009b), differentiating actual items from generalizations might be comparatively harder in the visual modality, either due to genuine modality differences, or because the presentation rate is about four times faster in auditory statistical learning experiment. Crucially, however, as in Endress and Mehler's (2009b) experiments, including explicit segmentation cues improved performance on word vs. phantom-word trials as well.

Experiment 6: Do participants prefer unseen high-TP items to familiar low-TP items with segmented familiarizations?

In Experiment 6, the familiarization stream contained blank screens between words, as in Experiment 5. However, in contrast to Experiment 5, participants had to choose between words and part-words, and between phantom-words and part-words.

One would probably expect participants to choose both words and phantom-words over part-words for three reasons. First, in the auditory modality, partic-

ipants reject items that straddle prosodically defined word boundaries (Langus et al., 2012; Shukla et al., 2007; Shukla, White, & Aslin, 2011); as a result, they should not accept items either that straddle a boundary defined by a blank screen either. Second, Endress and Mehler (2009a) and Endress and Wood (2011) showed that, in both audition and vision, participants accept items that have "correct" items at their edges. Given that both words and phantom-words have correct edge shapes, they might be preferred to part-words for this reason. Third, Endress and Mehler (2009b) and Experiment 2 above show that phantom-words are preferred to part-words even after continuous familiarizations; as such, there is no reason to expect that they might not be preferred after segmented familiarizations

Materials and methods. Experiment 6 was identical to Experiment 2c (i.e., using stimuli from Brady et al. (2008)), except that words during familiarization were separated by a 1 s blank screen.

13 Catalan/Spanish bilinguals (10 females, 3 males, mean age 20.5, 18-23) took part in Experiment 6.

Results and discussion. As shown in Figure 10, participants preferred words to part-words ($M = 89.74\%$, SD $= 15.18\%$), $t(12) = 9.44$, $p < .0001$, D $= 2.6$, $CI_{95} = 80.57\%, 98.92\%$. The likelihood ratio in favor of the alternative hypothesis was 6.29×10^{18}. They also preferred phantom-words to part-words ($M = 97.12\%$, $SD = 7.49\%$), $t(12) = 22.67$ $p < .0001$, Cohen's $d = 6.3$, $CI_{95} = 92.59\%, 101.6\%$. The likelihood ratio in favor of the alternative hypothesis was 1.46×10^{111}.

An ANOVA with Trial Type as a within-participant predictor yielded a significant main effect, $F(1,12) = 5.66$, $p = .035$, $\eta_p^2 = .32$.[13]

Experiment 6 showed that participants prefer both words and phantom-words over part-words when the familiarization stream comprises blank screens between words. These results confirm that, when boundary cues are given, participants are sensitive to them, and prefer triplets that have "correct" items in their edges. However, they leave open the question of whether words were actually memorized. We will come back to this question in the General Discussion.

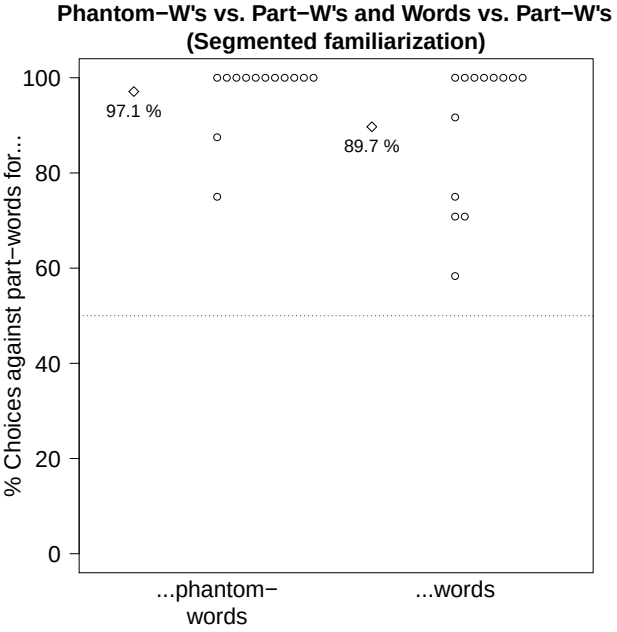

Figure 10. Results of Experiment 6. Circles represent individual participants and diamonds sample averages. In contrast to Experiment 2, the familiarization stream contained blank screens of 1 s after each word.

The nature of statistical learning

As discussed in the introduction, statistical sequence learning mechanisms can be partitioned into two classes: bracketing and clustering mechanisms (Goodsitt et al., 1993; see also Thiessen et al., 2013, for a review). In the case of word segmentation, bracketing mechanisms insert boundaries between words, and thus presumably require additional mechanisms to place items in memory. Clustering mechanisms, in contrast, group syllables together, which creates chunks that can be memorized. In this section, we will show that the current as well as Endress and Mehler's (2009b) and Perruchet and Poulin-Charronnat's (2012) results rule out any purely distributional clustering mechanisms. We will first provide general arguments, and then illustrate these general arguments using one of the most prominent clustering models (Perruchet & Vinter, 1998).

Distributional clustering cannot weigh TPs higher than frequency: A general argument

In this section, we argue that a clustering mechanisms cannot prefer phantom-words to part-words because assuming otherwise would lead to a serious contradiction. Let $\mathcal{M}$ be a purely distributional clustering mechanism that places sequences from the input into memory. As a result, it is more familiar with items it has heard than with items it has not heard. Let us assume that $\mathcal{M}$ weighs TPs higher than frequency information. As the previous discussion shows, this implies that there exist items that have not been encountered and yet are preferred to items that have been encountered. (Concretely, phantom-words should be preferred to part-words.) However, the existence of such items

[13] We did not analyze word vs. part-word trials as a function of the part-word type because we did not record the part-word types.

contradicts the assumption that $\mathcal{M}$ is clustering mechanism that retrieves items from memory, because attested items should be more familiar than unattested ones.

More formally, no clustering model that retrieves items from memory can accept unattested items. Let $\mathcal{M}$ be such a mechanism, and let us assume that it accepts an unattested item *XYZ*. If so, *XYZ* must have been memorized. However, this is only possible if *XY* has been followed by *Z*, or if *X* has been followed by *YZ*. Either way, the sequence *XYZ* would be attested, which contradicts the assumption that *XYZ* is unattested. Hence, $\mathcal{M}$ cannot accept such items.

As a result, no purely distributional clustering algorithm exists that (i) places items in some form of memory store, and (ii) weighs TP information higher than frequency information, at least if the recognition process is faithful. That said, it is certainly possible that phantom-words are easier to perceive because its component mono- and bisyllabic chunks are stronger than the component chunks of part-words. Likewise, it is possible that the familiarization sequences are analyzed by other processes simultaneously with a clustering mechanisms, that the preference for phantom-words is due to these additional mechanisms. While this might be so, these possibilities illustrate the point that phantom-words have not been stored in memory but are composed of other chunks, stressing again that our results are problematic for a memory-based chunking mechanism.

The result that unheard high-TP items are preferred over heard low-TP items thus rules out that humans are endowed with a purely distributional clustering mechanism that is used to populate the lexicon. These results also place important constraints on how purely distributional bracketing mechanisms are used. A bracketing strategy might well weigh TPs higher than frequency information. However, given such a strategy, we are left with three possibilities that can be illustrated with Ngon et al.'s (2013) data. As discussed above, they showed that 11-month-olds prefer syllable combinations that are frequent in French over those that are not, even if neither syllable combination is a real word, but they have no preference for real words over frequent syllable combinations. On the positive side, and as we discussed in the context of Graf-Estes et al.'s (2007) data, such results are consistent with the view that TP-based processes *prepare* learners to acquire words for when they are presented in situations that are conducive for word learning. On the negative side, however, such results might also suggest that distributional strategies are either not used to populate the lexicon, or that they lead to spurious lexical entries.

Distributional clustering cannot weigh TPs higher than frequency: Illustrations with PARSER (Perruchet & Vinter, 1998)

To make the arguments above more concrete, we now illustrate them with a specific computational clustering model. Following Perruchet and Poulin-Charronnat's (2012) suggestion that the PARSER model (Perruchet & Vinter, 1998) can account for the results with Italian and French speakers, we also use PARSER, one of the best known clustering models. We explore a wide range of model parameters and ask the model to choose between words, phantom-words and part-words, yielding 252,004 simulated experiments.

PARSER recursively chunks units from a continuous stream. For instance, upon encountering A and then B, it might create a unit AB. If, later on, the unit AB is followed by C, it might create a new unit ABC. Units that recur are strengthened, while spurious units are eliminated through decay and interference.

In the simulations below, PARSER was presented with a stream of symbols representing the syllables (or visual items). To simulate the fading in of the speech streams typical of artificial grammar learning experiments, we removed the first two syllables of the first word from the speech stream. We also repeated the same simulations without removing the initial and the final syllables. Following such familiarizations, we recorded the items the model had retained in its memory store as well as their memory strength.

As with humans, the model was tested on two different test lists. In one test list, it had to choose (i) between words and phantom-words and (ii) between words and part-words. In the other test list, the model had to choose (i) between words and part-words and (ii) between phantom-words and part-words. In each trial for each test list, we compared the weight of the two test items in the lexicon. For example, if a test trial comprised a word and a part-word, we would assign a value of 1 to the trial if the weight of the word in the lexicon was higher, of 0 with the weight of the part-word was higher, and of .5 if the two weights were identical. These scores were then averaged across the entire test list for a simulated participant (see below), in the same way as they are averaged for actual human participants.

PARSER has five parameters: the maximal number of units considered, the increment in memory strength upon encountering a unit, the threshold for an item to be removed from the lexicon, the initial weights of the syllables, the forgetting rate and the interference rate. While Perruchet and Poulin-Charronnat (2012) used a well chosen parameter set to model their data, we systematically varied the forgetting rate and the interference rate, while keeping the other variables at their original

values. Specifically, while the original PARSER model used a forgetting rate of .05 and an interference rate of .005, we varied the forgetting rate from 0 to .1 and the interference rate from 0 to .1, both in 251 equidistant steps.

The combinations of parameters and test lists yielded 251 (forgetting rates) × 251 (interference rates) × 2 (fading: initial and final syllables removed vs. kept in place) × 2 (test lists: word vs. part-word and phantom word or word and phantom-word vs. part-word) = 252,004 "experiments." Each experiment was run with 50 random initializations, representing 50 participants. Performance in each experiment was analyzed as with actual human participants.

A number of simulated experiments needed to be excluded from analysis due to undefined variances. This typically happened for relatively high forgetting and/or interference rates (because the model remembered hardly *anything* from the familiarization streams). However, in total, we excluded about .04% of the simulated experiments. The counts of excluded experiments are given in Table 3.

Results.

Test-lists comparing words to part-words and phantom-words. As shown in Figure 11a and b, the model significantly preferred words to part-words and words to phantom-words in 90% of the experiments where fading was implemented, and in 87% of the experiments where it was not. In the remaining experiments, the model had a numeric but non-significant preference for words.

Crucially, in 39% of the simulations with fading, the preference for words over phantom-words was significantly stronger than for words over part-words; in the remaining 61% of the simulations, the preference for words over phantom-words was numerically, but not significantly, stronger. The corresponding numbers for simulations without fading were 36% and 64%, respectively. As shown in Figures 12a, 12d, 13a and 13d, the preference for words over phantom-words was stronger than that over part-words in nearly 100% of the simulations with small forgetting and interference rates. Presumably, larger forgetting or interference rate tend to remove part-words (but also words) from the lexicon, which makes it relatively easier to reject part-words. Hence, in contrast to actual Italian and French speakers (with speech stimuli) and the Spanish/Catalan bilinguals here, there were no experiments where the preference for words over part-words was greater than the preference for words over phantom-words, while the preference for words over phantom-words was larger in a substantial proportion of the simulations.

A random-factor meta-analysis across experiments turned out to be impossible due to some simulated experiments with a variance of zero; as a result, the estimate of the between-experiment variance would not be defined (Borenstein, Hedges, Higgins, & Rothstein, 2010). To summarize the simulations in a quantitative way nonetheless, we computed the effect size (Cohen's d) for each experiment as well as the associated standard error across experiments. As shown in Figure 14, the mean effect size for the word vs. part-word discrimination was lower than the mean effect size for the word vs. phantom-word discrimination, in all experiments. As a result, the word vs. phantom-word discrimination was substantially easier for PARSER, while the opposite was true for actual humans (Endress & Mehler, 2009b; Perruchet & Poulin-Charronnat, 2012, and the data reported here).

Test-lists comparing words to part-words and phantom-words to part-words. As before, and as shown in in Figures 11a and b, the model significantly preferred words to part-words in 89% of the simulations where fading was implemented, and 87% of the simulations where it was not.

Crucially, the model significantly preferred part-words to phantom-words in 51% of the experiments where fading was used, and numerically but not significantly preferred part-words to phantom-words in the remaining 49%. The corresponding numbers for simulations without fading are 44% and 56%, respectively. Further, and as shown in Figures 12b, 12e, 13b and 13e, almost 100% of the simulations prefer part-words to phantom-words for small forgetting and interference values, while Figures 12c, 12f, 13c and 13f show that the preference for words over part-words was almost always greater than the preference for phantom-words over part-words. In contrast to human participants, there were no experiments where phantom-words were preferred. Further, and again in contrast to human participants, the preference for words over part-words was greater than the preference for phantom-words over part-words in all simulations.

As shown in Figure 14, the mean effect size for the phantom-word vs. part-word discrimination was between -.27 and -.30 (SE = .002). In contrast to actual humans, PARSER had a preference for part-words over phantom-words.

Discussion. The simulation results illustrate our general arguments above that a clustering mechanism with a memory component is inconsistent with a preference for phantom-words over part-words, as well as with a higher preference for words over part-words compared to phantom-words. Contrary to earlier claims, the empirical results thus refute the PARSER model.

In contrast to humans, the model prefers part-words

Table 3
Counts of excluded simulated experiments (among a total of 252,004) due to undefined variances.

	Test list	
	Words vs. Part-words/ Words vs. Phantom-Words	**Words vs. Part-words/ Phantom-Words vs. Part-words**
With fading	24	25
No fading	32	32

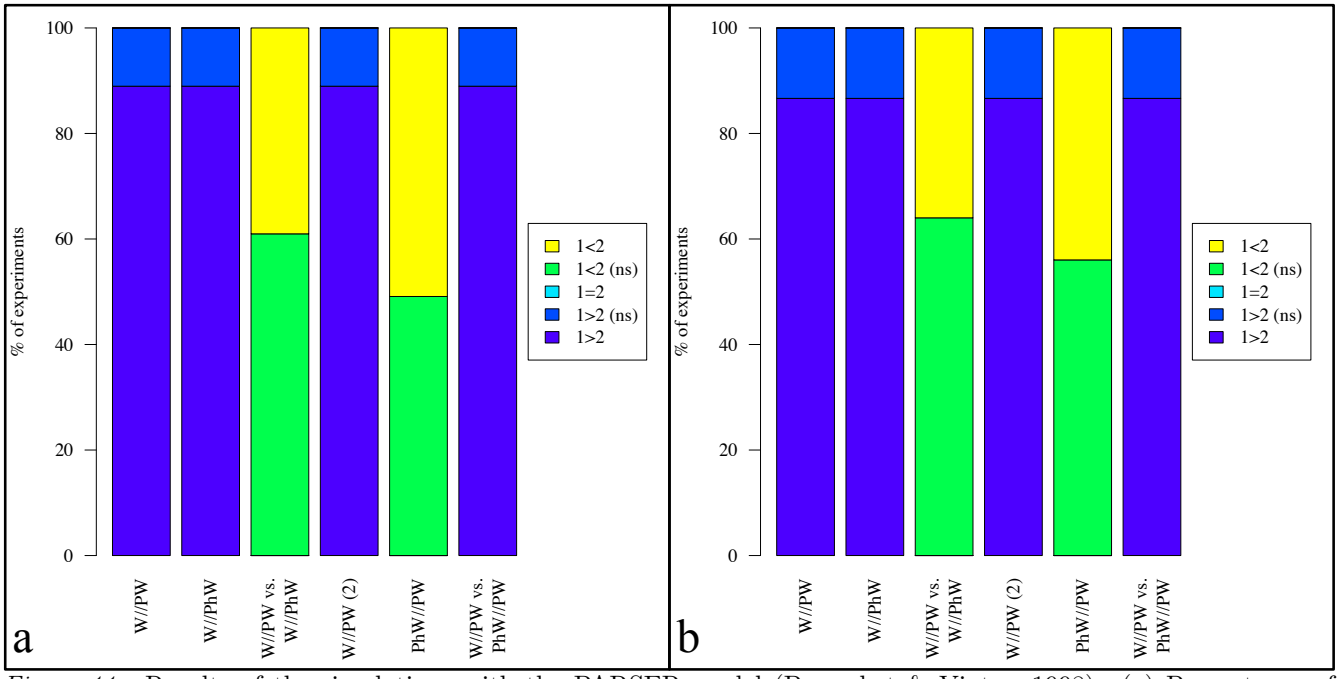

Figure 11. Results of the simulations with the PARSER model (Perruchet & Vinter, 1998). (a) Percentages of outcomes in the simulated experiments where the stream is faded in and out. The shades represent whether there is a significant preference for the first item type over the second type item in a comparison ("1>2"), whether there is a numeric but not significant advantage for the first item type over the second one ("1>2 (ns)"), whether there is strictly no preference ("1=2"), whether there is a numeric but not significant advantage for the second item type over the first one, ("1<2 (ns)"), or whether the latter difference is statistically significant ("1<2 (ns)"). W, PW, and PhW stand for words, part-words and phantom-words, respectively. For example, in the phantom-word vs. part-word comparison (fifth column), the model significantly prefers part-words to phantom-words in 62 % of the experiments ("1<2"), numerically but not significantly prefers part-words to phantom-words in 26 % of the experiments ("1<2 (ns)", and is at exactly 50% in another 11 % of the experiments ("1=2"). (b) As (a), but for familiarizations where the stream was not faded in and out.

to phantom-words (where humans prefer phantom-words), and finds the word vs. part-word discrimination harder than the word vs. phantom-word discrimination (where humans show the opposite effect). That said, for large forgetting or interference rates, words were not preferred over phantom-words to a greater extent than over part-words. This, however, is presumably because, with large interference or forgetting rates, part-words are not remembered by the model. In other words, the model predicts that phantom-words are either easier to reject than part-words, or that there is no differences between these item types. Both results are inconsistent with actual human behavior.

As a result, these simulations illustrate the finding above that purely distributional clustering mechanism with a memory component should not weigh TPs higher than frequency information.

General discussion

We investigated whether and how distributional cues such as TPs allow learners to place high-TP items in memory. Do distributional computations always lead to memory representations? And do such mecha-

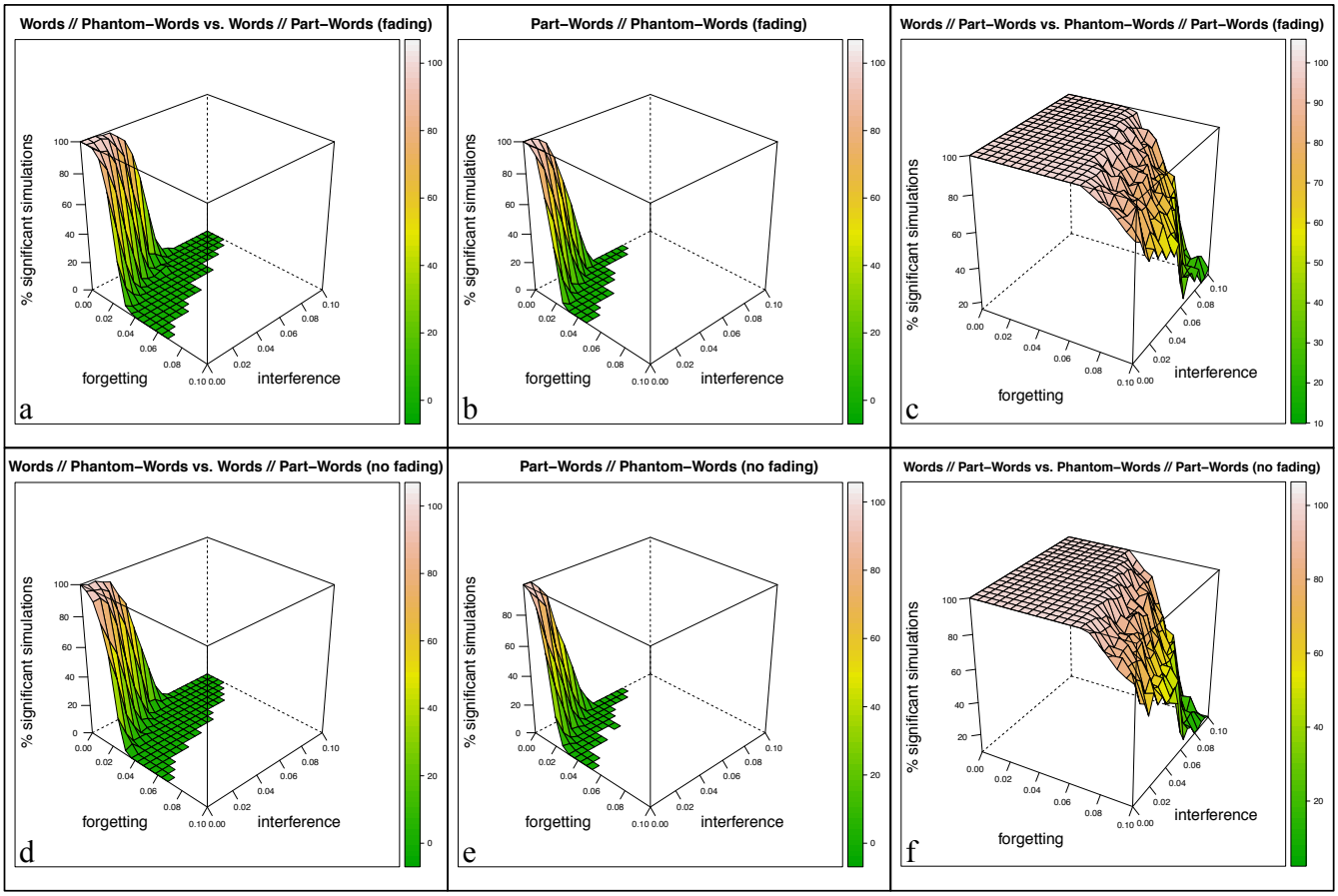

Figure 12. Model performance in terms of the percentage of simulations showing a preference for different test items. The results are shown as a function of the forgetting rate and the interference rate. The 251 forgetting and interference rates were binned into 20 intervals each. (Top) Simulations where the initial and the final syllables of the familiarization stream were faded in and out. (Bottom) Simulations where the syllables are not faded in and out. (a,d) Percentage of simulations where the preference for words over phantom-words was significantly larger than that for words over part-words. (b,e) Percentage of simulations with a significant preference for part-words over phantom-words. (c,f) Percentage of simulations were the preference for words over part-words was significantly larger than for phantom-words over part-words.

nisms operate according to clustering or chunking principles? Our point of departure were the experiments by Endress and Mehler (2009b) and Perruchet and Poulin-Charronnat (2012), who reported that, after familiarization with a continuous speech stream, both Italian and French speakers find it easier to discriminate high-TP items from low-TP items than to discriminate attested high-TP items (words) from unheard high-TP items (phantom-words), even though French and Italian speakers differ in whether they are sensitive to frequency information at all.

We first asked to what extent statistical learning leads to memorization in the visual modality. We use the visual modality for two reasons. First, it is plausibly less affected by language-specific influences than the verbal modality, and might thus bypass the language-specific differences that have plagued previous studies. Second, humans (and other animals) clearly need to learn and retain visual sequences, but it is unknown to what extent statistical learning leads to memorization in the visual modality.

In our experiments, we used visual stimuli, drawn from either Fiser and Aslin (2002a) or Brady et al. (2008). (While the results with the two stimulus sets were rather similar, a combined analysis of our experiments revealed a very small advantage for Fiser and Aslin's (2002a) stimuli.[14]) Experiments 1 and 2 were

[14]To compare performance with the two stimulus sets, we combined all word vs. part-word trials across all experiments that were run with both types of stimuli (i.e., Experiments 1, 2, 3 and 5), and submitted the percentage of

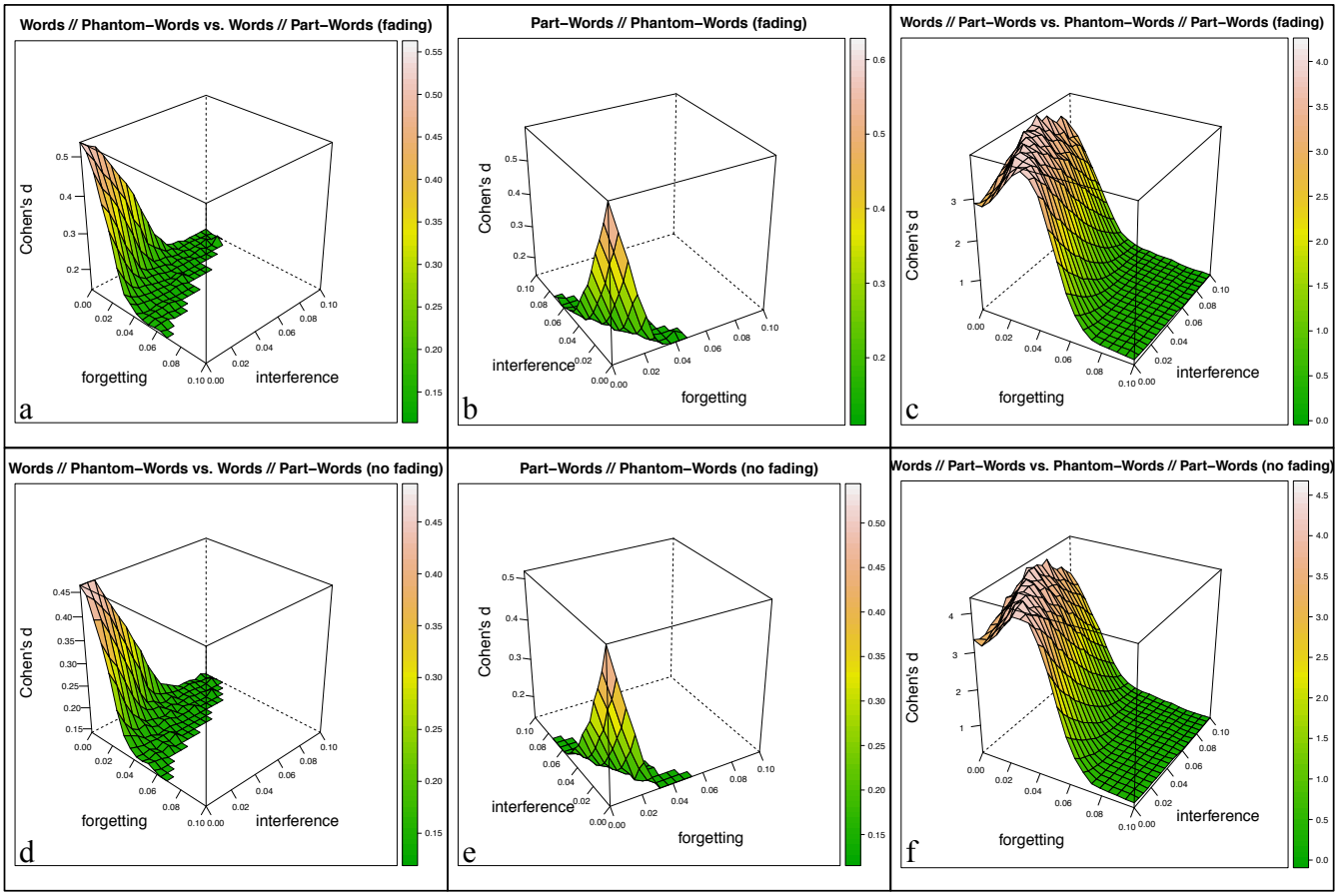

Figure 13. Model performance in terms of the effect sizes of different comparisons among test items. The results are shown as a function of the forgetting rate and the interference rate. The 251 forgetting and interference rates were binned into 20 intervals each. (Top) Simulations where the initial and the final syllables of the familiarization stream were faded in and out. (Bottom) Simulations where the syllables are not faded in and out. (a,d) Cohen's d of the difference between the preference for words over phantom-words and the preference for words over part-words; positive values indicate a stronger preference for words over phantom-words. (b,e) Cohen's d for the preference for part-words over phantom-words. (c,f) Cohen's d of the difference between the preference for words over part-words compared to that for phantom-words over part-words. Positive values indicate a stronger preference for words over part-words.

globally consistent with Endress and Mehler's (2009b) results: words were preferred to part-words, words were not preferred to phantom-words, and phantom-words were preferred to part-words, and the word vs. part-word discrimination was easier than the word vs. phantom-word discrimination. However, Experiments 1 and 2 also revealed two surprising results (albeit with very small effect sizes that might be type I errors): phantom-words were preferred to words, and the phantom-word vs. part-words discrimination was easier than the word vs. part-word discrimination. Experiment 3 and 4 were similar to Experiments 1 and 2, except that the familiarization streams were played twice. However, contrary to our expectations, performance did not even improve on the word vs. part-word

correct responses to an (arguably unbalanced) ANOVA with the between-participant predictors Stimulus Set and Familiarization Type (segmented vs. continuous). Both the main effect of Familiarization Type, $F(1,127) = 348.1$, $p < .0001$, $\eta_p^2 = .609$, and of Stimulus Set, $F(1,127) = 5.75$, $p = .017$, $\eta_p^2 = .01$, reached significance. Excluding the data from Experiment 3 (where participants were at chance) yielded similar results. While visual inspection of the data suggests that the effect of Stimulus Set was relatively systematic across experiments, it also shows that it was rather small. Further, in unpublished experiments using simultaneous presentation of stimuli (as opposed to the sequential presentation in the current experiments), and using much stronger TPs, performance was better with the Brady et al. (2008) stimuli than with other non-sense shapes. As a result, before conclud-

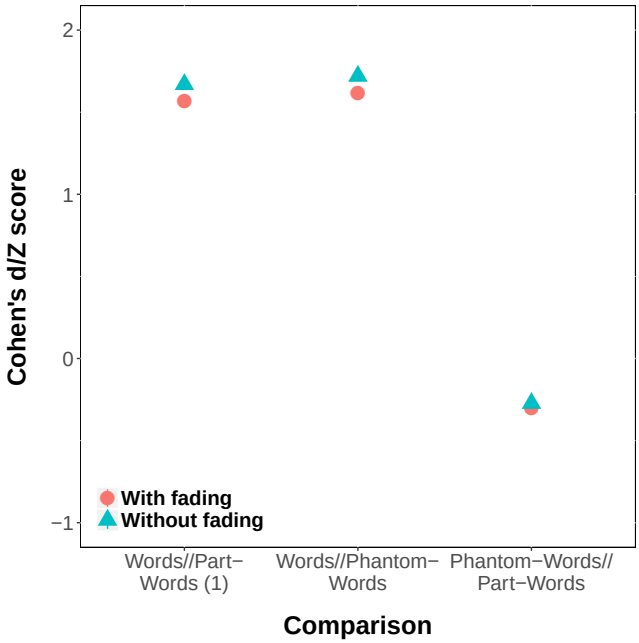

Figure 14. Results of the simulations with the PARSER model (Perruchet & Vinter, 1998). Average effect sizes (Cohen's d) of the word vs. part-word, word vs. phantom-word and phantom-word vs. part-word discrimination. The 95% confidence intervals of the effect sizes are shown but hard to see due to their small size. According to the PARSER model, the word vs. phantom-word discrimination should be easier than the word vs. part-word discrimination, and participants should have a preference for part-words over phantom-words. Both predictions are inconsistent with the empirical results.

discrimination, and became numerically worse.

In Experiments 5 and 6, the familiarization streams contained blank screens of 1 s between words. This manipulation improved performance, and established a weak preference for words over phantom-words, suggesting that, as in the auditory case, additional, non-statistical cues help tracking actually occurring items.

Taken together, these results are globally consistent with previous results with Italian speakers: Participants were more familiar with phantom-words than with part-words, found word vs. phantom-word discriminations harder than word vs. part-word discriminations, and, at least in some experiments, there was no preference for words over phantom-words. However, there are two caveats. First, some of the current results are somewhat unpredictable, and might reflect type I or type II errors, suggesting that statistical learning might be less robust than commonly believed. Second, results with French speakers (Perruchet & Poulin-Charronnat, 2012) as well as with Catalan/Spanish speaker (Langus & Endress, under review) show that words are sometimes preferred to phantom-words. As a result, boosting statistical learning performance might improve performance on the word vs. phantom-word trials as well. At minimum, however, the current results as well as the earlier results all demonstrate that phantom-words are relatively difficult to reject, and certainly much harder to reject than part-words. As a result, even subtle TP differences count more than substantial differences in chunk frequency.

Based on these and previous results, we suggest that these data are problematic for purely distributional clustering mechanism with a memory component. In an illustration with PARSER (Perruchet & Vinter, 1998), a prominent clustering model, we showed that this model is inconsistent with previous findings with Italian- and French-speaking adults (Endress & Mehler, 2009b; Perruchet & Poulin-Charronnat, 2012), as well as with the current results. Contrary to actual human data, PARSER prefers part-words to phantom-words, and finds it easier to choose between words and phantom-words than to choose between words and part-words. These simulation results are in line with our more general argument that purely distributional clustering mechanisms that have a memory component cannot weigh TPs higher than the frequency information about chunks.

We suggest that these results (i) are problematic for purely distributional clustering mechanisms, and (ii) suggest that distributional cues might prepare learners to acquire recurring items once they are encountered in more conducive learning situations than fluent continuous streams.

Possible mechanisms for extracting recurrent units from continuous sequences

As discussed in the introduction, Goodsitt et al. (1993) partitioned possible word segmentation mechanisms into "bracketing" mechanisms (that use cues to insert boundaries), and "clustering" mechanisms (that use cues to chunk items such as syllables). The finding that even relatively subtle TP differences count more than frequency of chunks is problematic for purely distributional clustering mechanisms that place items in memory, as such a mechanism cannot prefer unattested items such as phantom-words to attested items such as words. This general point is illustrated by the failure of PARSER (Perruchet & Vinter, 1998) to account for the

ing that arbitrary shapes are more conducive for statistical learning than meaningful objects, more targeted experiments need to address this issue more directly.

better performance on word vs. part-word trials compared to word vs. phantom-word trials, and for the preference for phantom-words over part-words. As a result, Endress and Mehler's (2009b), Perruchet and Poulin-Charronnat's (2012) and the current results are difficult to reconcile with the class of distributional clustering mechanism.

These results are particularly problematic for any theory that holds that distributional cues are used to place items in memory. In fact, frequency effects are one of the oldest and most robust findings in psycholinguistics (see, among many others, Cattell, 1886; Forster & Chambers, 1973; Solomon & Postman, 1952). If the output of distributional mechanisms has only a limited sensitivity to chunk frequency (at least for chunks longer than two syllables), the resulting representations seem to have different properties from actual items such as words.

The same conclusion follows from a memory perspective. Chunking is one of the central concepts in memory (see, among many others, Chase & Simon, 1973; Cowan, Chen, & Rouder, 2004; Feigenson & Halberda, 2008; Rosenberg & Feigenson, 2013; Simon, 1974). For example, chunking the letters C, I and A into the chunk CIA substantially facilities memorization. As such, if the output of distributional computations has only a limited sensitivity to such chunks, it seems to have different properties from the mechanisms we use to store items in memory.

These results also severely constrain how purely distributional bracketing mechanisms might be used, and leave us with two possibilities. Given that participants prefer unattested high-TP items over attested low-TP items, distributional strategies might either not be used to populate the lexicon, or might be used to memorize word candidates, while also memorizing numerous spurious lexical entries. This conclusion is strengthened by Ngon et al.'s (2013) result that French infants have no preference for frequent syllable combination over actual words.

Importantly, TPs might still play an important role in word learning. For example, if our interpretation of Graf-Estes et al.'s (2007) results above is correct, TPs might be useful by making it easier to memorize high-TP items once the items are presented in such a way that they can actually be memorized (e.g. in isolation). We thus predict that, in experiments like Graf-Estes et al.'s (2007), there might not only be an advantage in sound-vision associations for words over part-words, but also for phantom-words over part-words. In other words, learners would not be able to memorize any word-like items based on TPs alone.[15] Rather, TPs might prepare learners to acquire words once they are presented in more conducive learning situations.

It should be noted that these conclusions remain valid even if learners never face the problem of discriminating words from phantom-words, as the results so far are problematic for the entire class of distributional clustering mechanism. Be that as it might, Endress and Mehler's (2009b) and Perruchet and Poulin-Charronnat's (2012) results suggest that preferences in auditory forced choice tasks reflect sensitivities to the relevant cues (be they distributional or not), but that they are not necessarily diagnostic of memory representations that will eventually populate the lexicon. As a result, more diagnostic tests need to be developed.

Which cues and memory mechanisms can we use to extract items from sequences?

How do learners memorize the output of word segmentation mechanisms? As mentioned above, Endress and Mehler's (2009b), Perruchet and Poulin-Charronnat's (2012) as well as the current results indicate that learners do not use a purely distributional clustering mechanism that places items in memory. We will now suggest that TP-based mechanisms might not have the right format for forming lexical representations, and, more generally for memorizing items from sequences.

Sequences can be memorized in at least two ways (see e.g., Henson, 1998, for a review; see also Hitch, Burgess, Towse, & Culpin, 1996; Ng & Maybery, 2002; Page & Norris, 1998). First, one might encode the transitions among items in a sequence. This type of memory is called chaining memory. For example, in the sequences ABCD, such a mechanism would learn that A goes to B, B to C and C to D. TPs are a probabilistic version of this memory mechanism. Second, learners might encode the positions of the items relative to the edges of the sequence. For example, in the sequence ABCD, this mechanism would learn that A came first, D came last, and that B and C had some position relative to the first and the last one. Below, we will call this mechanism the "positional" mechanism (though Henson (1998) called it the ordinal mechanism). Empirical and computational results suggest that humans have both mechanisms (and not only one of them as often claimed in the memory literature), but that they dissociate in multiple ways (e.g., Endress & Bonatti, 2007; Endress & Mehler, 2009a; Endress & Wood, 2011; Endress & Bonatti, 2016; Marchetto & Bonatti, 2013).

[15] An alternative interpretation of these results would be that the role of TPs is to exclude low-TP items as word candidates. However, even with this account TPs could not be used to memorize items, as their role would just be to weed out inappropriate low-TP items.

Based on evidence from speech errors (e.g., R. Brown & McNeill, 1966; A. S. Brown, 1991; Dell, 1984; Kohn et al., 1987; Koriat & Lieblich, 1974, 1975; MacKay, 1970; Rubin, 1975; Tweney & Zaruba, 1975), artificial grammar learning (e.g., Endress, Scholl, & Mehler, 2005; Endress & Mehler, 2010) and formal linguistics (e.g., McCarthy & Prince, 1993; Nespor & Vogel, 1986; Selkirk, 1986), we have suggested that the representations of words (or of linguistic sequences in general) rely on the positional mechanism (Endress & Hauser, 2010; Endress & Mehler, 2009b; Endress, Nespor, & Mehler, 2009). Based on experiments with brain-damaged patients, Fischer-Baum, McCloskey, and Rapp (2010) and Fischer-Baum, Charny, and McCloskey (2011) drew the same conclusion for the memory representations of written words.

However, if positional and chaining memories rely on different and dissociable mechanisms, and if linguistic sequences are encoded using a positional mechanism, then tracking TPs cannot allow learners to place items in memory because a chaining mechanism cannot be used to create positional memory representations.[16]

In line with this view, Endress and Mehler (2009b) showed that participants preferred words to phantom-words after a familiarization with speech streams where short silences were inserted between words, or where the last syllable of each word was lengthened; according to Endress and Mehler (2009b), this was because either manipulation created edges, and thus allowed for the creation of positional memory representations (see also Endress & Bonatti, 2007). Experiment 5 replicated this result in the visual domain, albeit with much poorer performance.

Further evidence for the view that edge-cues are linked to memorization of stimuli comes from Shukla et al.'s (2007, 2011) experiments with adults and infants. In their crucial adult experiments, participants were familiarized with speech streams comprising intonational contours. Crucially, high-TP items were either aligned with the intonational phrase boundaries, or straddled the phrase boundary. Unsurprisingly, participants choose high-TP items over low-TP items when the high-TP items were aligned with the phrase boundary. However, when they straddled the phrase boundary, high-TP items were either not preferred, or even rejected (depending on the experiments). Crucially, control experiments showed that TPs were tracked equally well irrespective of whether the items were aligned with the phrase boundary or not.[17] These results thus suggest that the kinds of mechanisms that are recruited when edge-cues are given differ from those involved in TP-computations, and that the former might also be involved in the memorization of items.

Further, in actual learning situations, positional edge information might not only be provided directly by prosodic boundary cues, but also indirectly, via positional phonotactics. Previous research has shown that adults and infants can learn which phonemes are allowed at the edges of words (e.g., Chambers, Onishi, & Fisher, 2003; Onishi, Chambers, & Fisher, 2002) or utterances (Sohail & Johnson, 2016), and such phonotactic learning is more flexible at word-edges than at other positions within words (Endress & Mehler, 2010). Monaghan and Christiansen (2010) proposed a computational model, where learners keep a list of diphone pairs at utterance boundaries. This list is constructed from those words that are known to the learner at a given stage, and is used to guide segmentation. Using this simple approach, Monaghan and Christiansen (2010) found excellent segmentation performance that, to our knowledge, compares favorably with all other segmentation mechanisms that have been proposed so far, and, importantly, is based on information at edges of words.

However, the current results also show that better evidence is needed to show that edge-cues lead to memorization. Specifically, in Experiment 6, participants were familiarized with a stream where words were separated by a blank screen, and still preferred phantom-words to part-words. As mentioned above, these results might have three mutually non-exclusive explanations. First, as in Shukla et al.'s (2007) experiments, participants might reject part-words because they straddle a boundary. Second, both humans and other animals can compute generalizations based on what occurs in the edges of items. In other words, they can compute prefixes and suffixes (e.g., Endress & Bonatti, 2007; Endress & Hauser, 2011; Endress & Mehler, 2009a; Endress et al., 2010; Endress, Cahill, Block, Watumull, & Hauser, 2009). As phantom-words have "legal" initial and final shapes, they might be endorsed for this reason. Third, phantom-words have higher TPs than part-words. However, given that the current results do not allow us to tease apart these explanations, it will be important to provide more direct evidence that, once edge-cues are provided, the segmented items are actually memorized.

[16] In contrast, strategies relying on isolated words, on words at utterance edges and on prosody might be compatible with a positional memory encoding, and thus with the kind of memory we hypothesized to be used for memorizing linguistic sequences.

[17] Specifically, when, after the same auditory familiarization, the test items were presented as written items (which was possible due to the orthographic transparency of the participants' native language), both aligned and straddling high-TP items were recognized equally well.

A message of caution: the reliability of statistical learning

In addition to these theoretical conclusions, the rather unpredictable aspects of our results suggest a message of caution regarding the reliability of statistical computations. In fact, there is now a body of evidence suggesting that statistical learning might be less robust than it seems.

First, Frost et al. (2015) argued that, in typical statistical learning experiments, about a third of the participants perform more or less at chance level; relying on statistical learning alone for finding word boundaries would thus lead to a sizable proportion of infants with word learning difficulties.[18]

Second, Johnson and Tyler (2010) suggested that TP-based segmentation is effective only in particularly simple situations (see also Mersad & Nazzi, 2012). Specifically, they showed that infants readily segmented artificial languages where all words had the same length, but failed when words had different lengths (see e.g., Sohail & Johnson, 2016; though this was not the main focus of their experiments). Third, while different statistical measures such as forward TPs, backward TPs, mutual information and so forth perform well in different languages (e.g., Gervain & Guevara Erra, 2012; Saksida, Langus, & Nespor, in press), learners have no way to know which measure to choose before knowing the words of their native language to begin with. Fourth, computational studies have shown this kind of statistical learning to be much less effective for word segmentation than prosodic cues.[19] Fifth, even if we assume that statistical learning is effective on transcribed corpora, it is unclear how successful a statistically-based segmentation strategy would be in real speech, due to the presence of other speech cues such as prosody. In fact, when such speech cues are in conflict with statistical cues, speech cues typically count more in both adults and infants (e.g., Johnson & Jusczyk, 2001; Johnson & Seidl, 2009; Shukla et al., 2007, 2011). As a result, other cues in the speech signal might well override a sensitivity to TPs. Hence, despite a growing body of evidence that supports a statistical learning approach to language acquisition, it remains unclear how universal TPs really are.

Of course, statistical learning experiments are arguably short with respect to the years over which language acquisition unfolds, and might be more reliable with more extensive experience. However, the current results also show that increased exposure does not necessarily improve performance (though it seems to consolidate TPs in other experiments; Endress & Bonatti, 2007; Peña et al., 2002).

Conclusions

Taken together, the current results as well as earlier results with Italian and French speakers with auditory stimuli allow for one strong conclusion, and two weaker ones. First, even when TP differences are relatively subtle, TPs are weighed higher than frequency information, in Italian speakers, French speakers, and, with visual stimuli, in Spanish/Catalan bilinguals. As such, these findings rule out any purely distributional bracketing mechanism that places items in memory.

Second, learners are clearly sensitive to the frequency of individual items in sequences (for evidence in speech sequences, see e.g. Gervain, Nespor, Mazuka, Horie, & Mehler, 2008; Hochmann, Endress, & Mehler, 2010). However, it is unclear whether they are also sensitive to frequency of syllable groups or groups of shapes when familiarized with continuous streams with no explicit segmentation cues. At minimum, our results suggest that discriminations based on group frequency is relatively hard, which clashes with the important role of frequency effects in natural languages.

Third, the current results are consistent with the suggestion that sequences might be stored using a positional memory mechanism that encodes items with respect to their edges, but they do not provide strong evidence for them. Rather, they suggest that the use of choice tasks (or the corresponding methods with infants) is unlikely to answer the question of which items are memorized, and in which format. This probably requires subtler methods, which would then allow us to answer the question of how recurring items are segmented and memorized from continuous streams.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old

[18] This is not to say that, to acquire language, these infants would need to use a completely different set of mechanisms from their statistically more successful peers but it does suggest that one needs to ask to what extent statistical computations are robust enough to be used in real language acquisition.

[19] Gambell and Yang (2006) compared the effectiveness of prosodic segmentation to TP-based segmentation. While their prosodic algorithm yielded precision and recall scores close to or in excess of 90%, precision in their statistical learner was 41.6%, while recall was 23.3%, and thus much poorer than their simple stress-based segmentation mechanism. However, one can trade precision against recall by changing TP threshold at which a word is accepted (Swingley, 2005), and, in some languages, TP-based segmentation appears to be more successful than in others; Saksida et al., in press.

infants. *Psychological Science*, *9*, 321–324.

Aslin, R. N., Woodward, J., LaMendola, N., & Bever, T. (1996). Models of word segmentation in fluent maternal speech to infants. In K. Demuth & J. L. Morgan (Eds.), *Signal to syntax. bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Mahwah, NJ: Erlbaum.

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, *83*(2), 167-206.

Beckman, M., & Pierrehumbert, J. B. (1986). Intonational structure in japanese and english. *Phonology Yearbook*, *3*, 15-70.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. doi: 10.1002/jrsm.12

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298-304. doi: 10.1111/j.0956-7976.2005.01531.x

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14325–14329. doi: 10.1073/pnas.0803390105

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

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

Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33-44.

Brentari, D., González, C., Seidl, A., & Wilbur, R. (2011). Sensitivity to visual prosodic cues in signers and nonsigners. *Language and Speech*, *54*(1), 49–72.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, *109*(2), 204–223.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*, 325–337.

Cattell, J. M. C. M. (1886). The time it takes to see and name objects. *Mind*, *41*, 63–65.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69-77.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2–3), 221–268.

Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: a logical sequel to miller (1956). *Psychological Science*, *15*(9), 634–640. doi: 10.1111/j.0956-7976.2004.00732.x

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(5), 1119-30. doi: 10.1037/0278-7393.30.5.1119

Cutler, A., Oahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*(2), 141–201.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22.

Dell, G. S. (1984). Representation of serial order in speech: evidence from the repeated phoneme effect in speech errors. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *10*(2), 222–233.

Endress, A. D. (2010). Learning melodies from non-adjacent tones. *Act Psychol*, *135*(2), 182–190.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*(2), 247–299.

Endress, A. D., & Bonatti, L. L. (2016). Words, rules, and mechanisms of language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(1), 19–35.

Endress, A. D., Cahill, D., Block, S., Watumull, J., & Hauser, M. D. (2009). Evidence of an evolutionary precursor to human language affixation in a nonhuman primate. *Biology Letters*, *5*(6), 749–751.

Endress, A. D., Carden, S., Versace, E., & Hauser, M. D. (2010). The apes' edge: positional learning in chimpanzees and humans. *Animal Cognition*, *13*(3), 483-495. doi: 10.1007/s10071-009-0299-8

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177-199.

Endress, A. D., & Hauser, M. D. (2011). The influence of type and token frequency on the acquisition of affixation patterns: Implications for language processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(1), 77-95. doi: 10.1037/a0020210

Endress, A. D., & Mehler, J. (2009a, Nov). Primitive computations in speech processing. *The Quarterly Journal of Experimental Psychology*, *62*(11), 2187–2209.

Endress, A. D., & Mehler, J. (2009b). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*(3), 351-367.

Endress, A. D., & Mehler, J. (2010). Perceptual constraints in phonotactic learning. *Journal of Experimental Psychology. Human Perception and Performance*, *36*(1), 235–250.

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, *13*(8), 348–353.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology. General*, *134*(3), 406-19.

Endress, A. D., & Wood, J. N. (2011). From movements to actions: Two mechanisms for learning action sequences.

Cognitive Psychology, *63*(3), 141–171.

Feigenson, L., & Halberda, J. (2008). Conceptual knowledge increases infants' memory capacity. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(29), 9926–9930. doi: 10.1073/pnas.0709884105

Fenlon, J., Denmark, T., Campbell, R., & Woll, B. (2008). Seeing sentence boundaries. *Sign Language & Linguistics*, *10*(2), 177-200.

Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: first language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*(2), 477-99. doi: 10.1016/j.cognition.2008.04.002

Fischer-Baum, S., Charny, J., & McCloskey, M. (2011). Both-edges representation of letter position in reading. *Psychonomic Bulletin and Review*, *18*(6), 1083–1089. doi: 10.3758/s13423-011-0160-3

Fischer-Baum, S., McCloskey, M., & Rapp, B. (2010). Representation of letter position in spelling: evidence from acquired dysgraphia. *Cognition*, *115*(3), 466–490. doi: 10.1016/j.cognition.2010.03.013

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499-504.

Fiser, J., & Aslin, R. N. (2002a). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(3), 458-67.

Fiser, J., & Aslin, R. N. (2002b). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(24), 15822-6. doi: 10.1073/pnas.232472899

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: statistical learning of visual feature hierarchies. *Journal of Experimental Psychology. General*, *134*(4), 521-37. doi: 10.1037/0096-3445.134.4.521

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*, 627–35.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125. doi: 10.1016/j.cognition.2010.07.005

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117–125. doi: 10.1016/j.tics.2014.12.010

Gambell, T., & Yang, C. (2006). *Word segmentation: Quick but not dirty* (Tech. Rep.). New Haven, CT: Yale University.

Garcia, J., Hankins, W. G., & Rusiniak, K. W. (1974). Behavioral regulation of the milieu interne in man and rat. *Science*, *185*(4154), 824-31.

Gervain, J., & Guevara Erra, R. (2012). The statistical signature of morphosyntax: a study of Hungarian and Italian infant-directed speech. *Cognition*, *125*(2), 263–287. doi: 10.1016/j.cognition.2012.06.010

Gervain, J., Nespor, M., Mazuka, R., Horie, R., & Mehler, J. (2008). Bootstrapping word order in prelexical infants: a Japanese-Italian cross-linguistic study. *Cognitive Psychology*, *57*(1), 56–74. doi: 10.1016/j.cogpsych.2007.12.001

Glicksohn, A., & Cohen, A. (2011). The role of gestalt grouping principles in visual statistical learning. *Attention, Perception and Psychophysics*, *73*(3), 708–713. doi: 10.3758/s13414-010-0084-4

Glover, S., & Dixon, P. (2004). Likelihood ratios: a simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin and Review*, *11*(5), 791–806.

Goodsitt, J., Morgan, J. L., & Kuhl, P. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language*, *20*(2), 229-52.

Graf-Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, *18*(3), 254-60. doi: 10.1111/j.1467-9280.2007.01885.x

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53-64.

Hay, J. F., Pelucchi, B., Graf Estes, K., & Saffran, J. R. (2011). Linking sounds to meanings: infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106. doi: 10.1016/j.cogpsych.2011.06.002

Henson, R. (1998). Short-term memory for serial order: The Start-End Model. *Cognitive Psychology*, *36*(2), 73-137.

Hitch, G. J., Burgess, N., Towse, J. N., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *49*, 116–139.

Hochmann, J.-R., Endress, A. D., & Mehler, J. (2010). Word frequency as a cue for identifying function words in infancy. *Cognition*, *115*(3), 444-457. doi: DOI: 10.1016/j.cognition.2010.03.006

Hollingworth, A., & Henderson, J. M. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition*, *7*(1-3), 213-235. doi: 10.1080/135062800394775

Hollingworth, A., & Henderson, J. M. (2003). Testing a conceptual locus for the inconsistent object change detection advantage in real-world scenes. *Memory and Cognition*, *31*(6), 930–940.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567.

Johnson, E. K., & Seidl, A. H. (2009). At 11 months, prosody still outranks statistics. *Developmental Science*, *12*(1), 131-41. doi: 10.1111/j.1467-7687.2008.00740.x

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*(2), 339-45. doi: 10.1111/j.1467-7687.2009.00886.x

Kohn, S. E., Wingfield, A., Menn, L., Goodglass, H., Glea-

son, J. B., & Hyde, M. (1987). Lexical retrieval: The tip-of-the-tongue phenomenon. *Applied Psycholinguistics*, *8*(3), 245–266.

Koriat, A., & Lieblich, I. (1974). What does a person in a 'TOT' state know that a person in a 'don't know' state doesn't know? *Memory and Cognition*, *2*(4), 647–655.

Koriat, A., & Lieblich, I. (1975). Examination of the letter serial position effect in the "TOT" and the "don't know" states. *Bulletin of the Psychonomic Society*, *6*(5), 539–541.

Langus, A., & Endress, A. D. (under review). Are listeners sensitive to frequency information across languages.

Langus, A., Marchetto, E., Bion, R. A. H., & Nespor, M. (2012). Can prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and Language*, *66*(1), 285 - 306. doi: http://dx.doi.org/10.1016/j.jml.2011.09.004

MacKay, D. G. (1970). Spoonerisms: the structure of errors in the serial order of speech. *Neuropsychologia*, *8*(3), 323-50.

Marchetto, E., & Bonatti, L. L. (2013). Words and possible words in early language acquisition. *Cognitive Psychology*, *67*(3), 130 - 150. doi: http://dx.doi.org/10.1016/j.cogpsych.2013.08.001

McCarthy, J. J., & Prince, A. (1993). Generalized alignment. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology 1993* (pp. 79–153). Boston, MA: Kluwer.

Mersad, K., & Nazzi, T. (2011). Transitional probabilities and positional frequency phonotactics in a hierarchical model of speech segmentation. *Memory and Cognition*, *39*(6), 1085–1093. doi: 10.3758/s13421-011-0074-3

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*(3), 303-315. doi: 10.1080/15475441.2011.609106

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*(3), 545–564. doi: 10.1017/S0305000909990511

Nespor, M., & Vogel, I. (1986). *Prosodic phonology.* Foris: Dordrecht.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of personality and social psychology*, *28*(1), 28–38.

Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of personality and social psychology*, *35*(12), 847–862.

Ng, H. L., & Maybery, M. T. (2002). Grouping in short-term verbal memory: Is position coded temporally? *Quarterly Journal of Experimental Psychology: Section A*, *55*(2), 391–424.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34. doi: 10.1111/j.1467-7687.2012.01189.x

Onishi, K. H., Chambers, K. E., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition*, *83*(1), B13-23.

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in speech processing. *Journal of Memory and Language*, *53*(2), 225–237.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(7), 2745–2750. doi: 10.1073/pnas.0708424105

Page, M. P., & Norris, D. (1998). The primacy model: a new model of immediate serial recall. *Psychological Review*, *105*(4), 761-81.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437-42.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244-7. doi: 10.1016/j.cognition.2009.07.011

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604-7. doi: 10.1126/science.1072901

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition*, *36*(7), 1299–1305. doi: 10.3758/MC.36.7.1299

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*(4), 807–818. doi: http://dx.doi.org/10.1016/j.jml.2012.02.010

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*, 246–63.

Pilon, R. (1981). Segmentation of speech in a foreign language. *Journal of Psycholinguistic Research*, *10*(2), 113 - 122.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*(2, Pt.1), 304–308. doi: 10.1037/h0028558

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, *21*(4), 803-814.

Rosenberg, R. D., & Feigenson, L. (2013). Infants hierarchically organize memory representations. *Developmental Science*, *16*(4), 610–621. doi: 10.1111/desc.12055

Rubin, D. C. (1975). Within word structure in the tip-of-the-tongue phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *14*(4), 392–397.

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

Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: evidence for developmental reorganization. *Developmental Psychology*, *37*(1), 74-85.

Saffran, J. R., Johnson, E., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27-52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–21.

Saksida, A., Langus, A., & Nespor, M. (in press). Co-occurrence statistics as a language dependent cue for speech segmentation. *Developmental Science*.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*(1), 51-89.

Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: edge alignment facilitates target extraction. *Developmental Science*, *9*(6), 565-573. doi: 10.1111/j.1467-7687.2006.00534.x

Seidl, A., & Johnson, E. K. (2008). Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *Journal of Child Language*, *35*(1), 1-24.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure.* Cambridge, MA: MIT Press.

Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology Yearbook*, *3*, 371–405.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193-247.

Shatzman, K. B., & McQueen, J. M. (2006a). Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science*, *17*(5), 372-7. doi: 10.1111/j.1467-9280.2006.01714.x

Shatzman, K. B., & McQueen, J. M. (2006b). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception and Psychophysics*, *68*(1), 1-16.

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*(1), 1-32. doi: 10.1016/j.cogpsych.2006.04.002

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(15), 6038–6043. doi: 10.1073/pnas.1017617108

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120. doi: 10.1016/j.jml.2015.02.001

Simon, H. A. (1974). How Big Is a Chunk? *Science*, *183*, 482-488.

Slone, L., & Johnson, S. (2015). Statistical and chunking processes in adults' visual sequence learning. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 2218–2223). Austin, TX: Cognitive Science Society. Paper presented at the annual meeting of the cognitive science society.

Sohail, J., & Johnson, E. K. (2016). How transitional probabilities and the edge effect contribute to listeners' phonological bootstrapping success. *Language Learning and Development*, 1-11. doi: 10.1080/15475441.2015.1073153

Solomon, R. L., & Postman, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, *43*(3), 195–201.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86-132. doi: 10.1016/j.cogpsych.2004.06.001

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: a two-process account of statistical learning. *Psychological Bulletin*, *139*(4), 792–814. doi: 10.1037/a0030801

Toro, J. M., & Trobalón, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception and Psychophysics*, *67*(5), 867-75.

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology. General*, *134*(4), 552-64. doi: 10.1037/0096-3445.134.4.552

Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(1), 195–202.

Tweney, S., Ryan D.and Tkacz, & Zaruba, S. (1975). Slips of the tongue and lexical storage. *Language and Speech*, *18*(2), 388–396.

Van de Weijer, J. (1999). *Language input for word discovery* (MPI series in psycholinguistics, 9). Max Plank Institute for Psycholinguistics, Nijmegen.

van Alphen, P. M., & van Berkum, J. J. A. (2010). Is there pain in champagne? semantic involvement of words within words during sense-making. *Journal of cognitive neuroscience*, *22*, 2618–2626. doi: 10.1162/jocn.2009.21336

Weinstein, Y., McDermott, K. B., & Chan, J. C. K. (2010). True and false memories in the drm paradigm on a forced choice test. *Memory*, *18*(4), 375 - 384.

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, *16*(2), 80–84.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*(1), 3-21.

Appendix

Follow-up analysis of the effect of Trial Type in Experiment 2

To assess the reliability of the effect of Trial Type in Experiment 2, we performed a number of follow-up analyses. First, visual inspection of the data shows that the distribution for the phantom-word vs. part-word comparison is clearly discrete. We thus ran a binomial regression on the trial-by-trial data, using Trial Type as a fixed factor, and random intercepts and slopes for both Participants and Stimulus Sets. (In *R* notation, the model specification was *correctness* ~ *trialType*+(1+

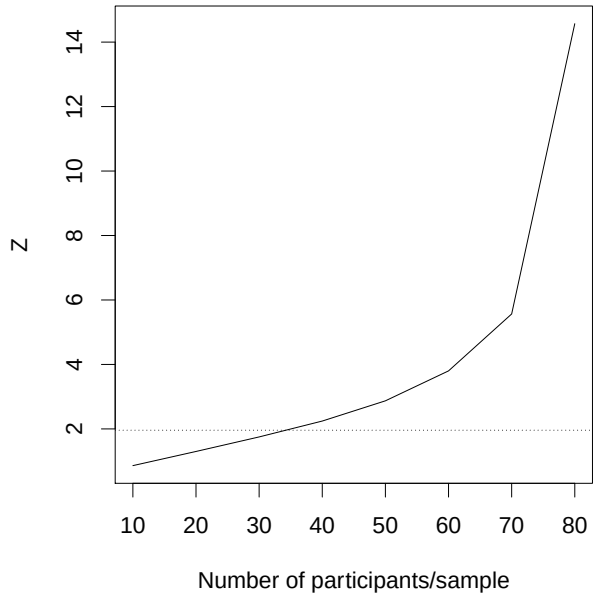

Figure A1. Additional analysis of Experiment 2. Z scores (Cohen's d) of the difference between the word vs. part-word and phantom-word vs. part-word discrimination as function of the sample size.

$trialType|participant)+(1+trialType|stimulusSet)$. In this analysis, Trial Type emerged as a significant predictor, $Z = 2.27$, $p = .023$.

Second, we subtracted the score for the phantom-word vs. part-word comparison from that of the word vs. part-word comparison for each participant, yielding 82 differences for 82 participants. For example, if a participant preferred phantom-words to part-words on 60% of the trials, and words to part-words on 55% of the trials, the resulting difference would be 5%. We generated 100,000 samples of the difference with sample sizes of 10, 20, 30, 40, 50, 60, 70, or 80. For each sample size, we took random samples of that size, and calculated the mean difference in each sample. We then calculated the mean and the standard deviation across samples at a given sample size, and used them to compute Z scores of the average difference at each sample size. Based on these Z scores, we asked at which sample size the Z score would cross the critical value for the .05 significance level (i.e., 1.96). As shown in Figure A1, a sample needs to include about 35 participants for the average difference to be reliably different from zero. This suggests that the difference is carried by a subset of the participants. In line with this conclusion, the effect size of the difference is just .25.