Available online at www.sciencedirect.com
ScienceDirect
Cognition 105 (2007) 247–299
COGNITION
www.elsevier.com/locate/COGNIT

Rapid learning of syllable classes from a perceptually continuous speech stream ☆, ☆☆

Ansgar D. Endress [a,b,*], Luca L. Bonatti [b,c]

[a] *Laboratoire de Sciences Cognitives et Psycholinguistique, EHESS-ENS-CNRS, Paris, France*
[b] *International School for Advanced Studies, Trieste, Italy*
[c] *University of Paris VIII at Saint Denis, France*

Received 12 July 2005; revised 14 August 2006; accepted 9 September 2006

Abstract

To learn a language, speakers must learn its words and rules from fluent speech; in particular, they must learn dependencies among linguistic classes. We show that when familiarized with a short artificial, subliminally bracketed stream, participants can learn relations about the structure of its words, which specify the classes of syllables occurring in first and last word positions. By studying the effect of familiarization length, we compared the general predictions of associative theories of learning and those of models postulating separate mechanisms for quickly extracting the word structure and for tracking the syllable distribution in the stream. As predicted by the dual-mechanism model, the preference for structurally correct items was negatively correlated with the familiarization length. This result is difficult to explain by purely associative schemes; an extensive set of neural network simulations confirmed this difficulty. Still, we show that powerful statistical computations operating on the stream are available to our participants, as they are sensitive to co-occurrence statistics among non-adjacent syllables. We suggest that different learning mechanisms analyze speech on-line: A rapid mechanism extracting structural information about the stream, and a slower mechanism detecting statistical regularities among the items occurring in it.

☆ This manuscript was accepted under the editorship of Jacques Mehler.

☆☆ This research was supported by an Allocation de Moniteur Polytechnicien fellowship to ADE and by the HFSP Grant RGP 68/2002. We are grateful to Ghislaine Dehaene-Lambertz, Jacques Mehler, Marina Nespor, Marcela Peña, Alessandro Treves, Judit Gervain and Valerie Lesk for discussions and comments on an earlier version of this manuscript.

* Corresponding author. Tel.: +39 040 3787 603; fax: +39 040 3787 615.
E-mail address: ansgar.endress@polytechnique.org (A.D. Endress).

doi:10.1016/j.cognition.2006.09.010

Keywords: Artificial grammar learning; Language acquisition; Dual mechanism models; Symbol manipulation

1. The multi-faced problem of language learning

Infants learn their native language fast and accurately despite the complex computations involved in language learning. One of the challenges they face is to learn the words of a language. Although adults perceive speech as a discrete sequence of words, no reliable acoustic cues indicate word boundaries in a speech stream. Therefore, to construct a lexicon, infants must first individuate the sound stretches in the continuous speech signal that form words.

It has long been known that sensitivity to statistical cues in the speech signal, in particular to transition probabilities (TPs), could in principle solve this problem (e.g., Harris, 1955; Hayes & Clark, 1970, but see Yang, 2004).[1] Recent research has shown that adults and infants can indeed segment a continuous speech stream when the only cue to word boundaries is that TPs are high within words and low between words (e.g., Aslin, Saffran, & Newport, 1998; Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996).

Sensitivity to statistical cues such as TPs might help infants to construct a lexicon. However, the lexicon is not a mere list of word-meaning pairs, but also contains a vast amount of structural information. For example, the lexical entry of the verb "give" must contain a specification of the relationship between the specific lexical entry (GIVE) and the arguments it takes (the giver, the given and the goal). Not only is such structural information commonly assumed in linguistic theories (e.g., Cook & Newson, 1996), but it also appears to be a requirement for word learning in the first place (Gillette, Gleitman, Gleitman, & Lederer, 1999; Landau & Gleitman, 1985). Structural information also plays a prominent role in morphology and syntax. Both require not only to individuate single words, but also to represent structural relations between sublexical morphemes or syntactic word classes. Since infants will eventually be able to use the lexicon productively, they cannot just memorize the words and sentences they have heard, but have to generalize grammatical and morphological regularities.

In spite of the importance of structural information in language, little is known about the processes underlying its acquisition. Indeed, it is still a matter of much debate whether learners do extract structural information, or whether they only track associations among items. Even in a well-studied case such as that of past tense formation, most studies either focus on adult speakers with years of linguistic experience, or try to model speakers' production by computer simulations. Hence, these

[1] TP is the conditional probability of encountering a syllable after having encountered another syllable. After having encountered the syllable /don/, there is a high probability of encountering the syllable /key/ because "donkey" is a word. More formally, conditional probabilities like $P(\sigma_{i+1} = /key/ \mid \sigma_i = /don/)$ are high within words, and low between words (σ denotes syllables in a speech stream).

studies can reveal only indirectly what psychological mechanisms may underlie the acquisition of linguistic structures. In this paper, we focus instead on the kinds of computations that can be observed *on-line* when participants learn from speech-like stimuli, and on the kinds of abstract structures that can be learned under these conditions.

2. From speech to rules: Types of generalization and mechanisms of rule extraction

Speech is the raw material from which both words are segmented and structural dependencies among words are learned. It is thus plausible that cues in the speech stream can be exploited not only in order to identify words, but also to learn some of the regularities found in grammar. Bracketing cues may be of particular relevance. Morgan (1986) formally showed that the complexity of sentences learners would need to process for inducing a grammar can be reduced if the input is bracketed, that is, if words that form syntactic groups are acoustically grouped together. In line with this proposal, Morgan, Meier, and Newport (1981) showed that adults were better at learning dependencies between word classes in visual word sequences when words were grouped in a structurally meaningful way.[2] Together with the results showing the importance of statistical computations in word segmentation, these theoretical and empirical arguments suggest that, even during on-line processing, a speech signal may be analyzed by both statistical and non-statistical computations.

This conclusion has been supported by recent evidence. Peña, Bonatti, Nespor, and Mehler (2002) investigated whether generalizations can be drawn from a simple continuous speech stream. They familiarized participants with a syllable stream composed of a concatenation of trisyllabic nonce words. In each nonce word (from now on, just "word"), the first syllable predicted the last syllable with certainty, whereas the middle syllables varied, yielding words of the form A_iXC_i.

While adults and infants can use TPs between adjacent syllables for segmenting a speech stream (e.g., Saffran, Newport, et al., 1996; Saffran, Aslin, et al., 1996), adult participants in Peña et al.'s (2002) experiments could not rely on such information. Instead, they could exploit TPs between non-adjacent syllables to identify words. Furthermore, because all words in the speech streams used by Peña et al. instantiated a rule, participants might have used also non-adjacent TPs to extract the regularity to which words complied. In fact, Peña et al. (2002) showed that participants could compute distant TPs, but only for segmenting the speech stream and not for generalizing the dependency between the first and the last syllable of words, although, in principle, TPs would be sufficient for capturing the dependency. Far from being

[2] However, these authors used written words presented simultaneously with their "referents" (that is, visual symbols). These referents shared also visual properties if the corresponding words belonged to the same class; hence, the cues to word classes available in these experiments may not be available to language learners. By contrast, acoustical properties may be. Indeed, several authors have proposed that words that belong to the same category may share some acoustical properties (e.g., Kelly, 1996;Selkirk, 1996; Morgan, Shi, & Allopenna, 1996; Shi, Werker, & Morgan, 1999).

helpful, an increased familiarity with the stream was even detrimental. After listening to a continuous stream for 30 min, not only did participants fail to generalize, but, when they had to choose between novel items respecting the generalization and familiar items violating it, they selected the statistically preferred, but structurally incorrect, items. Only when words were separated by subliminal silences could participants generalize the dependency between the first and the last syllable. In that condition, in striking contrast with the failure to detect the generalizations after exposure to 30 min of the continuous stream, a familiarization of only 2 min was sufficient to induce the generalizations. Therefore, the subliminal segmentation cues appeared to be required for the generalizations to be drawn. Peña et al. (2002) concluded that participants learn dependencies of the form "If the first syllable is /pu/, then the last syllable is /ki/" (henceforth called *A_iC_i-rules*), and that these computations are unlikely to be performed by a statistical mechanism.

Several morphological processes resemble A_iC_i-rules. Parasynthesis is an example. In Italian, the adjective "arrosso" and the verb "rossire" do not exist, but the verb "arrossire" ("to blush") does. Hence, this verb is created by simultaneously adding a prefix (ad-) and the verbal ending (-i-re) to a morphological root. Gender and number agreements also present the kind of long distance relationships among items that seems to call for A_iC_i-rules. In several gender-marked languages, articles, nouns and adjectives agree with an intervening lexical root (e.g., IL bambinO – "the child", masculine singular – vs. LA bambinA – feminine singular – vs. LE bambinE — feminine plural). However, several morphological and syntactic regularities appear to be more general than A_iC_i-rules. In morphology for example, the imperfect and the aorist in Classical Greek have a common prefix while the suffixes depend on the person and differ between the two past tenses; hence, several prefixes can be combined with several suffixes, and there is no one-to-one correspondence between prefixes and suffixes. Likewise, semantically modifying prefixes like /re/ in English or French can be used productively with virtually all verbs. They can be combined with all possible suffixes of a paradigm. Furthermore, learners are able to apply A_iC_i-rules even if they are unlikely to have encountered examples of such generalizations. For instance, in French the simple past is rare in spontaneous speech. Speakers may never have encountered a combination of the prefix /re/ with, say, the suffix for the second person plural in simple past, yet they are able to produce this combination if they need to do so. These examples suggest that several morphological processes involving prefixation and suffixation do not use fixed combinations of prefixes and suffixes, and are thus not readily described by A_iC_i-rules.

Also many syntactic regularities are more general than A_iC_i-rules. An English noun-phrase can be formed by, say, a determiner, an arbitrary modifier (e.g., an arbitrary instance of an adjective phrase) and an arbitrary instance of a noun phrase. A verb phrase can contain an arbitrary instance of the verb class, an arbitrary instance of the noun class as the direct object and further constituents selected on the basis of their class membership. In short, A_iC_i-rules are restricted to dependencies between particular items, whereas in morphology and syntax dependencies are often defined between *classes of items*. A language learner must find a way to learn such dependencies.

Table 1
Summary of the main test item types used

Item Type	Structure	Explanation
Word	A_iXC_i	Words appearing in the stream
Part-word	C_iA_jX (type 12) or XC_iA_j (type 21)	Item from the stream straddling a word boundary
Rule-word	$A_iX'C_i$	As words, but with new middle syllable
Class-word	$A_iX'C_j$	As rule-words, but with first and last syllables from different families

Peña et al.'s (2002) results are compatible with both kinds of dependencies. Participants in their experiments may have learned A_iC_i-rules, but they may also have learned that the first and the last position in a word are variables that take their values from distinct classes; these classes are formed by the syllables that occur in the first and the last position, respectively. Participants could thus have processed words with the structure A_iXC_i as particular instances of items conforming to this class-based regularity. For mnemonic purposes, we will call this class-based regularity a "*class-rule*".[3]

Which kind of generalizations did participants extract in Peña et al.'s (2002) experiments with segmented streams? If they learned a class-rule rather than A_iC_i-rules, then they should accept new words that conform to the class-rule but not to A_iC_i-rules. That is, they should accept as legal items with the structure $A_iX'C_j$; A_i and C_j always occurred, respectively, in the first and third positions of words in the stream but never in the same word, and X' is a syllable that occurred in the stream but never in the middle-position of words. We will call items with the structure $A_iX'C_i$ *rule-words*, items with the structure $A_iX'C_j$ *class-words*, and items that occurred in the stream but straddled word boundaries *part-words* (see Table 1). In the present experiments, we test whether participants exposed to subliminally segmented streams not only accept rule-words, as shown by Peña et al. (2002), but also class-words. In order to test this hypothesis, we ask participants to choose between class-words and part-words.

The issue of what generalizations can be learned online while listening to a speech stream is related to another problem: Does language acquisition recruit a single learning mechanism, or does it exploit different mechanisms for extracting words and rules? Associative mechanisms such as those involved in extracting words from continuous speech are exclusively sensitive to the (conditional) frequency of occurrence of particular syllables. Such sampling mechanisms should stabilize towards reliable responses only after a certain amount of exposure to the stream: the bigger the sampled set is, the more accurate the estimates of the distributions in the population will be. Therefore, if participants learn class-based regularities after exposure to short streams such as those used by Peña et al. (2002), it is unlikely that they do so by virtue of the simple associative mechanism with which they can extract words from the

[3] More precisely, we mean by "regularity entailing syllable classes" the regularity that syllables of one class could occur initially, and syllables of another class could occur finally.

speech stream. Rather, a device capable of extracting structural relations on the basis of few examples seems to be more adequate to explain how participants can be led to prefer items they have never heard to items they repeatedly experienced during familiarization. Thus, finding that class-rules can be learned from short exposure might suggest the existence of two learning mechanisms: one, statistically driven, for sampling the co-occurrence statistics of basic units in a stream, and another one, capable of extracting more structural information.

Although other models computing more powerful statistical relations beyond syllables may explain such a finding without resorting to a dual mechanism hypothesis, they would naturally predict that any relation extracted by sampling the input should be strengthened when more evidence becomes available. This prediction may allow one to decide between the two alternatives. We will return to this point later.

Dual route models are well known also in other areas of language (e.g., Baayen, Dijkstra, & Schreuder, 1997; Marslen-Wilson & Tyler, 1997; Marcus et al., 1992; Marcus, Brinkmann, Clahsen, Wiese, & Pinker, 1995; Prasada & Pinker, 1993; Pinker, 1991; Pinker & Prince, 1988). However, these models are mostly concerned with the kinds of representations necessary to account for language competence, rather than with the learning mechanisms needed to attain such representations. For example, a model like Pinker's (1991) rule-and-exceptions model is compatible with the view that both rules and exceptions are learned by virtue of a single, inductive, learning mechanism, because it focuses more on the final result of the learning process than on the mechanism generating it.

Rather than on the differences in representations at the endpoints of learning, we focus here on the nature of the learning mechanisms that adult speakers can recruit on-line to extract such representations. We test the hypothesis that at least two distinct mechanisms, performing different kinds of computations, are active for word extraction and rule extraction. We will call it the *More than One Mechanism Hypothesis*, or for short, the *MOM hypothesis*.

The MOM hypothesis makes a clear but counterintuitive prediction. As class-rules may be extracted on the basis of few examples, class-words should be familiar quickly, and the familiarity with the class-rule should change only slightly with increased familiarization. In contrast, the memory representations of the speech sequences actually encountered in the stream should become strengthened as the stream gets longer. Therefore, the MOM hypothesis predicts that, when participants are familiarized with streams such as those used by Peña et al. (2002), the preference for items conforming to the class-rule (but not encountered during familiarization) to items encountered during familiarization (but not conforming to the class-rule, such as part-words) should be more pronounced with shorter streams than with longer streams, whether continuous or segmented. This phenomenon should occur even if the speech sequence straddles word boundaries, like in part-words, because the consolidation of the memory traces for the actual sequence of syllables should override any disruption possibly caused by the presence of the subliminal gaps. Hence, class-words should be preferred to part-words after familiarizations with shorter streams but not with longer streams, even if the streams are subliminally segmented.

A single-mechanism account based on the detection of statistical regularities would predict the opposite. Associations become strengthened with longer exposure; hence, if the generalizations were computed by an associationist algorithm, participants should generalize the class-rule better when familiarized with a long stream as opposed to a short stream. We tested these predictions by familiarizing participants with streams of various durations.

3. An overview of the experiments

In all the experiments reported here, we familiarized participants with streams that consisted of concatenations of trisyllabic nonce words, as in Peña et al. (2002). In each nonce word, the first syllable predicted the last syllable with probability 1, whereas the middle syllable was variable. In some experiments, the stream was continuous, whereas in other experiments subliminal silences separated the words. The experiments are summarized in Table 2. In Experiments 1 and 2, we asked whether participants could learn a class-rule. After the familiarization, participants were asked to decide which of two trisyllabic items was more likely to be a word of the language. The test pairs were composed of a class-word and a part-word; while the latter occurred during the stream but violated the class-rule, the former conformed to the class-rule but did not occur during familiarization. If the MOM hypothesis is correct, generalizations should be available only after familiarizations with segmented streams; participants should generalize class-rules when the familiarization contained (even subliminal) gaps between words (Experiment 1) but not when the stream was continuous (Experiment 2). In Experiments 3–5, we contrasted the predictions of a purely associationist mechanism with those of the MOM hypothesis. We familiarized participants with a 2 min, a 30 min and a 60 min stream containing silences between words. If the MOM hypothesis is correct, the preference for the class-words should decrease for long familiarization durations.

Table 2
Summary of the experiments

Experiment	Silence betw. words	Stream duration	Test items	Preference for
1	+	10′	Class-words vs. part-words	Class-words
2	–	10′	Class-words vs. part-words	No preference
3	+	2′	Class-words vs. part-words	Class-words
4	+	30′	Class-words vs. part-words	No preference
5	+	60′	Class-words vs. part-words	Part-words
6	+	2′	$A_{\mathrm{i}}C_j$-X vs. XA_iC_j	No preference
7	+	10′	Class-words vs. $A_iX'A_j$ $C_iX'C_j$	Class-words
8	+	2′	Words vs. rule-words	Words
9	+	2′	Class-words vs. part-words[a]	Class-words
10	+	2′	Class-words vs. part-words[b]	Class-words
11	–	2′	Class-words vs. part-words[b]	No preference
12	+	2′	Rule-words vs. class-words	Rule-words
13	–	10′	Rule-words vs. class-words	Rule-words

[a] Both test items were surrounded by pure tones.
[b] These experiments controlled for possible phonotactic confounds in Experiments 1 and 2.

Experiments 6 and 7 controlled whether participants considered both the initial and the final syllable of nonce words to extract the generalizations, or only one of those. Experiment 8 asked whether participants generalized the class-rule, or simply responded on the basis of a partial encoding of the words in the familiarization stream. Experiment 9 tested whether alternative accounts based on statistical computations over both syllables and silences could explain the results. Experiments 10 and 11 controlled for possible phonological confounds in Experiments 1 and 2. Experiments 12 and 13 investigated the relationship between the A_iC_i-rules and the class-rule, and the role of statistical information in extracting them. Finally, an extensive set of neural network simulations explored to what extent a group of widely used statistical mechanisms could extract a class-rule.

4. Extracting regularities defined over syllable classes

Experiment 1 and 2 ask whether participants would accept class-words when exposed to a subliminally segmented (Experiment 1) or a continuous stream (Experiment 2). After familiarization, participants had to decide whether class-words or part-words looked more like items of the imaginary language. If they extracted a class-rule, they should prefer class-words; if this extraction is possible only when a stream is (at least subliminally) segmented, the preference for class-words should occur only for segmented familiarization streams.

4.1. Experiment 1

4.1.1. Materials and method

4.1.1.1. Participants. Twenty native speakers of Italian participated in the experiment (11 females, 9 males, mean age 23.3, range 21–29). In none of the experiments reported here did participants take part in more than one experiment.

4.1.1.2. Apparatus. The experiment was run on a computer running DOS and the EXPE programming language (Pallier, Dupoux, & Jeannin, 1997). Participants were tested individually in a quiet room. Stimuli were presented over headphones.

4.1.1.3. Materials. We used a language with nine nonce words grouped in three families, whose first and last syllable were, respectively, /pu/ and /ki/, /be/ and /ga/, and /ta/ and /du/. For all families, the middle syllable could be either /li/, /Ra/ or /fo/, yielding the following nine words with the structure A_iXC_i (bold-face indicating syllables that define families): **/pu**li**ki/**, **/pu**Ra**ki/**, **/pu**fo**ki/**, **/be**li**ga/**, **/be**Ra**ga/**, **/be**fo**ga/**, **/ta**li**du/**, **/ta**Ra**du/**, **/ta**fo**du/**. The stimuli were synthesized with the MBROLA speech synthesizer (Dutoit, Pagel, Pierret, Bataille, & Vreken, 1996), using the fr2 diphone base.[4] In

[4] Pilot tests with native participants showed that Italian native speakers find synthesized speech with the fr2 diphone base more intelligible than speech synthesized with the available Italian diphone bases. We thus decided to use fr2. Obviously, all phonemes we selected also exist in Italian.

order to avoid direct cues to word onsets, the stream was synthesized with increasing and decreasing amplitude ramps in the first and last 5 s, respectively. This ensured that the stream fades in and out at no point corresponding to either words or part-words. Words had a mean length of 696 ms (mean syllable duration 232 ms), and a pitch of 200 Hz.

4.1.1.4. Familiarization. During familiarization, participants were exposed to a syllable stream that lasted for approximately 10 min. The stream was a concatenation of the nine trisyllabic items. These items were separated by subliminal silences of 25 ms. Each word was repeated 100 times. Consecutive items could not belong to the same family, nor could they have the same middle syllable. This arrangement yielded TPs between any A_i and the adjacent X, or between any X and any adjacent C_i, of 0.33. Because repetitions of items belonging to the same family were excluded, the TP between the last syllable of any word and the first syllable of the following one was 0.5. The TP between any A_i and its C_i was always 1.

Participants were informed that they would hear a sequence of words from an imaginary language; they were instructed to listen carefully to the sequence. They were also informed about the nature of the task they would have to complete during test.

4.1.1.5. Test. Participants listened to 12 test pairs, presented twice in random order. Each test pair contained a class-word, that is, an item with the structure $A_iX'C_j$, and a part-word that occurred in the stream but spanned a word boundary. Part-words could be of two types. Half of the part-words comprised the last syllable of a word and two syllables of a succeeding word (hence they had the structure C_jA_iX, called part-words of *type 12*), and half comprised the last two syllables of a word and the first syllable of a succeeding word (hence they had the structure XC_jA_i, called part-words of *type 21*); in this way, half of the part-words began with a liquid and their last syllable started with a stop consonant, and half began with a stop consonant and their last syllable started with a liquid. Appendix A lists the test items of Experiments 1 and 2. The test pairs were presented once with the class-word first, and once with the part-word first, for a total of 24 trials. For each trial, items were presented with a 1.5 s interval. Participants were instructed to choose the item that they considered more likely to have occurred in the familiarization stream, and they were asked to guess if they were unsure. A new trial started 2 s after each response.

4.1.2. Results

Fig. 1 presents the results of Experiment 1. Participants preferred class-words to part-words ($M = 59.4\%$, $SD = 15.5\%$, $t(19) = 2.7$, $p = 0.014$).[5] There was no difference between the part-word types against which the class-words were tested ($t(19) = 0.42$, $p > 0.6$, ns, paired t-test). In order to assess the individual performances, we computed

[5] Unless otherwise stated, t-tests are two-tailed.

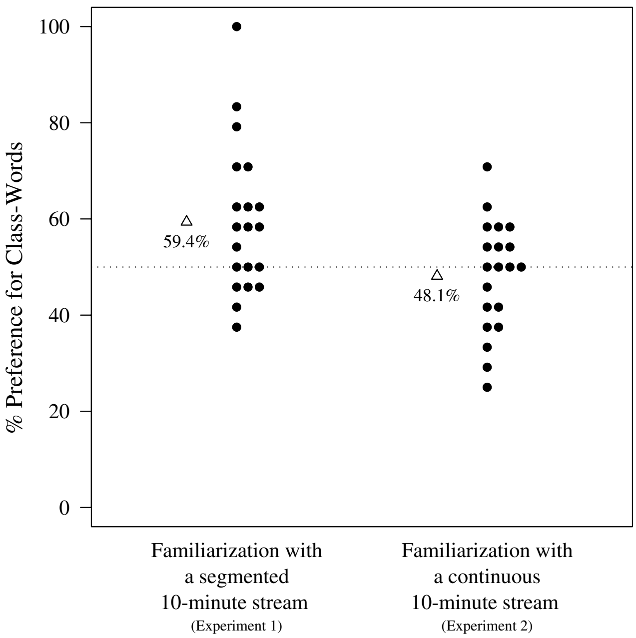

Fig. 1. Results of Experiment 1 (10 min familiarization with 25 ms silence between words) and Experiment 2 (10 min familiarization with a continuous stream). Dots represent the means of individual participants, triangles population averages and the dotted line the chance level of 50%. Participants prefer class-words to part-words after familiarization with a segmented stream, but not after familiarization with a continuous stream.

a one-tailed binomial test for each participant with their proportion of choices of class-words as dependent variable and 0.05 as significance threshold. Five participants preferred class-words by the binomial test, and none preferred part-words.

4.2. Experiment 2

4.2.1. Materials and method

Experiment 2 was identical to Experiment 1 except that the familiarization stream was continuous. Twenty native speakers of Italian participated in the experiment (13 females, seven males, mean age 22.4, range 19–29).

4.2.2. Results

Fig. 1 presents the results of Experiment 2. Participants showed no preference for either class-words or part-words ($M = 48.1\%$, $SD = 11.7\%$, $t(19) = -0.7$, $p > 0.48$, ns). There was no difference between the part-word types against which the class-words were tested ($t(19) = 0.18$, $p > 0.8$, ns). One participant preferred class-words by a binomial test, and two preferred part-words.

A joint ANOVA of the results of Experiments 1 and 2 with the presence of the subliminal silences as a between-subject factor showed that the preference for class-words in Experiment 1 was significantly higher than that in Experiment 2 ($F(1,38) = 6.7$, $p = 0.013$).

4.2.3. Discussion

In Experiment 1, participants preferred class-words to part-words; in Experiment 2, this preference disappeared. Because the only difference between the two experiments was the presence or absence of silent gaps between words, this result suggests that participants captured a class-rule when the familiarization stream was covertly segmented, but not when it was continuous. As in Peña et al. (2002), the subliminal segmentation cues separating words seem to have induced the generalizations, whereas such generalizations do not occur when bracketing cues are not available. However, the results of Experiments 1 and 2 also suggests that participants extracted abstract regularities defined over classes of items, defined by their positions within words, and not only relations between individual syllables as suggested by Peña et al. (2002).

5. Single versus dual mechanisms

Experiments 1 and 2 suggest that, when exposed to a stream containing subliminal bracketing cues, participants can extract regularities defined over syllable classes. They learn that the first and the last syllables of words have to be members of distinct classes, and not only dependencies among particular syllables. However, when bracketing cues are not present, the same amount of familiarization seems insufficient to attain those regularities. Peña et al. (2002, Experiment 1) showed that a 10 min familiarization with a continuous stream suffices to extract the words contained in it on the basis of non-adjacent relations among syllables. Thus, participant can compute statistical relations among non-adjacent syllables. Such computations could in principle allow them to capture the generalizations defined over word classes, were they directed to the right level of abstraction. However, it appears that participants cannot use the computational power they deploy over tokens to extract regularities defined over classes. Either they can compute statistical relations only among tokens in a stream, and the identification of token classes and their relations uses different mechanisms, or else, if classes can undergo the same statistical computations as tokens, the emergence of regularities among syllable classes may require different input characteristics, and may follow a different temporal course.

However informative, this conclusion does not tell us about the mechanisms extracting the generalizations from subliminally bracketed streams. The previous discussion suggests that the statistical computations that allow participants to segment the continuous stream in Peña et al.'s (2002) Experiment 1 are not responsible for the generalizations we found in our Experiment 1, but it is still possible that participants generalize by virtue of a single, associationist, mechanism not limited to processing co-occurrence statistics among tokens. Alternatively, the computations generating such results may be performed by distinct mechanisms, as the MOM hypothesis holds.

These possibilities make distinct predictions. If participants compute the generalizations by a single associationist mechanism, then they should benefit from an increase in exposure, because longer experience should strengthen the representations built by associative learning (whatever these representations may be). Therefore, the preference for class-words should positively correlate with the familiarization duration. In contrast, if the structural relations are extracted by a mechanism that quickly extracts generalizations among items in the signal, coexisting, and possibly competing, with a system tracking the syllable distribution, then participants should tend to prefer class-words to part-words when exposed to *short* familiarizations, and tend to lose this initial preference when the familiarization gets longer. This is because the familiarity with class-words would quickly reach ceiling, while the memory representations of part-words would keep being strengthened. We contrasted these predictions by familiarizing participants with streams of different duration in Experiments 3–5.

5.1. Experiment 3

A first prediction that may tell apart the single and dual mechanism hypotheses can be tested by using very short familiarization streams. If the MOM hypothesis is correct, no particular reduction in performance should be observed by reducing the familiarization, provided that the stream contains a few examples necessary (and sufficient) to project the generalization. By contrast, this reduction should negatively affect performance if an associative mechanism were the reason for the generalization attained in Experiment 1.

5.1.1. Materials and method

Experiment 3 was identical to Experiment 1 except that the familiarization lasted for 2 min. Twenty native speakers of Italian (11 females, nine males, mean age 26.9, range 19–46) participated in the experiment.

5.1.2. Results

Fig. 2 presents the results of Experiment 3. Participants preferred class-words to part-words ($M = 59.8\%$, $SD = 15.4\%$, $t(19) = 2.8$, $p = 0.01$). There was no difference between the part-word types against which the class-words were tested ($t(19) = 1.0$, $p > 0.31$, ns, paired t-test). Seven participants preferred class-words by a binomial test, and none part-words.

5.2. Experiment 4

The mirror prediction that could tell the two hypotheses apart is that the extraction of the generalization should get better with longer familiarizations if a single mechanism sensitive to the statistical distribution of the items in the stream were responsible for the extraction of the generalizations. The tendency should be opposite instead, according to the MOM hypothesis. We tested this prediction in Experiments 4 and 5.

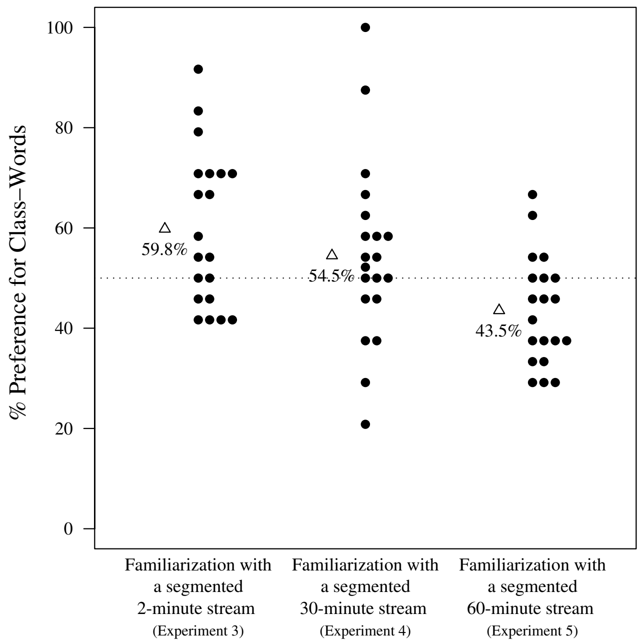

Fig. 2. Results of Experiments 3, 4 and 5. Dots represent the means of individual participants, triangles population averages and the dotted line the chance level of 50%. Participants prefer class-words to part-words after familiarization with a segmented 2 min stream, are at chance after familiarization with a segmented 30 min stream, and prefer part-words after familiarization with a segmented 60 min stream.

5.2.1. Materials and method

Experiment 4 was identical to Experiment 1 except that the familiarization stream was presented three times, for a total familiarization duration of 30 min. Participants could proceed to the next presentation of the stream by pressing a button. Twenty native speakers of Italian (16 females, four males, mean age 24.5, range 20–32) participated in this experiment.

5.2.2. Results

Fig. 2 presents the results of Experiment 4. Participants did not prefer either class-words or part-words ($M = 54.5\%$, $SD = 18.1\%$, $t(19) = 1.1$, $p = 0.283$, ns). There was no difference between the part-word types against which class-words were tested ($t(19) = 1.3$, $p > 0.19$, ns, paired t-test). Three participants preferred class-words by a binomial test, and two preferred part-words.

5.3. Experiment 5

5.3.1. Materials and method

Experiment 5 was identical to Experiment 1 except that the familiarization stream was presented six times, for a total familiarization duration of 60 min. As the familiarization was quite long, participants were allowed to take a pause between each stream presentation, and were allowed to walk around the laboratory if they felt so.When ready, they could proceed to the next presentation of the stream by pressing a button. Twenty native speakers of Italian (8 females, 12 males, mean age 24.3, range 21–33) participated in this experiment. Another 3 participants were excluded from analysis due to experimenter error.

5.3.2. Results

Fig. 2 presents the results of Experiment 5. Participants preferred part-words to class-words ($M = 43.5\%$, $SD = 10.9\%$, $t(19) = 2.7$, $p = 0.0154$). There was no difference between the part-word types against which class-words were tested ($t(19) = 1.6$, $p > 0.125$, ns). No participant preferred class-words by a binomial test, and 3 preferred part-words.

To assess the effect of familiarization length on the participants' preferences, we performed a series of common analyses of Experiments 1, 3, 4 and 5. These analyses are shown in Figs. 3 and 4.

We first entered the grand mean correct performance of each experiment in a regression analysis. Familiarization duration was negatively correlated with preference for class-words ($r = -0.986$, $F(1,2) = 68.5$, $p = 0.014$), and accounted for 95.8% of the variance. We then entered all subjects' individual means in a regression analysis.

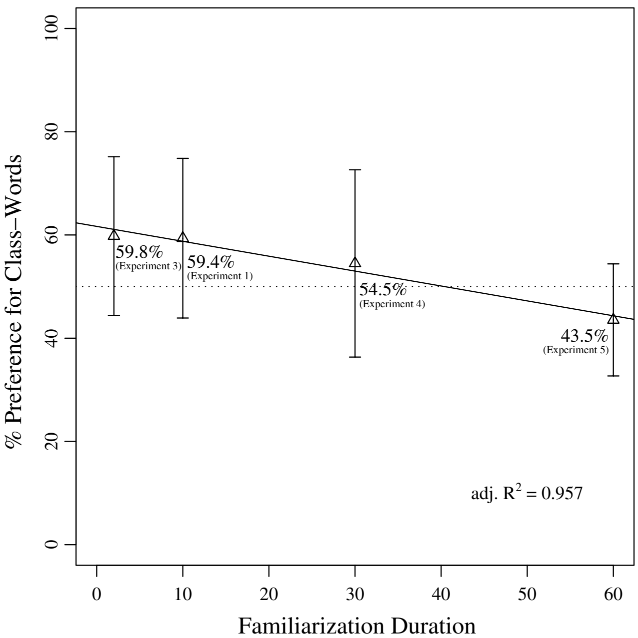

Fig. 3. Results of Experiment 1, 3, 4 and 5. Triangles represent population averages, the solid line is the regression line, and the dotted line represents chance level (50%). Error bars represent standard deviations from the means. Preference for part-words is negatively correlated with familiarization duration.

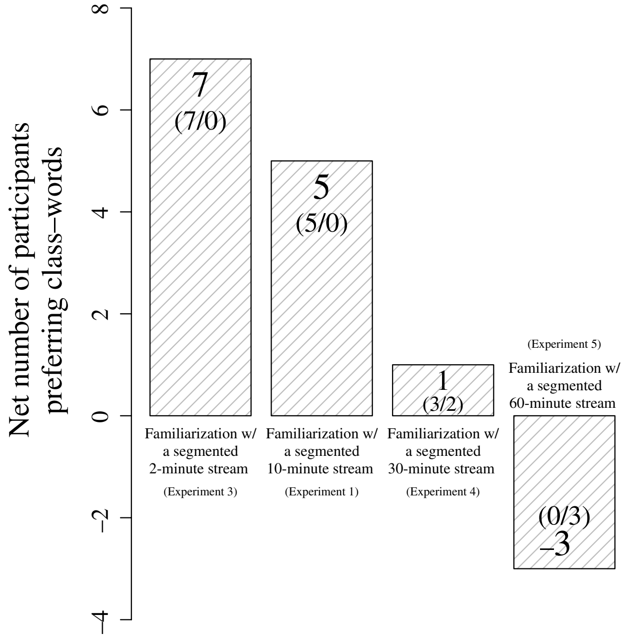

Fig. 4. Results of Experiment 1, 3, 4 and 5. Each box represents the net number of participants preferring class-words by an intra-participant binomial test (i.e., the difference between the numbers of participants preferring class-words and part-words, respectively); numbers in brackets indicate participants preferring class-words and part-words, respectively.

Despite the foreseeable much higher dispersion of the data, familiarization duration was still negatively correlated with the preference for class-words ($r = -0.399$, $F(1,78) = 14.8$, $p < 0.0003$), and accounted for 14.8% of the variance.

Finally, we compared the results of the experiments by computing the net number of participants preferring class-words in the four experiments, that is, the difference between the numbers of participants choosing class-words and those choosing part-words. The net number of participants preferring class-words was also correlated with familiarization duration ($r = -0.992$, $F(1,2) = 122.7$, $p = 0.008$).

5.3.3. Discussion

The MOM hypothesis predicts that a short familiarization should suffice for participant to project generalizations about the stream. It also predicts that the preference for class-words should decrease as the memory traces of the alternative choices (that is, the part-words) get consolidated. Experiments 3–5 confirmed both of these predictions. Experiment 3 showed that participants still prefer class-words despite a sharp reduction of familiarization. Experiments 4 and 5 showed that the preference for class-words was reduced by increasing the familiarization length. Indeed, joint analyses of Experiments 1, 3, 4, and 5 revealed a strong linear negative correlation between familiarization duration and preference for class-words. This suggests that increasing the familiarization duration has the effect of strengthening the memory trace of the associations between syllables in the stream. That long exposures favor part-word is also predicted by a theory arguing that a single associative mechanisms tracks relations between syllables. However, as these associations weaken with reduced exposure, a single mechanism associative model should predict that the participants' behavior should tend towards random performance as the exposure gets shortened, which is not what we found.

It is important to notice that the pattern of performance was obtained with segmented familiarization streams. Experiment 3 (and Peña et al.'s (2002) Experiment 5) showed that a 2 min stream is sufficient to induce generalizations. If they are potentially available with segmented streams, why did participants not show any evidence for generalizations in Experiments 4 and 5, where the familiarization stream was also segmented? Remember that, in our tests, participants always had to choose between part-words and class-words. Unlike class-words, part-words did appear during familiarization, and more often so in longer familiarizations. Hence, the memory traces of part-words should be strengthened over time, and participants should eventually come to prefer part-words, even though they preferred class-words after familiarizations with short streams. A dual mechanism model can easily account for this reversal in performance: As the memory representations of part-words may be still weak after short familiarization streams, the structure of the words may be the predominant property that participants may be able to extract quickly; in contrast, long streams – whether continuous (as in Peña et al.'s (2002) Experiment 4) or segmented (as in our Experiments 4 and 5) – consolidate the memory traces of part-words. This does not necessarily imply that the representation of the class-rule disappears over time (though we cannot rule out this possibility), but rather that the representations of the alternative choices become strengthened.

Seidenberg, MacDonald, and Saffran (2002) argued that Peña et al.'s (2002) results do not call for non-statistical learning because participants may rely on many statistical cues other than an abstract relation between syllables at distant position in a stream. If this were true, such cues should only be reinforced with increased familiarization. The fact that increased familiarization reduces performance suggests instead that statistical information plays a different role in learning different aspects of our language. It is not used to consolidate structure, but rather to identify items actually occurred in a stream. It is thus ideal to promote word segmentation, but not to explain sensitivity to structure.

6. Generalization versus perceptual biases

We suggested that Experiments 1 through 5 show that participants extract a class-based regularity when familiarized with subliminally bracketed streams. However, alternative possibilities requiring no sensitivity to structural information may also explain the results. For example, participants' preference for class-words may be based on covert perceptual biases rather than the extraction of a regularity. We now examine and exclude four possible bias-based explanations. In Experiments 6 and 7, we explore the possibility that participants did not extract a class-rule requiring simultaneous monitoring of non-adjacent syllables and their class memberships, but only attended to either the first or the last syllable of words, because of their salient positions. In Experiment 8, we ask whether the participants' "generalization" of the class-rule were in reality due to their failure to attend to word-medial syllables, and to encode words entirely. A third possibility is that the preference for class-words is based on TP computations between silences and syllables: although not overtly perceived, the silences in the stream may enter into the statistical computations and thus bias participants towards class-words without the need to postulate a class-rule. We explore this hypothesis in Experiment 9. Finally, Experiments 10 and 11 asked whether the subliminal silences were sufficient for inducing the class-rule, or whether a phonotactic confound in Experiments 1-5 in addition to the subliminal silences may explain participants' choices.

6.1. Experiment 6

A possible explanation of our results that would not require sensitivity for structural information might appeal to the salience of the initial and final syllables in words. The first (or last) syllable of class-words always appears in the position where it appeared during familiarization, whereas the first (last) position of part-words always contains a syllable that did not appear in initial (final) word position during familiarization. Therefore, if the subliminal gaps were used as cues for segmentation and participants only monitored whether the first or last syllables of test items appeared in that position during familiarization, neglecting every other information about the words and their structures, then they could have produced the pattern of results we observed in Experiments 1 and 3. Experiment 6 is designed to rule out this possibility.

After familiarization with a segmented 2 min stream, participants had to choose between items with the structure A_iC_jX and items with the structure XA_iC_j. If participants attended only to the last syllable of words, then they should choose items conforming to the structure XA_iC_j because they have the C_j syllable in final position and, by hypothesis, the initial syllable should be ignored. By the same argument, if participants attended only to the first syllable of words, they should choose items with the structure A_iC_JX. If instead participant monitored both the initial and the final syllables, then no preference should be observed.

6.1.1. Materials and method

6.1.1.1. Participants. Twenty native speakers of French (9 females, 11 males, mean age 22.8, range 20–27) participated in this experiment.

6.1.1.2. Familiarization. Participants were familiarized with the same covertly segmented 2 min stream used in Experiment 3.

6.1.1.3. Test. During test, participants had to choose between items with the structure A_iC_jX or items with the structure XA_iC_j. Both items in each test pair were built using the same syllables. Eighteen test pairs were presented twice in different orders.

The order of presentation was randomized between participants. Appendix B lists all the test items used in Experiment 6.

6.1.2. Results and discussion

As shown in Fig. 5, participants showed no preference ($M = 52.8\%$, $SD = 10.1\%$, $t(19) = 1.2$, $p = 0.234$, ns). One participant preferred items with the structure XA_iC_j by a binomial test, and none preferred items with the structure A_iC_jX.

If the generalizations in Experiments 1 and 3 were carried exclusively by the initial syllable, participants should have preferred items with the structure A_iC_jX to items conforming to XA_iC_j. If the generalizations were carried exclusively by the final syllable, the opposite preference should have been observed. Instead, participants showed no preference for either choice, suggesting that both initial and final positions are crucial to establish the generalization we observed in Experiments 1 and 3.

6.2. Experiment 7

Experiment 6 suggests that participants attend both to initial and final syllables. However, it does not rule out the possibility that, instead of learning two distinct syllable classes, participants may have learned one single syllable class including all syllables appearing at the edges, irrespectively of whether they appeared in 'A' or 'C' positions. Thus, they may prefer test items with one of those syllables in edge

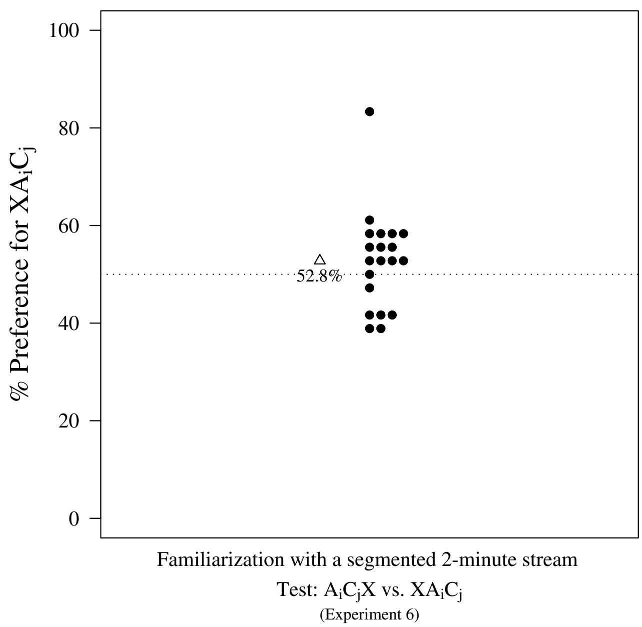

Fig. 5. Results of Experiment 6. Dots represent the means of individual participants, the triangle the population average and the dotted line the chance level of 50%. Participants showed no preference for items with the form A_iC_jX or XA_iC_j.

position, without monitoring whether it occurred at the beginning or at the end. Experiment 6 is compatible with this possibility. Experiment 7 addresses this concern. Participants were first familiarized with the segmented 10 min stream of Experiment 1. Then, they had to choose between class-words and two types of foils with the structures $A_iX'A_j$ or $C_iX'C_j$, respectively. If participants learned only one single class composed by the union of the 'A' and 'C' syllable classes, then they should not prefer class-words to either foil type. If, instead, they attended only to initial syllables, they should prefer class-words to foils of the form $C_iX'C_j$ but not to foils of the form $A_iX'A_j$, because the initial syllables "legal" in the latter foils. Likewise, if participants attended only to final syllables, they should prefer class-words to foils with the structure $A_iX'A_j$ but not to foils of structure $C_iX'C_j$. Finally, if participants extracted distinct classes for 'A' and 'C' positions, and attended to both initial and final syllables, they should prefer class-words to both foil types.

6.2.1. Materials and method

6.2.1.1. Participants. Twenty native Italian speakers (18 females, 2 males, mean age 22.3, range 20–27) were tested.

Familiarization. Participants were familiarized with the same covertly segmented 10 min stream of Experiment 1.

6.2.1.2. Test. During test, participants had to choose between class-words and foils with the structures $A_iX'A_j$ or $C_iX'C_j$, respectively. The test phase comprised 12 test pairs for each foil type. The order of presentation was randomized between participants. Appendix C lists the test items of Experiment 7.

6.2.2. Results and discussion

As shown in Fig. 6, participants preferred class-words to foils ($M = 56.1\%$, $SD = 10.2\%$, $t(19) = 2.7$, $p = 0.015$). There was no difference between the foil types to which class-words were compared ($t(19) = 0.74$ $p = 0.410$, ns, paired t-test). Three participant preferred class-words by a binomial test, and one preferred foils.

If the generalizations in Experiments 1 and 3 were carried exclusively by the initial syllable, participants should manifest no preference for class-words compared to foils with structure $A_iX'A_j$; likewise, if these generalizations were carried exclusively by the final syllable, they should not have preferred class-words to foils of the form $C_iX'C_j$. Finally, if participants had learned only one syllable class including all 'A' and 'C' syllables, they should have no preference for class-words against either foil type. Instead, participants preferred class-words to both foil types, suggesting that they learned distinct classes for initial and final syllables, and that they attended to both initial and final syllables.

6.3. Experiment 8

Another possible criticism of the conclusion that participants in Experiments 1 and 3 extracted a class-rule may hold that participants did not answer because they grasped the structure of the words of the language, but because they failed to encode

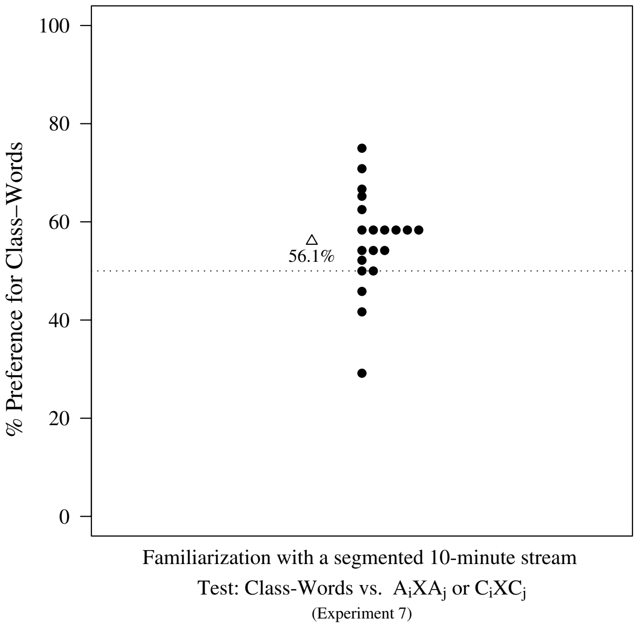

Fig. 6. Results of Experiment 7. Dots represent the means of individual participants, the triangle the population average and the dotted line the chance level of 50%. Participants preferred class-words to foils with the structures $A_iX'A_j$ and $C_iX'C_j$.

the words completely. In particular, the short experience with a 2 min stream may not have been sufficient to encode the words' middle syllables. Hence, participants might have responded on the basis of a partial representation of words, concentrating on the first or last syllable alone because all other information contained in the stream was not retained (e.g., Perruchet, Tyler, Galland, & Peereman, 2004).

It is difficult to see how Perruchet et al.'s (2004) criticism may explain our (or, for that matter, Peña et al.'s) results. Because participants have no preference for class-words when familiarized to a short continuous stream, the hypothesis would imply that participants would "ignore" the middle syllable of words when familiarized with a segmented, but not with a continuous, stream. It is difficult to see the rationale for this difference, but in principle it is possible that the segmentation cues help participants to encode the edge syllables only, while middle syllables may be remembered poorly, as any middle element in a list is.

In order to test for this possibility, we familiarized participants with a segmented 2 min stream. After this familiarization, they had to choose between words (that is, items that occurred in the stream) and rule-words. Rule-words are identical to words except for their middle syllable; in this position, rule-words contain a syllable that occurred in the stream but never in the middle of words. If participants ignored the middle syllable after a familiarization with a segmented stream, they should not prefer words to rule-words. In contrast, if they encoded also the middle syllables, they should prefer words to rule-words.

6.3.1. Materials and method

6.3.1.1. Participants. Fourteen native speakers of French (8 females, 6 males, mean age 22.1, range 18–26) participated in this experiment.

6.3.1.2. Method. Participants were familiarized with the same segmented 2 min stream used in Experiment 3. Then, they had to choose between words and rule-words. The 18 test pairs are given in Appendix D; each test pair was presented twice with different word orders.

6.3.2. Results and discussion

Fig. 7 shows that participants preferred words to rule-words ($M = 74.5$, $SD = 15.4$, $t(13) = 5.9$, $p < 0.00005$). Ten participant preferred words by a binomial test, and none preferred rule-words. Because the only difference between words and rule-words is in their middle syllables, participants should not show any preference if they failed to encode the middle syllables of words. The results of Experiment 8 show instead that an incomplete representation of the items in the stream, and in particular of their middle syllable is not the cause of the participants' choices. This suggests that the preference for structurally "correct" items observed in Experiments 1 and 3, as well as in Peña et al. (2002), is driven by a sensitivity to their structure, rather than by a failure to encode them.

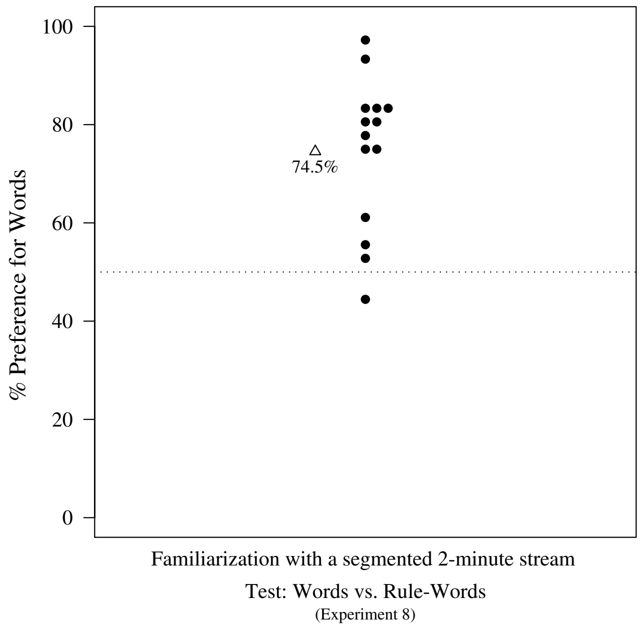

Fig. 7. Results of Experiment 8. Dots represent the means of individual participants, the triangle the population average and the dotted line the chance level of 50%. Participants prefer words to rule-words after familiarization with a segmented 2 min stream.

6.4. Experiment 9

Taken together, the results of Experiments 1–8 suggest that participants can extract a class-rule when familiarized with subliminally segmented streams. Such results seem difficult to explain on the basis of a single associationist model, because all measures of associative strength between syllables favor part-words. Part-words occurred during the stream, whereas class-words do not contain any chunks that appeared in the stream. Accordingly, by any measure of strength of associations between syllables (whether TPs, chunk strength, or others), part-words should be preferred to class-words. However, there might be another possibility. The subliminal silences contained in the segmented streams may also enter the computations of associations among items. Potentially, factoring subliminal silences in the determination of associations might explain preference for class-words over part-words without the need to capture regularities defined over syllable classes.

Peña et al. (2002, footnote 27) did address a version of this objection. They showed that computations of adjacent relations among silences and syllables could not account for the preference for rule-words found in their experiments. In particular, even when they asked participants to choose between rule-words and part-words containing the 25 ms silence that occurred during the familiarization phase (e.g., $C_i\#A_jX$, where # represents a 25 ms silence), participants still preferred rule-words to part-words. However, it may be claimed that the silences *surrounding* the test items, rather than the silences *within* part-words, are crucial for inducing a preference for class-words. If participants represent a class-word as $\#A_iX'C_j\#$, and if the representation of part-words also includes the silences, the TPs between silences and syllables in the test items may favor class-words over part-words regardless of the class-rule. Indeed, the sums of first, second and third order TPs, respectively, are higher in class-words than in part-words (Class-words: 1.33, 0, 1.33; part-words (both types): 0.33, 0, 0). Hence, participants could prefer class-words to part-words on the basis of TPs between silences and syllables.[6]

We investigated the possible role of the silences surrounding test items by exposing participants to the subliminally segmented stream of Experiment 3, but testing them with items immediately preceded and followed by a pure tone. Because this manipulation eliminates the transitions between silences and syllables, TPs to and from the silences are also eliminated. Hence, if participants still prefer class-words to part-words, TPs to and from the silences cannot explain the preference for class-words.

[6] It should be noted that in Peña et al.'s (2002) aforementioned experiment (footnote 27), TPs still favor at least half of the part-words compared to rule-words, suggesting that the silences surrounding test items are not important for the participants' choices, at least in that context. Counting reveals the following cumulative first, second and third order TPs, respectively: Rule-words: 1.33, 1.0, 1.33; Part-words of type 12: 1.66, 0.83, 0.33; Part-words of type 21: 1.66, 1.5, 0.33. Recall that part-words of type 21 contain two syllables from the first word and one from the second one, while part-words of type 12 contain one syllable from the first word and two from the second one.

6.4.1. Materials and method

6.4.1.1. Participants. Twenty native speakers of Italian (14 females, 6 males, mean age 22.8, range 19–41) participated in the experiment.

6.4.1.2. Method. Participants were familiarized with the segmented 2 min stream of Experiment 3. Then, in the test phase, they had to choose between class-words and part-words, immediately preceded and followed by 50 Hz tones with a duration of 232 ms. The tone duration corresponds to the average syllable duration; we chose this, instead of the duration of the segmentation cues, because 25 ms of pure tones before and after test items could be perceived as simple noise.

6.4.2. Results

As shown in Fig. 8, participants still preferred class-words to part-words ($M = 57.9\%$, $SD = 16.6\%$, $t(19) = 2.1$, $p = 0.046$). There was no difference between the part-word types against which the class-words were tested ($t(19) = 1.47$, $p = 0.157$, ns). Eight participants preferred class-words by a binomial test, and no participant preferred part-words.

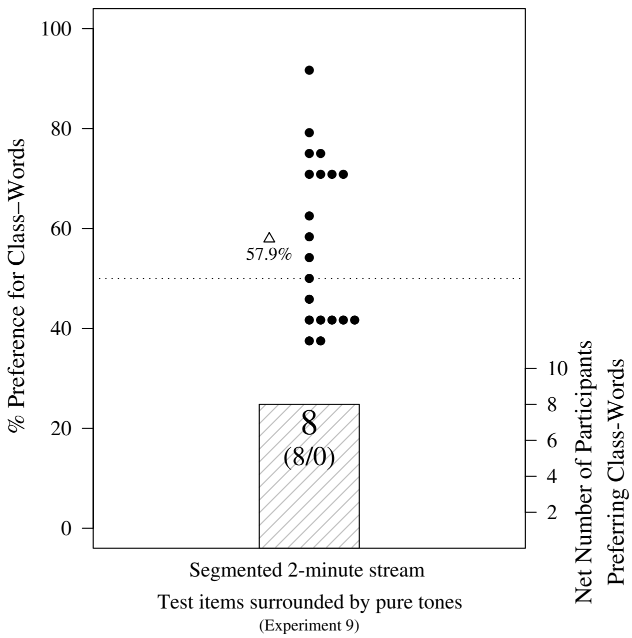

Fig. 8. Results of Experiment 9. Dots represent the means of individual participants, the triangle represents the population average and the dotted line the chance level of 50%. The box represents the net number of participants preferring class-words by an intra-participant binomial test (i.e., the difference between the numbers of participants preferring class-words and part-words, respectively); numbers in brackets indicate participants preferring class-words and part-words, respectively. Participants prefer class-words to part-words after a familiarization with a segmented 2 min stream even when the test items are surrounded by pure tones. Likewise, eight participants preferred class-words to part-words by a binomial test.

6.4.3. Discussion

Participants preferred class-words to part-words after familiarization with a segmented 2 min stream, although the test items were surrounded by pure tones. Because this manipulation obliterates TPs to and from silences surrounding test items, silences cannot play any crucial role to account for participants' preference for class-words found in Experiments 1 and 3. Experiment 9 demonstrates that associationist computations over syllables and silences alike cannot account for the preference for class-words.

Our conclusion does not imply that boundaries are irrelevant to explain the generalizations. After all, the subliminal silences are precisely boundaries that induce the generalizations. However, they do not contribute to computing TPs between items, nor are they represented as separate items, like syllables are. Indeed, boundaries have no predefined physical correlate. They can take the form of short silences, long silences, pure tones, and presumably many other physical configurations. In contrast, if syllables were replaced by arbitrary sounds or silences, this would clearly influence participants' responses. Hence, the computations induced by the boundaries are of a fundamentally different nature than the processes tracking syllable distributions.

7. Extraction of regularities and phonological factors

We have shown that participants seem to extract class-rules from short discontinuous streams. However, a possible alternative to our conclusion might appeal to the effect of prior information on participants' performance. Indeed, participants come to the experimental room loaded with their knowledge of their native language. Possibly, this knowledge may also influence our results. In particular, the words and class-words we used, as well as the words and rule-words from Peña et al.'s (2002) Experiments 2–5, begin and end with a stop consonant. If this pattern were favored by the statistical distribution of phonemes within Italian words, participants could have preferred class-words because they embody syllable transitions that are statistically preferred in their native language, rather than in virtue of the structure of the test items within the experiment. Whereas several researchers mentioned this as a possible criticism to Peña et al.'s (2002) work (e.g., Gómez & Maye, 2005; Newport & Aslin, 2004; Perruchet et al., 2004; Seidenberg et al., 2002), only Onnis, Monaghan, Richmond, and Chater (2005) provided empirical evidence that such confounding may play an important role in segmentation. With an elegant series of experiments, Onnis et al. (2005) tried to show that phonotactic knowledge is the real cause of the apparent ability of computing non-adjacent TPs documented by Peña et al. (2002). In their Experiment 4, Onnis et al. (2005) created a language in which words had continuants in initial position and plosives in medial and final positions, while maintaining most other aspects of Peña et al.'s (2002) Experiment 1 unchanged. The reasoning was that if participants preferred words in Peña et al.'s (2002) experiment, not because they computed distant TPs, but because they preferred sounds starting or ending with plosives, then they should prefer part-words with the new familiarization. After exposure to a continuous 10 min familiarization, participants indeed preferred part-words. Importantly, all the effect was carried by the comparison between words and part-words of structure CAX, which have structure stop-continuant-stop, whereas in the comparison between words and part-words of structure XCA (which have structure stop-stop-continuant) the difference was not significant.

As the familiarization stream in Onnis et al.'s (2005) experiments was continuous, their results may be more relevant to word segmentation than to the extraction of generalizations. However, it is easy to see how they could extend to our (and Peña et al.'s, 2002) results on generalizations. It may be argued that the particular phonological structure of the items of our language, and not sensitivity for its structure, may be the cause of participants' preferences for class-words. Indeed, although no computerized corpus is available for Italian, French and Spanish data may make this hypothesis plausible. We checked the frequencies of consonant-initial words in a French corpus (New, Pallier, Ferrand, & Matos, 2001). Eighteen percent of the words in the corpus start with one of the consonants that we used for word onsets, 15.84% with consonants that could be in the onset of medial syllables and 19.86% with consonants that could be in the onset of final syllables. These values are similar to those obtained from a Spanish corpus (M. Peña, personal communication) and may thus apply to Italian as well.

Yet, several points militate against the hypothesis that our and Peña et al.'s (2002) results can be explained away as byproducts of phonological confoundings. One first point is that Peña et al. (2002) did control for the possible effects of the phonological structure of their language on segmentation. They familiarized participants with continuous a stream in which, by changing the probability relations among the syllables, most words became part-words and vice versa (footnote 17; see also Bonatti, Peña, Nespor, & Mehler, 2006, for more details). For example, the stream was changed so that the item *puliki* (a word in the original stream) became a part-word in the new stream, and the item *ragapu* (a part-word in the original stream) became a word (see Appendix H for the complete material). As a result, in the control experiment no word began with initial plosives or had a stop-continuant-stop structure. In this control experiment participants still preferred words to part-words, although the effect was reduced. Thus, prior "phonotactic" knowledge does influence segmentation (as one may expect), but it is not responsible for participants' preferences in Peña et al. (2002) experiments.[7]

[7] Indeed, a comparison between Peña et al.'s (2002) control and Onnis et al.'s (2005) Experiment 4 reveals a close similarity between the test items. Hence, it is not the case that the simple elimination of the stop-continuant-stop structure of the words obliterates segmentation on the basis of long distance probabilities. It is more likely (as one would expect anyhow) that phonological and phonotactic factors are language-dependent, and French and English speakers' sensitivity to them will vary. This interpretation is confirmed by the previous observation that preference for part-words in Onnis et al.'s (2005) Experiment 4 was due to the preference for those part-word with stop-continuant-stop structure: apparently, this structure is more salient for English speakers. This, however, does not challenge Peña et al.'s (2002) conclusion that, over and above phonological and phonotactic biases, participants appear to take advantage of non-adjacent transition probabilities to segment a continuous speech stream. This conclusion is supported also by Onnis et al.'s (2005) Experiment 5: in this experiment, participants succeeded to segment on the basis of non-adjacent TPs when both words and part-words had similar phonological structure.

While the previous observations concerned the cues used for segmenting speech streams, our next point suggests that the aforementioned phonotactic and phonological biases cannot explain our generalization results either. Indeed, our Experiments 1 through 5 (where class-words were pitted against part-words) offer several internal controls for the effect of the stop-continuant-stop structure of words and test items. First, remember that the only difference between Experiments 1 and 2 is the presence of covert segmentation marks in the familiarization streams. As the phonology and phonotactics of the experiments is unchanged, we should expect both experiments to yield the same outcome if such factors were responsible for the results. However, in Experiment 1 participants succeeded to extract the generalization and in Experiment 2 they failed.

Then, consider the distribution of success and failures according to the types of part-words to which class-words are compared. In Experiments 1 through 5, half of the part-words begin with a liquid, and the other half with a stop consonant. If a preference for items beginning with a stop consonant influenced our results, then the preference for class-words should be more pronounced when class-words are compared to part-words beginning with a liquid than when they are compared to part-words beginning with a stop consonant. As we reported, there was no hint towards such a trend. Likewise, the last syllable of part-words starts with a liquid for half of them, and with a stop consonant for the other half. Hence, the preference for class-words should vary according to whether they are compared with part-words ending with liquids or ending with stops. No such effect was observed. It is thus unlikely that the preference for class-words is due to uncontrolled artifacts caused by participant's prior language knowledge. Rather, participants' preferences seemed to have been shaped by the experimental manipulations we introduced.

Still, a subtler variant of this objection could be envisioned. It is possible that participants may have learned to respond to class-words only because the streams contained the phonotactic cues mentioned above *in addition to* the silences contained in the covertly segmented streams. Indeed, work in language acquisition suggests that linguistic categories can be learned only when multiple, convergent, cues can be exploited (e.g., Gerken, Wilson, & Lewis, 2005; Gómez & Lakusta, 2004; Mintz, 2002; Monaghan, Chater, & Christiansen, 2005; Shi, Morgan, & Allopenna, 1998; Redington, Chater, & Finch, 1998; but see Cartwright & Brent, 1997). In our experiments, the effect of phonotactic confounds may have cumulated with the subliminal pauses to conjure a preference for class-words not induced by sensitivity to structure. Possibly, participants may not extract the class-rule in the absence of these additional phonotactic cues; in other words, they may learn the syllable classes only because they map onto phonological categories (e.g., stops).

We assessed this possibility in Experiments 10 and 11. In them, we used words that did not start or end systematically with stop consonants. Participants were familiarized with a segmented stream in Experiment 10 and a continuous stream in Experiment 11. If the particular phonological structure of our test items were responsible for the results of Experiments 1 and 3, we should not replicate them with a material lacking such cues.

7.1. Experiment 10

7.1.1. Materials and method

This experiment was identical to Experiment 1 except that (i) like in Experiment 3, participants were familiarized with a 2 min segmented stream, and (ii) the familiarization words were **li**pi**fe**, **li**mu**fe**, **li**ga**fe**, **to**pi**du**, **to**mu**du**, **to**ga**du**, **ba**pi**so**, **ba**mu**so**, and **ba**ga**so**. One word family had continuant consonants in first and last position; a second family started with a stop consonant but ended with a continuant; and a third family had stop consonants in both initial and final positions. Furthermore, the syllables in medial position started with stop consonants in two cases and with a continuant consonant in another case. Thus, no family had stop-continuant-stop structure. This mixture of phonological indexes characterized also the class-words of the test phase, and the part-words against which class-words were compared (see Appendix E). Sixteen native speakers of Italian (12 females, 4 males, mean age 23.4, range 19–30) participated in the experiment.

7.1.2. Results

As shown in Fig. 9, participants preferred class-words to part-words ($M = 68.2\%$, $SD = 14.7\%$, $t(15) = 5.0$, $p < 0.0002$). There was no difference between the part-word types against which the class-words were tested ($t(15) = 1.91$, $p = 0.076$, ns). Six participants preferred class-words by a binomial test, and none part-words.

7.2. Experiment 11

7.2.1. Materials and method

Experiment 11 was identical to Experiment 10 except that participants were familiarized with a continuous stream. Sixteen native speakers of Italian (8 females, 8 males, mean age 24.6, range 21–32) participated in the experiment.

7.2.2. Results

As shown in Fig. 9, there was no preference for class-words or part-words ($M = 54.9\%$, $SD = 13.1\%$, $t(15) = 1.5$, $p = 0.15$, ns). One participant preferred class-words by a binomial test, and one part-words. A joint ANOVA of the results of Experiments 10 and 11 with the presence of the subliminal silences as a between-subject factor showed that the preference for class-words in Experiment 10 was significantly higher than that in Experiment 11 ($F(1,30) = 7.3$, $p = 0.012$).

7.2.3. Discussion

Participants preferred class-words to part-words when familiarized with a segmented stream whose words did not have the stop–nonstop–stop consonant structure. This preference disappeared when they were familiarized with a continuous stream. These results closely mirror those of Experiments 1 and 2, despite the changes in the phonotactic structure of the familiarization. Hence, the "phonotactic" regularities present in Experiments 1 and 2 do not appear to be the crucial cues to successfully project a class-rule. Instead, the presence of several examples of a structure

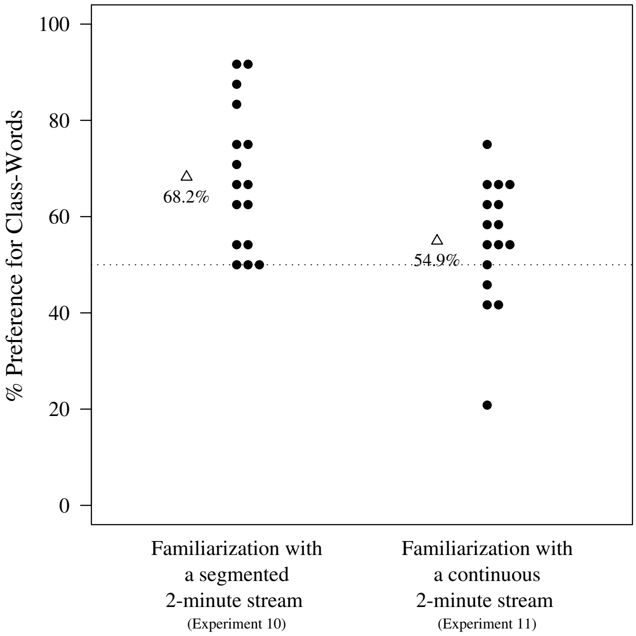

Fig. 9. Results of Experiment 10 (2 min familiarization with 25 ms silence between words) and Experiment 11 (2 min familiarization with a continuous stream). Dots represent the means of individual participants, triangles population averages and the dotted line the chance level of 50%. Participants prefer class-words to part-words after familiarization with a segmented stream also when words start with consonants from different natural classes. In contrast, participants do not prefer class-words after familiarization with a continuous stream.

within a short segmented speech stream seems to be the invariant element across the experiments that allows projecting the generalizations.

8. The relation between A_iC_i-rules and class-rules

While Peña et al. (2002) suggested that participants learn A_iC_i-rules when exposed to subliminally segmented streams, the preceding experiments suggest that participants learned a class-based regularity. Whereas we pitted class-words against part-words to ask whether participants extracted the class-rules, Peña et al. (2002) used rule-words and part-words to test whether A_iC_i-rules had been learned. Notice that, besides conforming to an A_iC_i-rule, rule-words also conform to the class-rule, but – in contrast to class-words – the TP between their first and last syllables is 1. In the following experiments, we ask whether a preference for rule-words necessarily implies that A_iC_i-rules are proper rule-like regularities represented independently of the corresponding class-rule, or whether such a preference may also be due to a sensitivity to the high TPs between their first and last syllables that is computed on top of the class-rule.

An analogy with ordinary linguistic knowledge may clarify the difference between the two hypotheses. Native speakers of a language such as English know the word classes "noun" and "verb" and how they can be used to compose sentences. This kind of knowledge is properly syntactical. However, they also know that some instances of the class "noun" are more likely to be combined with specific instances of the class "verb": they know that "fish" is more likely to occur with the verb "swim" than with the verb "fly", and that the opposite holds for "bird". This knowledge is not grammatical, and is likely acquired by sampling the co-occurrences of exemplars of the classes "verb" and "noun".

In the same vein, participants in our experiments might have extracted a class-rule, but might also have noticed the frequent co-occurrence of initial and final syllables in A_iC_i-rules. Thus, the generalizations in Peña et al.'s (2002) experiments (i.e., the A_iC_i-rules) could be explained in terms of a class-rule, together with a statistical sensitivity to TPs between non-adjacent syllables. Participants might have extracted the A_iC_i-rules because they extracted the class-rule and, on top of it, also noticed that the combinations between initial and final syllables that occurred in A_iC_i-rules were particularly frequent.

We tested these possibilities by studying the conditions under which participants prefer rule-words to class-words. Previous data and the current experiments suggest that, with familiarizations of artificial streams such as those used in our experiments, generalizations are available only after familiarizations with segmented, but not with continuous streams. Therefore, if A_iC_i-rules are proper rule-like regularities, extracted independently of the class-rule, then they should be available only after familiarizations with segmented, but not with continuous, streams. Alternatively, A_iC_i-rules may result from statistical computations independent of the generalizations, and may be seen as particular, frequently instantiated items conforming to the class-rule. We thus asked whether participants prefer rule-words to class-words after being exposed to either a subliminally segmented (Experiment 12) or a continuous stream (Experiment 13). If rule-words are preferred to class-words only after a familiarization with a segmented stream, we would have evidence that A_iC_i-rules are independently represented generalizations. In contrast, a preference for rule-words to class-words after a familiarization with a continuous stream would raise the possibility that A_iC_i-rules may not be independently represented generalizations but may arise from a sensitivity to non-adjacent TPs computed on top of the class-rules (or, if A_iC_i-rules are independently represented, that they are extracted by different mechanisms from the class-rules).

8.1. Experiment 12

8.1.1. Materials and method

8.1.1.1. Participants. Twenty native speakers of French (13 females, 7 males, mean age 22.3, range 19–27) participated in Experiment 12.

8.1.1.2. Procedure. Participants were familiarized to a 2 min subliminally segmented stream (as in Experiment 3). They were then tested with 24 test trials. In each trial, they had to choose between a rule-word and a class-word. Rule-words were constructed by inserting a syllable between the initial and the final syllable of one family (e.g., between

/pu/ and /ki/) that had never occurred in this position. Therefore, they were like class-words except that the first and the last syllable belonged to the same family, and conformed both to the A_iC_i-rules and the class-rule. Most importantly, whereas all TPs in class-words are 0, in rule-words the first syllable predicts the last syllable with certainty.

Half of the rule-word/class-word-pairs overlapped in their first two syllables and differed only in the last syllable; the other half overlapped in their last two syllables and differed in their initial syllable. Each pair was presented three times. Appendix F presents the test items used in Experiment 12.

8.1.2. Results

As shown in Fig. 10, participants preferred rule-words to class-words ($M = 60.0\%$, $SD = 10.4\%$, $t(19) = 4.3$, $p < 0.00036$). There was no difference between test pairs with initial or final overlap ($t(19) = 0.892$, $p > 0.38$, ns, paired t-test). Ten participants preferred rule-words by a binomial test, and none preferred class-words.

8.2. Experiment 13

8.2.1. Materials and method

8.2.1.1. Participants. Twenty native speakers of French (11 females, 9 males, mean age 26.2, range 18–38) participated in Experiment 13.

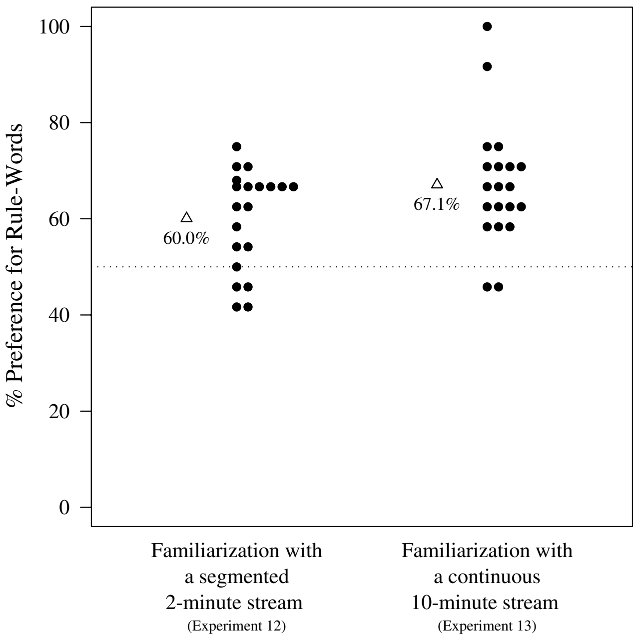

Fig. 10. Results of Experiments 12 and 13. Dots represent the means of individual participants, triangles population averages and the dotted line the chance level of 50%. Participants prefer rule-words to class-words when familiarized with a segmented 2 min stream (Experiment 12) or a continuous 10 min stream (Experiment 13).

8.2.1.2. Procedure. Participants were familiarized with a continuous 10 min stream, and then tested with the same test rule-word/class-word pairs used in Experiment 12.

8.2.2. Results

As shown in Fig. 10, participants preferred rule-words to class-words ($M = 67.1\%$, $SD = 12.7\%$, $t(19) = 6.0$, $p < 0.00001$). There was no difference between test pairs with initial or final overlap ($t(19) = 0.16$, $p > 0.874$, ns, paired t-test). Eight participants preferred rule-words by a binomial test, and none preferred class-words.

Participants tended to prefer rule-words more in Experiment 13 than in Experiment 12 but this trend failed to reach significance ($F(1,38) = 3.6$, $p = 0.064$, ns).

8.2.3. Discussion

Experiment 13 shows that, after being exposed to a continuous stream, participants prefer rule-words to class-words. Because the familiarization was continuous, this preference does not signal that participants extracted any generalization. Rather, our (and Peña et al.'s) results suggest that generalizations are not available after continuous familiarizations. Thus, a more plausible explanation for the preference for rule-words compared to class-words is that statistical computations over non-adjacent TPs drive the participants' responses. Just as in Peña et al.'s Experiment 4 part-words were preferred to rule-words because part-words actually occurred in the continuous familiarization stream (whereas rule-words did not), in our Experiment 13, rule-words may be preferred to class-words because the high TPs between their first and third syllables were attested in the stream, whereas no such long distance relations existed for class-words. That is, during a continuous familiarization, participants may compute statistical relations among syllables, whether adjacent or not, and mold their choices between rule-words and class-words according to these statistical relations. While these relations were adjacent TPs in Peña et al.'s Experiment 4, participants may have tracked non-adjacent TPs in our Experiment 13, leading to the preference for rule-words compared to class-words.

This explanation suggests that the results of Experiment 12, where participants preferred rule-words to class-words after a familiarization with a segmented stream, may be interpreted in two different ways. Under one interpretation, the preference for rule-words may require no independent generalizations; exposed to a segmented stream, participants may quickly extract a class-rule (as shown by Experiments 1 and 3), but, within the class-frame, they may also compute associations between the first and the last syllables of the familiarization items. Thus, they may notice that particular 'A' and 'C' syllables frequently co-occur, and favor rule-words as statistically frequent "instantiations" of class-words. Under this interpretation, no separate representation of A_iC_i-rules is required and preference for rule-words is essentially the result of the same statistical computations participants perform on a continuous stream. Under the other interpretation, A_iC_i-rules may be independently represented generalizations, maybe directly extracted by a fast extraction mechanism like the one we have described, or resulting from the pruning of class-rules to the narrower A_iC_i-rules. If so, participants may use cues in the signal to construct different types of generalizations, and respond to one or the other according to the test items presented in the experiment.

Our data are compatible with both hypotheses (although parsimony favors the hypothesis that the preference for rule-words to class-words arises from the sensitivity to non-adjacent TPs, and that A_iC_i-rules are not independently represented). In any case, under both hypotheses, a preference for A_iC_i-words may be the result of a two-step process, in which class-rules are attained first, and then statistical processes, operating on top of the class-rules, may detect that particular 'A' and 'C' syllables frequently co-occur.

Experiments 12 and 13 also offer a further control militating against the claim that phonological confounds, and not the projection of generalizations (with segmented streams) or the computation of non-adjacent TPs (with continuous streams) are responsible for Peña et al.'s (2002) and our results (e.g., Newport & Aslin, 2004; Onnis et al., 2005; Perruchet et al., 2004; Seidenberg et al., 2002). In particular, Perruchet et al. (2004) argued that participants are unable to compute non-adjacent TPs between syllables from a continuous stream, and that Peña et al.'s (2002) results are entirely due to confounds of some sort, although Perruchet et al. (2004) were unable to isolate any such factor. In Experiments 10 and 11, we have already shown that the specific phonetic characteristics of the words composing the familiarization stream does not explain participants' choices for novel items structurally well formed against familiar items structurally ill formed.

Experiments 12 and 13 offer a more complete control for phonetic and phonological factors in participants' preferences. Instead of manipulating the familiarization, as Peña et al. (2002, footnote17) and our Experiments 10 and 11 did, Experiments 12 and 13 keep the familiarization fixed but factor out the phonetic and phonological cues by using test items that share all these cues. Rule-words and class-words in a test pair overlap in either the first two syllables or the last two syllables. The syllables that differ between a rule-word and a class-word always belong to the same syllable class (either A-syllable or C-syllable). Thus, both rule-words and class-words have identical phonological and phonetic properties. In particular, the first consonants of the syllables of both rule-words and class-words share the structure Stop-Liquid-Stop. Hence, in order to prefer rule-words to class-words in the absence of any phonological differences between them, participants must have tracked relations between non-adjacent syllables. In particular, the fact that participants preferred rule-words to class-words in Experiment 13, after familiarization with a continuous stream, is a clear demonstration that they are sensitive to TPs between non-adjacent syllables.

The conclusion that in certain conditions participants can compute non-adjacent relations among items from a continuous speech stream and that they capture abstract regularities defined over classes of items from a segmented stream can also turn into powerful constraints on the available models of language learning and speech segmentation. We now turn to this topic.

9. Neural networks and generalizations

According to the MOM hypothesis, both associationist and non-associationist mechanisms may analyze speech streams: associationist mechanisms may track the syllable distribution, while a non-associationist mechanism may be responsible for extracting generalizations. A possible alternative would hold that a single associationist mechanism can explain both the fast extraction of generalizations and its loss with long familiarizations. We will now explore how a plausible associationist alternative could account for our data. We will use a Simple Recurrent Network (SRN; Elman, 1990) because it is widely used in cognitive modeling and is relatively simple. This is not to say that associationist mechanisms could be reduced to SRNs; nevertheless, we will argue that these results transcend this particular model, and that other associationist mechanisms will behave like SRNs with respect to our experiments.

A SRN is a three-layer feed-forward network augmented by a copy-layer. In each time step, the activation of the hidden layer is copied to the copy-layer; at the next time step, both the input units and the copy units feed into the hidden layer. The hidden layer receives a trace of its past activations as input, and is therefore sensitive to temporal dependencies. Variations of this model have been used to simulate many aspects of language, such as grammar acquisition (e.g., Elman, 1990), aspects of linguistic performance (e.g., Christiansen & Curtin, 1999), the learning of simple rules (e.g., Altmann, 2002), and the learning of some formal languages (e.g., Rodriguez, 2001). The network has usually the task to predict the next element in a sequence. In our case, it will have to predict the next element in the familiarization streams, and will then be tested on the test items of our experiments. For each test item, we will record the prediction for the third syllable.

In this section, we ask whether an SRN can learn to recognize class-words with segmented or continuous input. In order to sample the parameter space of the model, we investigated this issue by running 100,800 simulations in two different SRNs.

9.1. Architecture and training

Syllables were represented by pair-wise orthonormal nine- or ten-dimensional binary vectors, depending on the simulations. The networks were presented with "syllable sequences" in this format whose statistical properties were the same as in the experiments reported above; all streams contained 100 repetitions of each word, yielding 900 words in total. The network was trained with the backpropagation algorithm to predict the next element in the sequences.

It is not immediately obvious how the network should represent the "silences" that were present in the artificial speech stream participants listened to, nor what elements it should predict given a stream in which silences are explicitly represented. Because the network predicts the next element of a sequence, it may be considered that its task is to predict the next syllable of the sequence, or else, if the next element is a space in the sequence, the space itself. Peña et al.'s (2002) data (footnote 27) and our Experiment 9 suggest that participants ignore the presence of the space in the test items. Therefore, the closest match with participants' representations of the test items seems to be a network that always predicts the next syllable in a sequence, and simply uses "empty" vectors as segmentation marks.

Table 3
Summary of the simulation types used with a Simple Recurrent Network

Silence Representation	Target after Silence	Test items start with Silence	Silence Included in Part-Words
none (continuous)	–	–	–
0-vector	0-vector	–	no
0-vector	Syllable after silence	–	no
Extra unit	Silence	no	no
Extra unit	Silence	yes	no
Extra unit	Silence	no	yes
Extra unit	Silence	yes	yes

For each simulation type, the parameter space of the network was sampled extensively.

In order to explore how different representations of the input streams and the test items influence the network's ability to issue correct predictions, we tested several alternatives. Four different syllable streams were used, differing according to whether silences between words were explicitly represented or not, and to how they were represented. One stream was continuous, like in Experiment 2. In two other streams, silences were represented as "empty" vectors (that is, vectors where all activations were 0, henceforth called 0-vectors). In one of them, the network had to predict the 0-vector itself, whereas in the other it had to predict the A-syllable *following* the 0-vector. That is, given the subsequence $A_iX^{\mu}C_i0A_jX^{\nu}C_j$…, the network had to predict the silence (i.e., the "0") in one kind of simulations, and the *Aj* following the silence in the other. In the fourth stream, silences were represented by an extra unit.

The stream with the silences represented as an extra symbol was used in four different types of simulations. In the first type, silences were not included during test; in the second type they were, and therefore all test items started with a silence. The other two types of simulations were inspired by the experiment reported by Peña et al. (2002) in Footnote 27. In this experiment, test part-words were created with a silence between the 'A' and 'C' syllables, in order to control if the generalizations were due to co-occurrence statistics between syllables and silences. Thus, the third type of simulations had part-words including silences, just as in Peña et al.'s (2002) experiment. The fourth type of simulations was identical to the third, except that all test items started with silences (like in the second type of simulations). These simulation types are summarized in Table 3.

We used two networks. One had five and the other 27 hidden units. The rationale for choosing different numbers of hidden units was that this parameter may influence the network performance; indeed, choosing too few hidden units may limit the network's representational power while choosing too many hidden units may lead the network to simply memorize the stimuli (e.g., Bishop, 1995; Geman, Bienenstock, & Doursat, 1992). Our hidden layers remain in a range that has been used successfully for different tasks (see e.g., Hare, Elman, & Daugherty, 1995, for a slightly lower ratio of hidden units to output units than in our "small" network; see e.g., Elman, 1993, for a higher ratio than in our "big" network); as it turns out, the results were very similar for both networks, suggesting that this parameter may have only a limited significance for our simulations.

In order to sample the parameter space of the network, we used 360 different combinations of learning rates, numbers of training cycles and momenta. For each parameter set, we ran 20 simulations (representing 20 participants). During test, we recorded the prediction for the third syllable of test items. When the test items included silences, we asked the network to always predict the third syllable, not the silence, just as participants in Peña et al.'s (2002) experiment seem to do.

For each test item type (e.g., class-words), we used the network's success at predicting the last syllable as a measure of its familiarity with this item type. We chose this measure because only the prediction for the third syllable is meaningful for rule-words and class-words (the test item types we used). The first syllable cannot be predicted (because there is nothing it could be predicted from), but also the second syllable is unpredictable in these items. It is only the third syllable that can be predicted in principle, either because of the high TP between the first and the last syllable or because the network may have learned the class the last syllable has to belong to. Using the prediction of the last syllable as a measure of the network's familiarity with test items has been used successfully even in other simulations of simple generalizations (e.g., Altmann, 2002). While the prediction task is probably not a faithful psychological model of the tasks participants were confronted with, we are only interested in the relative performance for class-words compared to part-words, assuming that it reflects the relative familiarity of the model with these test item types. Thus, for each test item, we recorded the normalized mean activation for the target syllables (see below).

Participants were confronted with class-words as test items. In these items, the 'A' and 'C' syllables always belonged to different families, having structure A_iXC_j. In the simulations, we used the same pairings of 'A' and 'C' syllables for the class-words as in our experiments. Thus, in order to simulate the network's state of activation for class-words, we recorded the mean activation for the two C syllables of that did not follow the A syllable in the words of the familiarization. For part-words of type 12 (i.e., with the form C_iA_jX), all 'X' syllables were targets, while for part-words of type 21 (i.e., with the form XC_iA_j), only the 'A' syllables that did not belong to the same family of the preceding 'C' syllable were target. Finally, for rule-words, the target syllable was the 'C' syllable that occurred together with its 'A' syllable in the words presented during familiarization.

We then submitted to ANOVAs the normalized mean activations of the target syllables, with item-type (part-words of type 12 vs. class-words, part-words of type 21 vs. class-words, and rule-words vs. class-words) as between-subject factor. Further details of the simulations and the analyses can be found in Appendix G.

9.2. Results and discussion

Fig. 11 presents the results of the simulations. For each simulation condition, momentum value, and network, we plotted the percentage of "experiments" significantly preferring one item type or the other. Each experiment is a set of 20 simulations with a common learning rate and number of cycles; "participants" in an experiment differ by the network initializations. The results show that the network

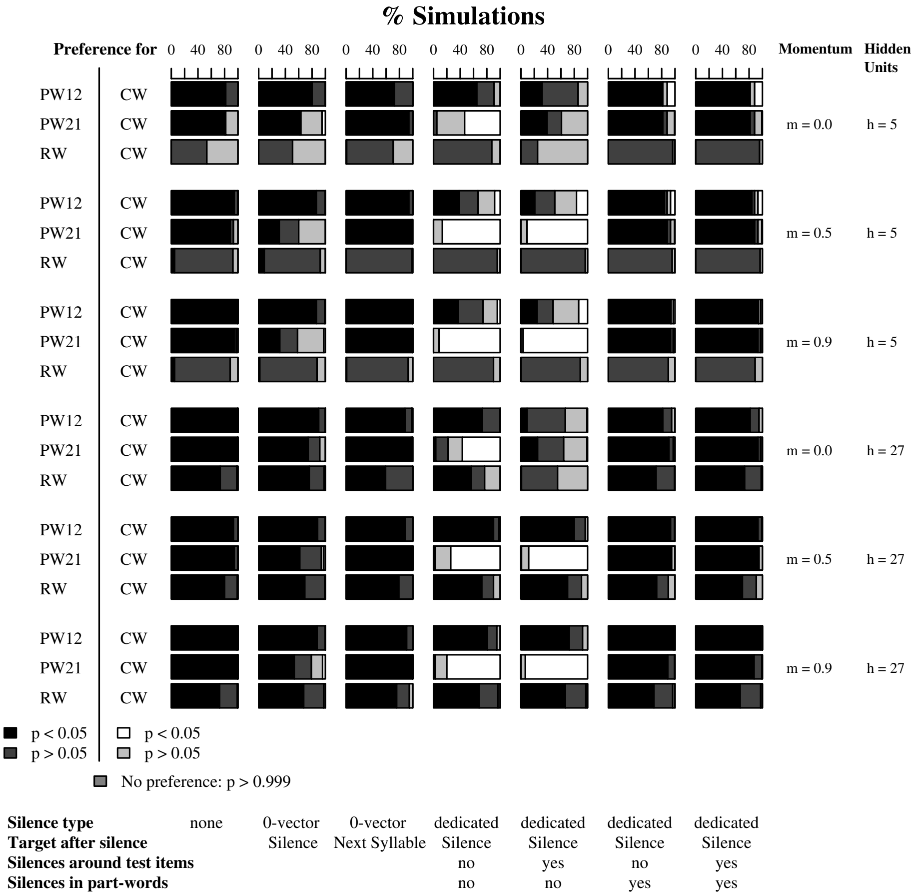

Fig. 11. Results of the simulations with a Simple Recurrent Network. Triplets of bars show comparisons between part-words of type 12 and class-words, part-words of type 21 and class-words, and rule-words and class-words (from top to bottom, respectively). Each row shows the result for a combination of a value for the momentum (0.0, 0.5 or 0.9) and a number of hidden units (5 or 27). Each column shows a type of simulation.

generally preferred part-words to class-words and rule-words to class-words. The preference for part-words to class-words suggests that a model only computing co-occurrence statistics cannot simulate the preference for class-words observed in our experiments. In contrast, the fact that rule-words can be preferred to class-words suggests that a SRN is sensitive to co-occurrences between non-adjacent items.

In some cases, the network did prefer class-words to part-words. However, this occurred only for some of those simulations where an extra-symbol represented silences *and* test part-words did not include silences, and even in those cases, only for part-words of type 21. Does this imply that the SRN can simulate participants' responses under some restricted conditions? There are several reasons to reject this conclusion. First, notice that the network predicts a difference between how part-words of type 12 and type 21 will stand the comparison with class-words; yet, in none of our experiments have we observed it. Second, Experiment 9 and the experiment reported by Peña et al. (2002) in Footnote 27 suggest that participants and the network behave differently even in those cases in which the network successfully simulates class-word preference. Indeed, while including the silences in part-words during test obliterates the preference for class-words completely in the SRN, including the silences in part-words does not prevent real participants to attain generalizations. So what is the cause of the network's limited "successes" in preferring class-words? It lies in a quirk of the representation induced by the familiarization onto the network that does not seem to affect participants. When silences are represented as extra-symbols during familiarization, the network learns that a silence follows a 'C' syllable with certainty. During the test phase, because the second syllables of part-words of type 21 are precisely 'C' syllables, the network will systematically predict an incorrect syllable, unless the silences are also included in the part-words. Instead, participants to our experiments were not affected by the presence or absence of silences in test part-words, whether they be of type 21 or of type 12. This suggests that the cause of the network's success for the few cases in which it did issue a preference for class-words has little to do with the cause of the corresponding human behavior. Thus overall SRNs, and with them other models based on co-occurrence statistics, seem unable to account for the preference for class-words we found in our experiments.

Because many statistical devices such as artificial neural networks tend to change behavior depending on the length of training, it is also worth investigating how the network simulates the dynamics of the preference for generalizations we documented. Although, as we saw, the network has difficulties at reproducing the preference for class-words in the first place, it may still be argued that the disappearance of such a preference with streams of longer duration may be simulated by the model as a consequence of over-learning (e.g., Bishop, 1995). Over-learning reflects the observation that, when a network is trained long enough, it will learn not only the "true" regularities, but also the regularities of the noise contained in the input; hence, if a network is trained for too long, its ability to capture generalizations may decrease because it starts generalizing not only the "true" regularity of its training examples, but also the regularity of the noise. While there is no obvious relationship between the amount of exposure this type of artificial neural network needs to learn a task and the amount required real by participants, we used the number of training cycles as a proxy for the familiarization length, and correlated the number of cycles with the preference for part-words.

The results of this analysis are shown in Fig. 12. The x-axis shows the natural log of the number of training cycles (because we sampled this parameter in an approximately logarithmic way), while the y-axis shows the mean F-value of the preference for class-words over part-words (both types pooled together, as in our experiments). Negative F-values indicate a preference for part-words. Overall, the simulations yielded no significant correlation, except for 14 correlations (11 positive and 3 negative). It is worth noting that none of the significant correlations yielded a reversal of the preference like the one we observed. In contrast to our participants, the networks preferred the same kinds of test items after short and long familiarizations, while participants had two clear but opposite preferences.

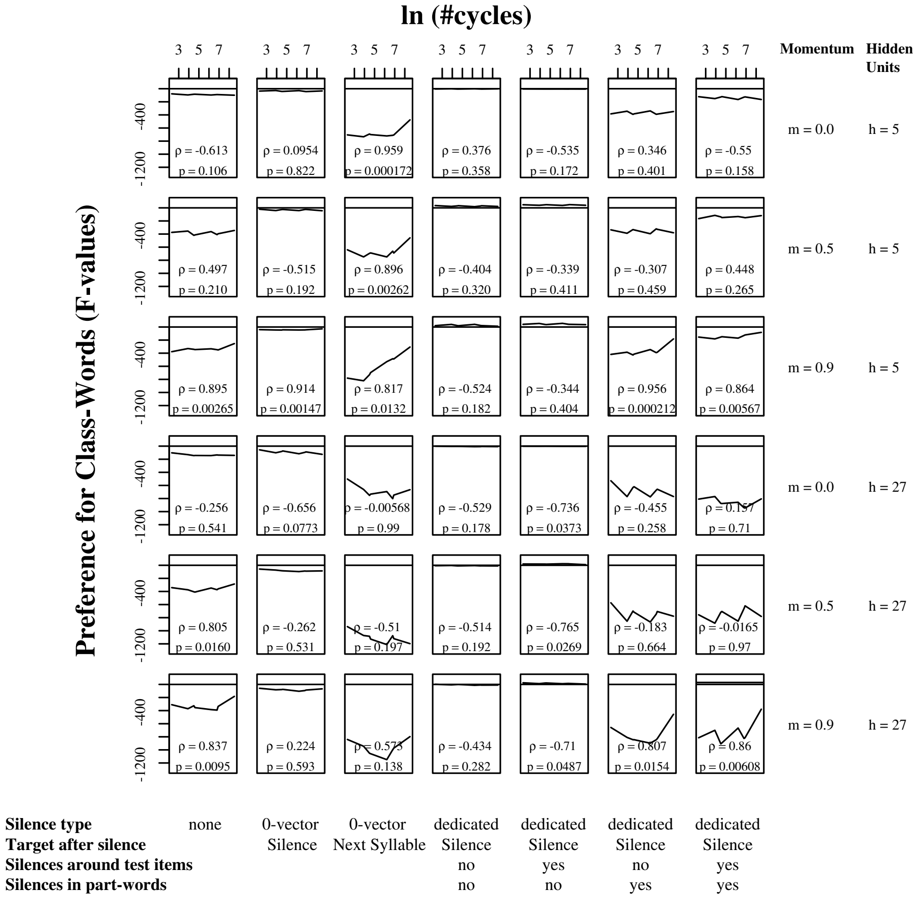

Fig. 12. Results of the simulations with a Simple Recurrent Network. The x-axis shows the natural log of the number of the training cycles, and the y-axis the F-value of the mean preference for class-words (negative values indicating a preference for part-words). Each row shows the result for a combination of a value for the momentum (0.0, 0.5 or 0.9) and a number of hidden units (5 or 27). Each column shows a type of simulation. In each cell, the correlation coefficient between the F-value and the number of training cycles and the corresponding p-value are given.

Interestingly, the 3 negative correlations were observed among the simulation types where the networks preferred class-words to part-words of type 21, namely when silences were represented by an extra-symbol during familiarization but not during test. This confirms our account of the origin of the “preference” in those marginal cases: the silences may disrupt the processing of statistical relations among syllables within part-words. Hence, it takes more training for the network to learn about these disrupted statistical dependencies. Once again, the fact that negative correlations occurred only in the peculiar combination in which silences were represented in familiarization but not in test items confirm that the network preferred part-words (at least of type 21) for reasons that are inadequate to explain our data.[8]

Together these results suggest that purely statistical mechanisms such as SRNs cannot account for the preference for class-words or for the negative correlation between the preference for class-words and the familiarization duration. In the few situations where it seems to account for these data, the network also makes predictions that are refuted by our experiments. Hence, we conclude that a single mechanism hypothesis, as implemented by a SRN or any associative device that extracts co-occurrences among items in the stream, is not adequate to explain our data.

10. General discussion

How rich is the repertoire of computational abilities that are required to learn structures as complex as those in language? While this question has been extensively studied from a formal point of view (e.g., Gold, 1967; Osherson, Stob, & Weinstein, 1986; Wexler & Cullicover, 1980), only recently have psycholinguistic investigations of language learning contributed significant advances. One important discovery has been that humans adults, infants and other animals possess powerful means of tracking the statistical distribution of items in a continuum, and can use such information to break a continuous speech stream into its constitutive words (e.g., Aslin et al., 1998; Hauser, Newport, & Aslin, 2001; Saffran, Newport, et al., 1996; Saffran, Aslin, et al., 1996). Heartened by these results, some researchers suggested that the same statistical abilities would suffice to learn all aspects of language (e.g., Bates & Elman, 1996). Another important result has been that, when exposed to similar continuous speech streams, participants cannot extract even fairly simple generalizations, even if the statistical information contained in the input is in principle sufficient to solve this task. At the same time, when the input is covertly bracketed, participants can capture the same generalizations that eluded them, and they can do so after a surprisingly short exposure to the signal (Peña et al., 2002).

In this paper, we investigated two related issues raised by these results. The first issue concerns the nature of the representations that can be extracted by passive exposure to artificial, subliminally segmented streams. We hypothesized that, when familiarized with such input, participants are not limited to processing dependencies between particular syllables (as proposed by Peña et al., 2002), but can extract a

[8] While our result show that the SRN cannot explain the dynamical aspect of our results, there are also other reasons to doubt whether an over-learning account could possibly explain the negative correlation between the preference for class-words and the familiarization duration. Indeed, over-learning arises because a network will not only learn the "true" regularities in the input, but eventually also the regularities of the noise contained in it (if trained long enough). However, for a mechanism that learns the class-rule from a distributional analysis of the streams, the only obvious candidates for playing the role of the noise are the co-occurrence statistics; after all, the cues that favor part-words to class-words are precisely such statistics. As Experiment 8 and 12 shows that this "noise" is learned perfectly well after only 2 min, it should not take 30 min for this noise to wipe out the preference for class-words; as the noise is learned earlier, it should also impair the generalizations earlier.

regularity entailing syllable classes. We suggested that these regularities are not unlike several morphological and grammatical rules that speakers have to master in order to learn a natural language. The second issue concerns the nature of the mechanisms projecting such generalizations. We argued that a single associationist mechanism does not seem to be able to account for how these regularities are extracted; in order to explain the class-based generalizations, a model invoking distinct learning mechanisms for extracting words and generalizations seems to be more adequate. We will discuss these points in turn.

10.1. Generalizations and statistical computations

To investigate the nature of the generalizations that can be extracted from a subliminally segmented stream, participants were familiarized with syllable streams containing nonce words with the structure A_iXC_i, where $A_i \ldots C_i$ were fixed syllable combinations, and X variable syllables. We asked whether participants could learn that members of one syllable class could occur as the first syllable of a word and members of another syllable class as the last syllable of a word. If so, the initial and the final syllable would act as variables quantified over distinct classes. After familiarization, we asked participants to choose either *class-words* (test items with the structure $A_iX'C_j$, where X' never appeared as a middle syllable of words in the stream) or *part-words* (syllable chunks that were encountered during familiarization but spanned a word-boundary). Class-words conformed to a class-rule but never appeared during familiarization, whereas part-words did not conform to the class-rule but appeared during familiarization and were therefore statistically favored.

Participants preferred class-words to part-words, thereby generalizing the regularity entailing classes to novel items, when they were familiarized with a stream containing subliminal silences (Experiment 1), but not when they were familiarized with a continuous, but otherwise identical, stream (Experiment 2). Experiments 10 and 11, as well as the internal controls implemented in our experiments, showed that such generalizations occurred independently of the particular phonotactic and phonological regularities contained in a stream, suggesting that participants can extract arbitrary syllable classes independent of the influence of their previous linguistic knowledge on the experiments.

Our results closely resemble Peña et al.'s (2002) results about the extraction of generalizations, with one crucial difference. Peña et al. (2002) suggested that participants extract A_iC_i-rules, that is, dependencies between syllable tokens of the form "If the first syllable is /pu/, the last syllable is /ki/". We provided evidence that participants can extract a regularity entailing syllable classes instead.

In order to study the relations between the A_iC_i-rules and class-rules, we asked under what conditions participants prefer rule-words to class-words. We showed that participants preferred rule-words to class-words after familiarizations with both continuous and segmented streams (Experiments 12 and 13). Because we (and Peña et al., 2002) showed that generalizations are available only after familiarization with segmented but not with continuous streams, the fact that participants prefer rule-words to class-words both with segmented and continuous familiarizations suggests that rule-words may not be independently represented. They may instead arise from the interaction between the extraction of generalizations over classes and statistical computations of non-adjacent relations among syllables. Such statistical computations appear to be performed on top of the generalizations.

These results speak directly to another controversial issue. Sensitivity to TPs between adjacent syllables is well documented (e.g., Aslin et al., 1998; Saffran, Newport, et al., 1996; Saffran, Aslin, et al., 1996). Peña et al. (2002) argued that participants can also compute non-adjacent TPs among syllables. Several authors have challenged this conclusion (e.g., Gómez & Maye, 2005; Newport & Aslin, 2004; Perruchet et al., 2004; Seidenberg et al., 2002), citing possible phonological or phonotactic confounds contained in Peña et al.'s (2002) experimental material. However, the rule-words and class-words in Experiments 12 and 13 shared most phonetic and phonological features. Thus, because the TP between the first and the last syllables is 1 in rule-words and 0 in class-words, participants' preference for rule-words over class-words after a familiarization with a *continuous* speech stream (Experiment 13) would be difficult to explain if human adults were not sensitive to second-order TPs.

This conclusion is compatible with several lines of evidence from sequence learning experiments (e.g., Ebbinghaus, 1885/1913), artificial grammar learning (e.g., Gómez, 2002), tip-of-the-tongue phenomena (e.g., Brown & McNeill, 1966) and research on hippocampal function (e.g., Dusek & Eichenbaum, 1997), all suggesting sensitivity to higher order relations. Taken together with this evidence, our results suggest that participants are indeed sensitive to TPs between non-adjacent items.

A system tracking TP distributions between adjacent and non-adjacent items may account for the ability to break the continuum into units even when the statistical relations among constituents are non-adjacent.[9] However, this is not necessarily the way participants learn structural relations between such constituents. We now turn to this issue.

10.2. What mechanisms extract the generalizations?

The MOM hypothesis posits that both associationist and non-associationist mechanisms analyze a speech stream, extracting different kinds of information. Three

[9] Newport and Aslin (2004) criticized this conclusion also on computational grounds, raising the possibility that grammar could be unlearnable if participants were sensitive to TPs between non-adjacent syllables; computing non-adjacent may lead to a "problem [that] grows exponentially" if the learning mechanism has to keep track of increasingly longer distance TPs (p. 129). We believe that this argument does not hold against our proposal. The argument assumes that TPs are stored in symbolic tables, which increase in size as the system tracks higher order probabilities. However, this is not the only possibility. Any system realizing correlational learning (e.g., Hebbian learning) in which syllable activations decay more slowly than syllable durations could be sensitive to second order TPs, because both adjacent and non-adjacent syllables would be active at the same time. Such a model is immune to the criticism of computational explosion. Moreover, our results suggest that extracting generalizations and sensitivity to TPs between adjacent and non-adjacent syllables are mediated by distinct processes. This can also address Newport and Aslin's (2004) computational argument: If TPs are not used to learn grammar to begin with, no learnability problem follows for grammar acquisition.

facts support the hypothesis. First, there is a sharp contrast between the successful generalizations after familiarization with a short subliminally segmented stream (Experiment 3), and the failure to generalize from longer continuous streams (Peña et al., 2002). Therefore, the mechanisms extracting the generalizations seem to operate only when the stream is bracketed, even if only subliminally. In contrast, the mechanisms tracking TPs between syllables do not appear to be subject to this limitation, and operate both on unsegmented and segmented speech streams.

The second fact comes from the dynamics of the generalizations. We showed that the preference for class-words over part-words decreases linearly with longer familiarizations, to the point that participants start preferring part-words even if the familiarization stream is covertly bracketed (Experiments 4 and 5). This result is predicted if a relatively fast, non-associationist mechanism were computing the generalizations and another component kept track of the statistical relations among items in the stream, reinforcing memory traces of the encountered tokens as experience with the stream accumulates. Acting on the basis of few examples, a fast component sensitive to the structure of tokens would be able to quickly project a generalization about the items in a stream, leading participants to accept as "legal" sequences that they have never heard. Hence, generalizations should be captured with short exposure to segmented streams, but the increasing weight of the items that did occur in the stream (sampled by a statistical mechanism) should eventually overtake the familiarity with the generalizations. This second factor would also explain the preference for part-words after familiarizations with longer streams as a consequence of the strengthening of memory traces for items encountered in the stream.

The third fact is related to the general logic of associationist computations, and was illustrated with our neural network simulations. Any measure of the strength of the associations between syllables (TPs, chunk strength, etc.) favors part-words to class-words, independently of how the syllables are represented and of any particular statistical model computing associations. As part-words appeared in the streams, while class-words contained three syllables from three different words, it is hard to see how a purely associative mechanism could possibly predict a preference for class-words. Appealing to silences in the stream as units of computation does not help either. As our Experiment 9 and Peña et al. (2002, footnote 27), show, participants' preferences for structurally correct but unheard items do not seem to be affected by the presence or the obliteration of silences in test items. Our simulations with a widely used associationist model, a Simple Recurrent Network, confirmed these considerations. In a wide range of the model's parameters, the network preferred part-words to class-words, and was able to simulate the preference for class-words only sporadically, when silences were represented with extra symbols in part-words during familiarization but not during test, and even in this case, only with part-words of type 21. Our Experiment 9 and the results by Peña et al. (2002, footnote 27) make these sporadic successful simulations psychologically implausible. We thus conclude that our participants showed a behavior that associationist devices such as SRNs cannot explain.

We also showed that even when the different effects of training length are taken into account, SRNs and other purely statistical schemes cannot explain our data.

Preference for class-words or part-words was generally flat as a function of the number of training cycles, and we never observed a reversal of preference from class-words to part-words.

There are other interpretations of our results that do not call for the MOM hypothesis. It is possible that participants exposed to short continuous streams respond only on the basis of a very partial representation of its words, thus monitoring only either the first or the last syllables of test items, which may be better represented or simply more salient. This strategy would favor class-words over part-words without the need for a mechanism sensitive to structural information. However, Experiments 6 to 8 exclude this possibility. Experiment 8 shows that participants do encode the items' middle syllables even after short familiarizations, and hence that their representation of words in the stream is rich enough to respond to the actual items, were they inclined to do so without paying attention to the structure of the test items. Experiments 6 and 7 show that participants do monitor both the first *and* the last position of test items in order to form their comparative judgments. Thus, participants' responses to our Experiments 1 and 3 cannot be explained away on the basis of deficient representations of test items or on the inability to monitor and compare all three positions of the test items.

It is also possible that models of distributional learning of linguistic categories may account for our results (e.g., Cartwright & Brent, 1997; Mintz, 2003; Mintz, Newport, & Bever, 2002; Redington et al., 1998). Applied to our experiments, such models would either record the context in which syllables can occur, and classify the syllables by comparing their contexts, or align the words from the streams in such a way that each syllable would get tagged as "first", "second" or "last". A distributional analysis of this tagged corpus would then lead to the extraction of the different syllable classes - for instance, concluding that *pu*, *be* and *ta* can occur as first syllables. While this is possible, we wonder to what extent such a proposal, if it had to account for our results as a whole, would count as an alternative to the MOM hypothesis.

Note first that models comparing the contexts of syllables can prefer class-words to part-words only if word boundaries are explicitly represented in the contexts (and if classes defined by word-boundaries are prioritized compared to classes defined by other syllables); hence also this class of models would essentially align the syllables in words. Let us now consider the constraints under which the proposal should operate. If it has to predict the outcomes of Experiments 1-5, distributional class learning should (1) get to the generalization on the basis of a limited input, and (2) move away from it when input increases. To succeed in (1), the alignment mechanism learning the appropriate categories must already be able to represent structural information without predefined physical correlates: information about word boundaries (which can be short silences, long silences, pure tones, or other arbitrary segmentation cues), information about structurally defined positions (beginnings of words, or ends of words, as an undefined number of syllables may appear between the first and last position of words), and information about the fact that arbitrary classes can be tied to these structurally defined positions. All such representational abilities would probably not be considered associative. If these are granted, then generalizations can also be represented as statistically instantiated relations between abstractly defined categories. This is to be expected: if there is a rule that determines how classes of stimuli follow each others, then instances of this rule will be frequent by definition in an highly regular input such as the one we used. Such a device possesses a great deal of pre-existing structure not extracted statistically, and can come to generalizations on the basis of few examples; one may want to call it "statistically-driven", but we believe that this is stretching the debate to the point where qualifying a mechanism as statistical may be a simple choice of terminology.

Finally, in order to succeed in (2), such a device must postulate that the same computations used to track syllable distributions do not operate on the same levels of representations that extract the generalization between syllable categories. If this were not the case, then every successive word encountered in longer streams would be further evidence validating the existence of the generalization. Hence, there would be no way for the distributional algorithm to backtrack and predict that long streams yield a preference for part-words over class-words. Thus, even assuming that the mechanism quickly extracting generalizations projects class-rules on the basis of a distributional analysis, the same analysis cannot explain the preference reversal across time. At a minimum, it must be postulated that two systems, one dedicated to analyze the relations among classes, and another keeping track of the distribution of the physical token encountered, operate on the stream. Even so, further assumptions are needed in order to explain why the former computation would be weakened by the second. Certainly, a single distributional analysis would find more confirming examples of the class-rule than of any word or part-word contained in the stream. Presumably, in order to address this last difficulty it is necessary to assume that there are distributional computations over abstract positions in the stream separate and different from the computations over single tokens. And then we would be offered essentially a dual account similar to the one we advocate here.

10.3. Artificial languages and natural language

The conclusion that processes of different types analyze speech signals is reminiscent of the proposal that regular and irregular inflectional morphology are mediated by qualitatively distinct representations (e.g., Baayen et al., 1997; Marslen-Wilson & Tyler, 1997; Pinker, 1991; Pinker & Prince, 1988). However, little evidence exists on the types of mechanisms that generate these different representations and on how such mechanisms are recruited in on-line tasks. Our results show that, even in a highly simplified artificial language, the effects of mechanisms of different nature can be observed, seemingly serving different purposes and exhibiting different temporal characteristics. Participants are sensitive to first and second-order TPs between syllables, but at the same time they can compute generalizations entailing syllable classes that associationist computations cannot account for in any obvious way. Importantly, these mechanisms can be recruited on-line and effortlessly from speech-like input alone, which may make such processes well suited for the situation a learner of a natural language faces.

Interestingly, in addition to entailing the presence of two types of processes, our results may be related to inflectional morphology for another reason. Learning syllables that can occur in word-initial or word-final positions resembles the processes of prefixation, suffixation and circumfixation. Our results may be taken to be a relatively direct demonstration of a symbolic mental operation that may support such processes. With much simpler stimuli than those used in studies of inflectional morphology, we showed that participants can learn on-line to apply an operation that resembles such morphological transformations.

How does the mechanism projecting the generalizations operate? Several hypotheses may be envisioned. One plausible hypothesis is to suppose that the generalizations may be mediated by a general mechanism representing syllables in words as variables, capable of operating under a variety of input conditions. Such a mechanism would be able to extract relations between such variables within their respective units. In this case, the fact that the silences are required for the generalizations to be drawn may be related to the bracketing hypothesis (e.g., Morgan, 1986). The silences may act as "markers" that define the units of an analysis. Such markers may be a prerequisite for dependencies between classes in speech to be analyzed, and this would explain the mechanism for generalization seems to only work over an already segmented input. The silences may also be required for inducing classes even on the hypothesis that the induction is the result of a distributional analysis. For such an analysis to succeed, multiple convergent cues to the classes appear to be required (e.g., Gerken et al., 2005; Gómez & Lakusta, 2004; Mintz, 2002; Shi et al., 1998), and the silences may be a cue making a distributional analysis possible.

Another hypothesis is that a more restricted mechanism, particularly sensitive to the relation among items in edge positions, is at work in our experiments. Endress, Scholl, and Mehler (2005) showed that simple generalizations (like whether certain structures contain repetitions) are easier to process when the repetitions are located in an edge position of a sequence than when they are located in a sequence-medial position (see also Endress & Mehler, under review). Because the syllable positions required for constructing the generalization in the present experiments were at the edges, it is possible that constrained "perceptual primitives", acting specifically on edge positions, extract the observed generalizations. Possibly, the existence of such primitives may be useful to explain some general linguistic observations such as, for example, why affixation is much more frequent than infixation across languages (e.g., Julien, 2002), or, more generally, why edges of (linguistic) constituents have to be aligned (e.g., McCarthy & Prince, 1993).

The possibility that our results may be limited to edge-positions does not vitiate the conclusion that they reflect class-based generalizations that cannot be easily explained by associationist computations. Rather, they demonstrate the action of two mechanisms: one computes TPs, while the other is most readily described in terms of non-associationist, class-based generalizations. While the mechanism responsible for generalization needs to be elucidated, our results show that speech is not analyzed by some monolithic mechanism, but that certain structural computations require particular, possibly subtle, properties of the input signal. Such selectivity may be crucial for language acquisition to succeed, and may open an avenue for investigating the computational tools involved in speech stream analysis.

Appendix A. Test items for Experiments 1–5

Table A1

Test items for Experiments 1 to 5

Class-words	Part-words
putaga	kibefo
puduga	likibe
pubedu	kitali
pugadu	Rakita
betaki	gapuRa
beduki	fogapu
bepudu	fogata
bekidu	gatali
tabeki	Radupu
tagaki	dupufo
tapuga	lidube
takiga	dubeRa

Appendix B. Test items for Experiment 6

Table B1

Test items for Experiment 6

Legal syllable	
Initial	Final
bedufo	fobedu
beduli	libedu
beduRa	Rabedu
bekifo	fobeki
bekili	libeki
bekiRa	Rabeki
pudufo	fopudu
puduli	lipudu
puduRa	Rapudu
pugafo	fopuga
pugali	lipuga
pugaRa	Rapuga
tagafo	fotaga
tagali	litaga
tagaRa	Rataga
takifo	fotaki
takili	litaki
takiRa	Rataki

Appendix C. Test items for Experiment 7

Table C1

Test items for Experiment 7

Class-word	Foil	Foil-type
putaga	putabe	$A_iX'A_j$
puduga	pudube	$A_iX'A_j$
pubedu	pubeta	$A_iX'A_j$
pugadu	pugata	$A_iX'A_j$
betaki	betapu	$A_iX'A_j$
beduki	bedupu	$A_iX'A_j$
bepudu	beputa	$A_iX'A_j$
bekidu	bekita	$A_iX'A_j$
tabeki	tabepu	$A_iX'A_j$
tagaki	tagapu	$A_iX'A_j$
tapuga	tapube	$A_iX'A_j$
takiga	takibe	$A_iX'A_j$
putaga	kitaga	$C_iX'C_j$
puduga	kiduga	$C_iX'C_j$
pubedu	kibedu	$C_iX'C_j$
pugadu	kigadu	$C_iX'C_j$
betaki	gataki	$C_iX'C_j$
beduki	gaduki	$C_iX'C_j$
bepudu	gapudu	$C_iX'C_j$
bekidu	gakidu	$C_iX'C_j$
tabeki	dubeki	$C_iX'C_j$
tagaki	dugaki	$C_iX'C_j$
tapuga	dupuga	$C_iX'C_j$
takiga	dukiga	$C_iX'C_j$

Appendix D. Test items for Experiment 8

Table D1

Test items for Experiment 8

Words	Rule-Words
puliki	pubeki
pufoki	pubeki
puliki	pugaki
puraki	pugaki
puraki	putaki
pufoki	putaki
beliga	bepuga
befoga	bepuga
beliga	bekiga
beraga	bekiga
beraga	beduga
befoga	beduga
talidu	tabedu
taradu	tabedu

(*continued on next page*)

Appendix D (continued)

Words	Rule-Words
taradu	takidu
tafodu	takidu
talidu	tagadu
tafodu	tagadu

Appendix E. Test items for Experiments 10 and 11

Table E1

Test items for Experiments 10 and 11

Class-words	Part-words
toliso	dubaga
tofeso	piduba
tobafe	dulipi
tosofe	muduli
balidu	sotomu
bafedu	gasoto
batofe	gasoli
badufe	solipi
libadu	mufeto
lisodu	fetoga
litoso	pifeba
liduso	febamu

Appendix F. Test items for Experiments 12 and 13

Table F1

Test items for Experiments 12 and 13

Rule-words	Class-words
beduga	beduki
bekiga	takiga
bepuga	bepudu
pubeki	pubedu
pugaki	tagaki
putaki	betaki
tabedu	tabeki
takidu	bekidu

Appendix G. Details about the simulations

All simulations were performed with the Stuttgart Neural Network Simulator (Version 4.2, http://www-ra.informatik.uni-tuebingen.de/SNNS/) compiled on an

Apple Dual G5 computer running Mac OS X. The simulations were automatically launched and analyzed by a set of SH, Gawk, Perl and R scripts.

G.1. Architecture

We used a Simple Recurrent Network (SRN; Elman, 1990) with 5 or 27 hidden units and nine input and output units. The simulations representing the silence as an extra-symbol used 10 input and output units.

G.2. Material

Syllables were represented by pair-wise orthonormal binary vectors. The networks were presented with "syllable sequences" in this format whose statistical properties were the same as in the experiments reported above; all streams contained 100 repetitions of each word, yielding 900 words in total. Four different syllable streams were used that differed in the presence or absence of "silences" between words and in the way silences were represented. One stream was continuous, like in Experiment 2. In two other streams, silences were represented as 0-vectors. In one of these streams, the network had to predict the 0-vector itself, in the other it had to predict the A-syllable following the 0-vector. In the fourth stream, silences were represented by an extra unit; the latter stream was used with four different test regimens (see below).

G.3. Training

The network was trained with the backpropagation algorithm to predict the next element in the sequence. In order to sample the parameter space of the network, all syllable streams were presented in 360 different sets of simulations with the learning rates: 1×10^{-5}, 5×10^{-5}, 9×10^{-5}, 1×10^{-4}, 5×10^{-4}, 9×10^{-4}, 1×10^{-3}, 5×10^{-3}, 9×10^{-3}, 1×10^{-2}, 5×10^{-2}, 9×10^{-2}, 1×10^{-1}, 5×10^{-1}, 9×10^{-1}, the numbers of training cycles: 10, 50, 90, 100, 500, 900, 1000, 5000, and the momenta: 0.0, 0.5, 0.9. Each parameter set was evaluated with 20 simulations, representing 20 participants. Since we used two networks and four different training streams, one of which was evaluated in four training regimen, this yielded 100,800 simulations.

G.4. Test

We assessed how well the networks predicted the third syllable for different test items; silence (that were included in the test items in some simulations) were not counted as syllables for determining which syllable had to be predicted. In the simulations without a special unit for the silences as well as in half of the simulations with such a unit, test items always started with their first syllable; in the remaining simulations with such a unit, test items started with the silence preceding the first syllable. Of primary interest is the relative performance for rule-words, class-words and part-words. Part-words could be of two types: either they consisted of the last syllable of the first word and the first two syllables of the second word ("type 12"), or they consisted of the last two syllables of the first word and the first syllable of the second word ("type 21"). This distinction is important since the two part-word types have different numbers of target syllables. Indeed, part-words of type 12 have three different target syllables since their last syllable is a X-syllable, and can thus take three possible values. In contrast, part-words of type 21 have two different target syllables since its last syllable is an A-syllable, and only A-syllables that do not belong to the family of the previous word are allowed.

G.5. Evaluation

The network performance was evaluated by computing the cosine of the angle between the target output (a vector where all target units were set to 1 and the other activations to 0) and the actual output; the cosine was then divided by the number of target syllables, yielding the normalized mean activation of the target syllables. These scores were averaged for all test items of a given type, each average representing a "participant"; 20 averages were obtained with different weight initializations. The performance for the different test item types was then compared in one-way ANOVAs. Of primary interest were comparisons between the performances for class-words and part-words of both types and between rule-words and class-words.

Appendix H. Test items in Peña et al.'s (2002) footnote 17

Table H1
Test used in Peña et al. (2002, footnote 17)

Experiment 1[a]		Control (footnote 17)[b]	
Words	Part-words	Words	Part-words
beliga	kitafo	*fogata*	**befoga**
tafodu	*Ragapu*	**likibe**	*dupufo*
pufoki	gapufo	*Rakipu*	**beRaga**
puRaki	*fogapu*	*Radupu*	**pufoki**
befoga	lidube	*fokita*	**taRadu**
talidu	likita	*ligabe*	*Gapuli*
taRadu	*Radube*	**lidube**	*kibeRa*
beRaga	kitaRa	*foduta*	**Talidu**
puliki	dubeRa	**Ragapu**	**Puliki**

The test items in this table were used in Peña et al.'s (2002) Eexperiment 1 and its control in footnote 17. Because words are organized into families, it is not possible to exactly invert all words and part-words. The control uses the closest possible match to a full inversion. Items in bold invert their word/part-word status exactly from the experiment to its control; items in italics change, at most, of one syllable in one position from the experiment to its control. Results of the experiments are reported at the bottom of the table.

[a] Preference for words: 73% ($p < 0.0005$).

[b] Preference for words: 58% ($p < 0.02$).

References

Altmann, G. T. (2002). Learning and development in neural networks – the importance of prior experience. *Cognition, 85*(2), B43–B50.

Aslin, R. N., Saffran, J., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324.

Baayen, R., Dijkstra, A., & Schreuder, R. (1997). Singulars and plurals in dutch: evidence for a parallel dual route model. *Journal of Memory and Language, 37*, 94–117.

Bates, E., & Elman, J. L. (1996). Learning rediscovered. *Science, 274*(5294), 1849–1850.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2006). How to hit Scylla without avoiding Charybdis: comment on Perruchet, Tyler, Galland, and Peereman (2004). *Journal of Experimental Psychology: General, 135*(2), 314–321 discussion 322–326.

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

Cartwright, T., & Brent, M. (1997). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition, 63*(2), 121–170.

Christiansen, M., & Curtin, S. (1999). Transfer of learning: rule acquisition or statistical learning? *Trends in Cognitive Science, 3*(8), 289–290.

Cook, V. J., & Newson, M. (1996). *Chomsky's universal grammer*. Oxford, UK: Blackwell.

Dusek, J., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences of the United States of America, 94*(13), 7109–7114.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Vreken, O. van der. (1996). The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of the fourth international conference on spoken language processing (Vol. 3, pp. 1393–1396)*. Philadelphia.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. New York: Teachers College, Columbia University <http://psychclassics.yorku.ca/Ebbinghaus/>.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition, 48*(1), 71–99.

Endress, A. D., & Mehler, J. (under review). Perceptual constraints in phonotactic learning. *Journal of Memory and Language*.

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), 406–419.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4*(1), 1–58.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32*(2), 249–268.

Gillette, J., Gleitman, H., Gleitman, L. R., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73*(2), 135–176.

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*, 447–474.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*(5), 431–436.

Gómez, R. L., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science, 7*(5), 567–580.

Gómez, R. L., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy, 7*(2), 183–206.

Hare, M., Elman, J. L., & Daugherty, K. (1995). Default generalization in connectionist networks. *Language and Cognitive Processes, 10*, 601–630.

Harris, Z. S. (1955). From phoneme to morpheme. *Language, 31*, 190–222.

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–B64.

Hayes, J. R., & Clark, H. H. (1970). Experiments in the segmentation of an artificial speech analog. In J. R. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley.

Julien, M. (2002). *Syntactic heads and word formation*. New York: Oxford University Press.

Kelly, M. (1996). The role of phonology in grammatical category assignments. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax. Bootstrapping from speech to grammar in early acquisition* (pp. 249–262). Mahwah, NJ: Lawrence Erlbaum.

Landau, B., & Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child*. Cambridge, MA: Harvard University Press.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: the exception that proves the rule. *Cognitive psychology, 29*(3), 189–256.

Marcus, G. F., Pinker, S., Ullman, M. T., Hollander, M., Rosen, T., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development, 57*(4), 1–182.

Marslen-Wilson, W., & Tyler, L. (1997). Dissociating types of mental computation. *Nature, 387*(6633), 592–594.

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

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition, 30*(5), 678–686.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*(1), 91–117.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*(4), 393–424.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition, 96*(2), 143–182.

Morgan, J. L. (1986). *From simple input to complex grammar*. Cambridge, MA: MIT Press.

Morgan, J. L., Meier, R., & Newport, E. L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior, 20*, 67–85.

Morgan, J. L., Shi, R., & Allopenna, P. (1996). Perceptual bases of rudimentary grammatical categories: Towards a broader conceptualization of bootstrapping. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax. Bootstrapping from speech to grammar in early acquisition* (pp. 263–283). Mahwah, NJ: Lawrence Erlbaum.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de donnees lexicales du français contemporain sur internet: Lexique. *L'Année Psychologique, 101*, 447–462 <http://www.lexique.org>.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive psychology, 48*(2), 127–162.

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

Osherson, D., Stob, M., & Weinstein, S. (1986). *Systems that learn, an introduction to learning theory for cognitive and computer scientists*. Cambridge, MA: MIT-Press.

Pallier, C., Dupoux, E., & Jeannin, X. (1997). EXPE: An expandable programming language for online psychological experiments. *Behavior Research, Methods, Instruments, and Computers, 29*, 322–327.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298*(5593), 604–607.

Perruchet, P., Tyler, M. D., Galland, N., & Peereman, R. (2004). Learning nonadjacent dependencies: no need for algebraic-like computations. *Journal of Experimental Psychology: General, 133*(4), 573–583.

Pinker, S. (1991). Rules of language. *Science, 253*(5019), 530–535.

Pinker, S., & Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*(1-2), 73–193.

Prasada, S., & Pinker, S. (1993). Generalizations of regular and irregular morphology. *Language and Cognitive Processes, 8*, 1–56.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425–469.

Rodriguez, P. (2001). Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Computation, 13*(9), 2093–2118.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language, 35*, 606–621.

Seidenberg, M. S., MacDonald, M. C., & Saffran, J. R. (2002). Neuroscience: Does grammar start where statistics stop? *Science, 298*(5593), 553–554.

Selkirk, E. (1996). The prosodic structure of function words. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax. Bootstrapping from speech to grammar in early acquisition* (pp. 187–213). Mahwah, NJ: Lawrence Erlbaum.

Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. *Journal of Child Language, 25*(1), 169–201.

Shi, R., Werker, J., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 72*(2), B11–B21.

Wexler, K., & Cullicover, P. (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences, 8*(10), 451–456.