Generalization in Artificial Language Learning: Modelling the Propensity to Generalize

Raquel G. Alhama (rgalhama@uva.nl)
Willem Zuidema (zuidema@uva.nl)
Institute for Logic, Language and Computation; University of Amsterdam

Abstract

Experiments in Artificial Language Learning have revealed much about the ability of human adults to generalize to novel grammatical instances (i.e., instances consistent with a familiarization pattern). Notably, generalization appears to be negatively correlated with the amount of exposure to the artificial language, a fact that has been claimed to be contrary to the predictions of a statistical mechanism (Peña, Bonatti, Nespor, and Mehler (2002); Endress and Bonatti (2007)). In this paper, we propose to model generalization as a three-step process involving: i) memorization of segments of the input, ii) computation of the probability for unseen sequences, and iii) distribution of this probability among particular unseen sequences. Applying two probabilistic models for steps (i) and (ii), we can already explain relevant aspects of the experimental results. We also demonstrate that the claim about statistical mechanisms does not hold when generalization is framed under the 3-step approach; concretely, a statistical model of step (ii) can explain the decrease of generalization with exposure time.

Keywords: artificial grammar learning; statistical learning; rule learning; computational modelling; cognitive modelling

Introduction

In the last twenty years, experiments in Artificial Language Learning have become increasingly popular for the study of the basic mechanisms that operate when subjects are exposed to language-like stimuli. Thanks to these experiments, we know that 8 month old infants can segment a speech stream by extracting statistical information (Saffran, Aslin, & Newport, 1996), and it has been shown that they can do it solely relying on the transitional probabilities between adjacent syllables (Aslin, Saffran, & Newport, 1998). This ability also seems to be present in human adults (Saffran, Newport, & Aslin, 1996), and to some extent in nonhuman animals like cotton-top tamarins (Hauser, Newport, & Aslin, 2001) and rats (Toro & Trobalón, 2005).

Peña et al. (2002) present a clever experimental design that is aimed to test generalization skills[1] in a segmentation task. The stimuli used in their experiments consists of a sequence of artificial *words* that obey a certain pattern (namely, a non-adjacent dependency between the first and last syllables of each word). This setup is particularly suitable to test whether words have been extracted, but also whether participants generalize to unseen sequences that are consistent with this pattern.

In the following section, we summarize the experiments reported by Peña et al., as well as some follow-up experiments (Endress and Bonatti (2007); Frost and Monaghan (2016)). We then present a way to think about the process of generalization that identifies 3 main steps: memorization of segments, computation of the probability of unseen sequences, and generalization to particular unseen sequences. Next, we use the Retention&Recognition model (Alhama, Scha, & Zuidema, 2016) to model the first step, and Simple Good-Turing (Gale & Sampson, 1995) for the second step. We show that modelling only the first two steps we can already explain the pattern of results found in the experiments. Finally, we discuss the implications of our study for hypothesis about the nature of the generalization mechanism.

[1]The type of generalization we are interested in is commonly referred to in the ALL literature as 'rule learning'. We prefer the term 'generalization' because 'rule-learning' can be confused with a particular theory of generalization that claims that the mental structures used in a generalization process have the form of algebraic rules.

Experimental Record

Peña et al. (2002) conduct a series of Artificial Language Learning (ALL) experiments in which French-speaking adults are familiarized to a synthesized speech stream consisting of a sequence of artificial *words*. Each of these words contains three syllables AXC such that the A syllable co-occurs with the C syllable, forming a non-adjacent dependency. The order of the words in the stream is randomized, with two constraints: (i) words belonging to the same 'family' (i.e., words with the same A and C syllables) do not appear consecutively, and (ii) words with the same middle syllable X do not appear consecutively.

stream	pulikiberagatafodupurakibefogatalidu ...
words A_iXC_i	puliki, beraga, tafodu, puraki, befoga, ...
part-words C_jA_iX, XC_iA_j	kibera, ragata, gatafo, fodupu, dupura, ...
rule-words A_iYC_i	pubeki, beduga, takidu, ...
class-words A_iYC_j	pubedu, betaki, tapuga, ...
rule*-words A_iZC_i	puveki, bezoga, tathidu, ...

Table 1: Summary of the stimuli used in the depicted experiments.

The participants are subsequently tested in a two-alternative forced choice task where they have to make choices between two items: a word (AXC) versus a *part-word* (an ill-segmented sequence of the form XCA or CAX, as shown in table 1), or a word versus a *rule-word* (a rule-obeying AYC sequence such that Y appears in the stream as an A or a C syllable). The participants were asked to choose

the item that looked more like a word from the artificial language they had been familiarized with.

In their baseline experiment, the authors expose the participants to a 10 minute stream of AXC words. In the subsequent test phase, the subjects show a significant preference for words over part-words, proving that the words could be segmented out of the familiarization stream. Next, the experiment is replicated, with the exception that the test now involves a choice between a part-word and a rule-word. The subjects' responses in this experiment do not show a significant preference for either part-words or rule-words, suggesting that participants do not generalize to novel grammatical sequences. However, when the authors insert micropauses of 25ms between the words, the participants do show a preference for rule-words over part-words. A shorter familiarization (2 minutes) containing micropauses also results in a preference for rule-words; in contrast, a longer familiarization (30 minutes) without the micropauses results in a preference for part-words. In short, the presence of micropauses seems to facilitate generalization to rule-words, while the amount of exposure time correlates negatively with this capacity.

Endress and Bonatti (2007) report a range of experiments with the same familiarization procedure used by Peña et al. However, their test for generalization is based on *class-words*: unseen sequences that start with a syllable of class "A" and end with a syllable of class "C", but with A and C not appearing in the same triplet in the familiarization (and therefore not forming a nonadjacent dependency).

From the extensive list of experiments conducted by the authors, we will refer only to those that test the preference between words and class-words, for different amounts of exposure time. The results for those experiments (illustrated in figure 1) also show that the preference for generalized sequences decreases with exposure time. For short exposures (2 and 10 minutes) there is a significant preference for class-words; when the exposure time is increased to 30 minutes, there is no preference for either type of sequence, and in a 60 minutes exposure, the preference reverses to part-words.

Finally, Frost and Monaghan (2016) show that micropauses are not essential for rule-like generalization to occur. Rather, the degree of generalization depends on the type of test sequences. The authors notice that the middle syllables used in rule-words might actually discourage generalization, since those syllables appear in a different position in the stream. Therefore, they test their participants with *rule*-words*: sequences of the form AZC, where A and C co-occur in the stream, and Z does not appear. After a 10 minute exposure without pauses, participants show a clear preference for the rule*-words over part-words of the form ZCA or CAZ.

In summary, the pattern of results in these experiments shows: i) generalization for a stream without pauses is only shown for rule*-words, but not for rule-words nor class-words; ii) the preference for rule-words and class-words is boosted if micropauses are present; iii) increasing the amount of exposure time correlates negatively with generalization to rule-words and class-words (with differences depending on the type of generalization and the presence of micropauses, as can be seen in figure 1). This last observation, which we call the *time effect*, is the phenomenon we want to account for in this paper.

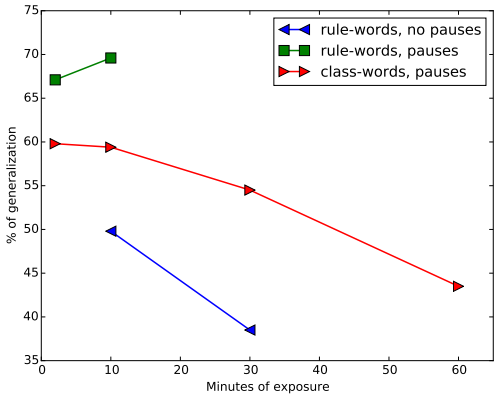


Figure 1: Percentage of generalization for rule-words and class-words, in the experiments reported in Peña et al. (2002) and Endress and Bonatti (2007), for different exposure times to the familiarization stream.

Understanding the generalization mechanism: a 3-step approach

Peña et. al interpreted their results as showing that generalization requires a separate mechanism (other than the statistical mechanism used for extracting the words). Endress and Bonatti claim that this separate mechanism cannot be based on statistical computations. The authors predict that a statistical mechanism would benefit from increasing the amount of exposure, based on the assumption that more evidence entails better representations that should encourage generalization. They call this the *More-than-One-Mechanism* hypothesis, or *MoM*.

We consider that, in order to formulate hypotheses about the workings of generalization, we should postulate a concrete framework that defines the main steps involved in it. We propose a conceptualization of generalization as a three-step process (illustrated in figure 2). According to the three-step approach, a model of generalization in ALL should involve the following steps:

(i) **Memorization:** Build up a memory store of segments with frequency information (i.e., compute subjective frequencies).

(ii) **Quantification of the propensity to generalize:** Depending on the frequency information from (i), decide how likely are other unseen types.

(iii) **Distribution of probability over possible generalizations:** Distribute the probability for unseen types com-

puted in (ii), assigning a probability to each generalized sequence.

Crucially, we believe that step (ii) has been neglected in cognitive approaches to generalization. This step accounts for the fact that generalization is not only based on the particular structure underlying the stimuli, but also depends on the statistic properties of the input. At this point, we can already reassess the MoM hypothesis: more exposure time does entail better representation of the stimuli (as would be reflected in step (i)), but its impact on generalization depends on the model used for step (ii). As we show later, modelling only steps (i) and (ii) we can already predict the *time effect*, based solely on statistical information of the input stream (although without step (iii) we cannot expect to obtain a precise quantitative fit to the data).

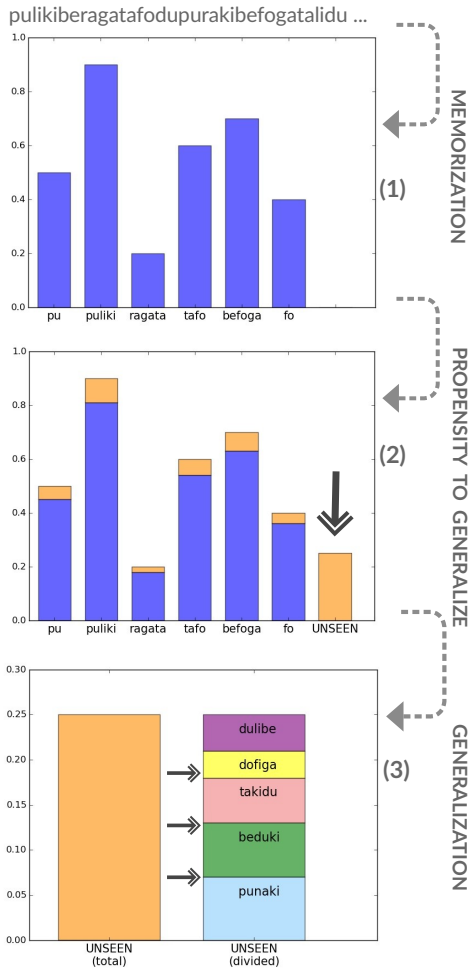
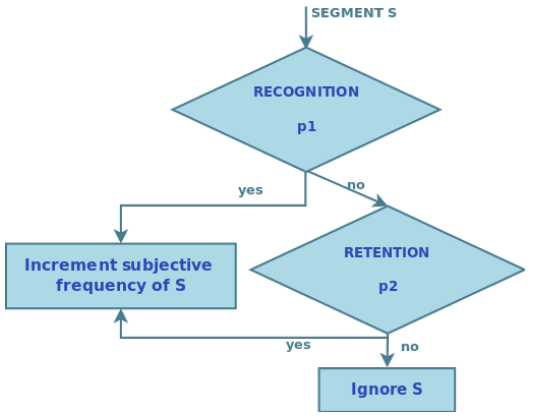


Figure 2: Three step approach to generalization.

Memorization of segments: the *Retention and Recognition* model

The Retention and Recognition model Alhama et al. (2016) was proposed as a model of memorization of segments from an auditory sequence. The model consists of an initially empty memory, and two mechanisms: retention and recognition. These mechanisms store segments from the input sequence in the memory, together with a count of how many times the segment has been retained or recognized. We refer to these counts as "subjective frequencies".

The model is presented with the ordered set of subsequences of the input, of any length. Each one of these segments is processed as shown in figure 3: first the recognition mechanism attempts to recognize the segment (that is, it attempts to determine whether the segment corresponds to one of the segments already in memory). The probability for successful recognition is P_1. If the attempt succeeds, the subjective frequency (*count*) of the segment in memory is increased with one. If the segment was not recognized, the model may still retain it, with probability P_2. If it does, the segment will be added to the memory (or, if already there from a previous iteration, its subjective frequency is increased with one). If not, the segment is ignored, and the next subsequence of the stream will be processed.



$$p_1(s) = (1 - B^{activation(s)}) \cdot D^{\#types}$$
$$p_2(s) = A^{length(s)} \cdot C^{\pi}$$
$$0 \leqslant A, B, C, D \leqslant 1; \quad \pi = \begin{cases} 0 \text{ after a pause} \\ 1 \text{ otherwise} \end{cases}$$

Figure 3: R&R:The Retention-Recognition Model

The probabilities of the model involve free parameters (A, B, C and D) that may be set based on empirical data. The recognition probability P_1 of a segment s depends on its $activation(s)$ in the internal memory (at the moment, the activation corresponds to the subjective frequency), so that segments with greater subjective frequency are easier to recognize. However, the number of different segments in memory

(#*types*) makes the recognition task more difficult. The retention probability P_2 is larger for shorter segments; in addition, the probability is attenuated (as indicated by parameter C) unless the segment is preceded by a pause.

Figure 4 shows the subjective frequencies computed by R&R after an exposure to a 10 minute familiarization, for an arbitrary parameter setting. As it can be seen, the distribution of subjective frequencies is skewed, with a prominent presence of part-words. We discuss later the effect of the skew in determining the propensity to generalize (step 2 of the proposed 3-step approach).

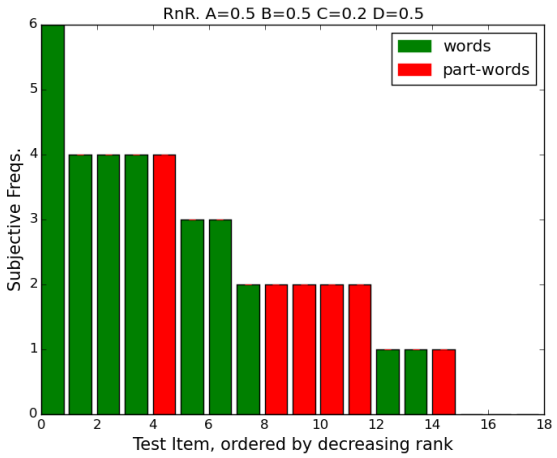


Figure 4: Subjective frequencies computed by the R&R model (A=0.5, B=0.5, C=0.2, D=0.5), for an exposure of 10 minutes (without pauses).

Quantifying the propensity to generalize: the Simple Good-Turing method

In probabilistic modelling, generalization must necessarily involve shifting probability mass from attested events to unattested events (referred to as *smoothing*). But how do we compute the amounts of probability mass that will be shifted, and thus the propensity to generalize?

Good and Turing (Good, 1953) define a method that, given a finite sample of a population of tokens belonging to different types, computes the probability that the next token drawn from that population belongs to a certain type, including the possibility of the token belonging to a type that was not included in the sample and was therefore unattested. We apply the Good-Turing method to subjective frequencies of the R&R model. Types are thus *segment types* (e.g., *talidu*), while tokens are particular occurrences of those segments (e.g., the first occurrence of *talidu* in the stream).

In the field of Natural Language Processing, Simple Good Turing (Gale & Sampson, 1995), a computationally efficient implementation of the Good-Turing method, is well known and widely used to smooth the probabilities (as computed with maximum likelihood estimation) of words in a language model, with the purpose of reserving some probability mass for words that did not appear in the finite corpus from which the model was inferred.

The method works as follows: we use the subjective frequencies r computed by R&R and, for each of them, we compute the frequency of that frequency (N_r), that is, the number of triples that have a certain subjective frequency r. The values N_r are then approximated by a continuous downward-sloping line in log space. These approximated values $E(N_r)$ are used to reestimate the frequencies according to (1).

$$r^* = (r+1)\frac{E(N_{r+1})}{E(N_r)} \tag{1}$$

The reestimated frequency r^* is then used to reestimate the probabilities:

$$P_r = \frac{r^*}{N} \tag{2}$$

The reestimated probabilities sum up to one when including the probability for unseen sequences. This probability is defined as follows:

$$P_0 = \frac{E(N_1)}{N} \tag{3}$$

This probability P_0 (also known as *missing mass*) corresponds to what we have called "propensity to generalize".

The Simple Good-Turing method is designed to ensure that the probability for unseen types is similar to the probability of types with frequency one. The propensity to generalize is therefore greater for distributions where most of the probability mass is for smaller frequencies. This obeys a rational principle: when types have been observed with high frequency, it is likely that all the types in the population have already been attested; on the contrary, when there are many low-frequency types, it may be expected that there are also types not yet attested.

Results

Memorization of words and part-words

First we analyze the effect of the different conditions (exposure time and presence of pauses) in the memorization of segments computed with R&R (step (i)). Figure 5 shows the presence of words and part-words in the memory of R&R after different exposure times (average out of ten runs of the model). As can be seen, the subjective frequencies of part-words increase over time, and thus, the difference between words and part-words decreases as the exposure increases.

The graph also shows that, when the micropauses are present, words are readily identified after much less exposure, yielding clearer diferences in subjective frequencies between words and part-words.

The results of these simulations are consistent with the experimental results: the choice for words (or sequences generalized from words) against part-words should benefit from shorter exposures and from the presence of the micropauses. Now, given the subjective frequencies, how can we compute the propensity to generalize?

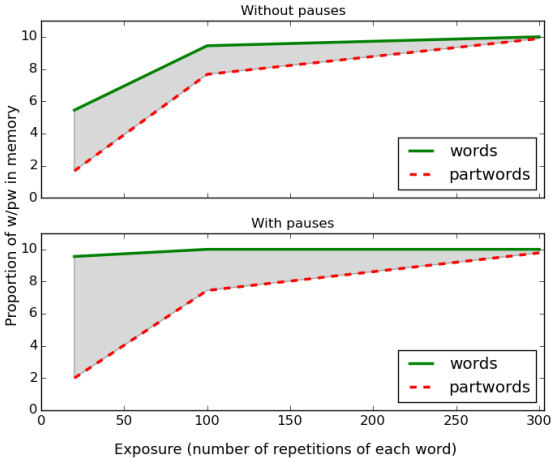


Figure 5: Average number of present words (green continuous line) and part-words (red dashed line) in 10 runs of a model with an arbitrary parameter setting (A=0.5 B=0.5 C=0.2 D=0.5).

Prediction of observed decrease in the propensity to generalize

Next, we apply the Simple Good-Turing method [2] to subjective frequencies computed by the R&R model. As shown in figure 6, we find that the propensity to generalize decreases as the exposure time increases, regardless of the parameter setting used in R&R. This result is consistent with the rationale in the Simple Good-Turing method: as exposure time increases, frequencies are shifted to greater values, causing a decrease in the smaller frequencies and therefore reducing the expectation for unattested sequences.

The results of these simulations point to a straightforward explanation of the experimental finding of a reduced preference for the generalized sequences: longer exposures repeat the same set of words (and partwords), and consequently, participants may conclude that there are no other sequences in that language - otherwise they would have probably appeared in such a long language sample.

It can be noticed in the graphs that the propensity to generalize is smaller for the micropause condition. The reason for that is that R&R identifies words quicker when micropauses are present, and therefore, the subjective frequencies tend to be greater. This is consistent with the results reported in Frost and Monaghan (2016), which show that micropauses are not needed for a certain type of generalization (concretely, for rule*-words). As the authors suggest, one plausible conclusion is that rule-words and class-words are constructed in a way that discourages generalization (due to the middle syllable occupying either an A or C position in the familiarization stream), but the micropauses compensate it by enhancing the salience of the initial and final syllables (A and C).

[2]We use the free software implementation of Simple Good Turing in https://github.com/maxbane/simplegoodturing.

Discussion

The experiments that we are addressing are all based on the same simple language, but the results form a complex mosaic: generalization is observed in different degrees depending on the amount of exposure, the presence of micropauses and the type of generalization (rule-words, class-words or rule*-words). We have approached the analysis of these results with the use of several tools: first, with the 3-step approach, a conceptualization of generalization that identifies its main components; second, with the use of R&R, a probabilistic model that already predicts some aspects of the results -and, importantly, generates a skewed distribution of subjective frequencies that is crucial for step (ii) (as we will discuss next); and third, with the Simple Good-Turing method for quantifying the propensity to generalize. We now discuss how we interpret the outcome of our study.

Framing generalization with the 3-step approach allowed us to identify a step that is usually neglected, namely, the computation of the propensity to generalize. We state that generalization is not only a process of discovering structure: the frequencies in the familiarization generate an expectation for unattested items, and the responses for generalized sequences must be affected by it. Moreover, this step is based on statistical information, proving that —contrary to other claims (Endress & Bonatti, 2007)— a statistical mechanism can account for the negative correlation with exposure time.

One issue to discuss is whether the results on the propensity to generalize depend on the use of the R&R model for computing the subjective frequencies. As we mentioned before, the frequency distributions computed by R&R are typically skewed. The Simple Good-Turing is actually designed to fit natural language frequencies, which are known to be Zipfian; although the distributions computed with R&R are not necessarily Zipfian, the skew makes them more suitable (a fact that becomes specially relevant with the use of artificial languages, which have frequency distributions very different from natural language).

To further illustrate this point, we have applied the Simple Good-Turing to frequency distributions computed with the Bayesian model presented in Goldwater, Griffiths, and Johnson (2006).[3] The results of three runs of the model are shown in figure 7. As it can be seen, the propensity to generalize is almost non-existent. Rather than decreasing, it either remains the same or even increases. The reason for this is the (almost) nonexistent amount of types with small frequency. As an example, all the runs of the model for the 10 minutes exposure result in a perfect memorization of the nine words, and no other segments; therefore, the resulting lexicon consists only of 9 types of frequency 100 (and consequently, a uniform frequency distribution without any degree of skew).

Finally, we reiterate that have accomplished our goal qualitatively. We capture the downward tendency of the propensity to generalize, but a model for step (iii) is required to

[3]We use the version of the model that exploits bigram word dependencies, with the default parameters.

also achieve a quantitative fit. This is however a longstanding question in linguistics and cognitive science. Thanks to identifying the three main components in generalization, we have been able to propose concrete models of the first two steps, and already explain much of the pattern of results.

Acknowledgments

This work was developed with Remko Scha, who sadly passed away before the finalization of this paper. We thank Carel ten Cate, Clara Levelt, Andreea Geambasu and Michelle Spierings for their feedback. We are also grateful to Raquel Fernández, Stella Frank and Miloš Stanojević for their comments on the paper. This research was funded by NWO (360-70-450).

References

Alhama, R. G., Scha, R., & Zuidema, W. (2016). Memorization of sequence-segments by humans and non-human animals: the retention-recognition model. *ILLC Prepublications*, *PP-2016-08*.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324.

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

Frost, R. L., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, *147*, 70–74.

Gale, W. A., & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, *2*(3), 217–237.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In *Proceedings of the annual meeting of the association for computational linguistics* (Vol. 44, pp. 673–680).

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*(3–4), 237–264.

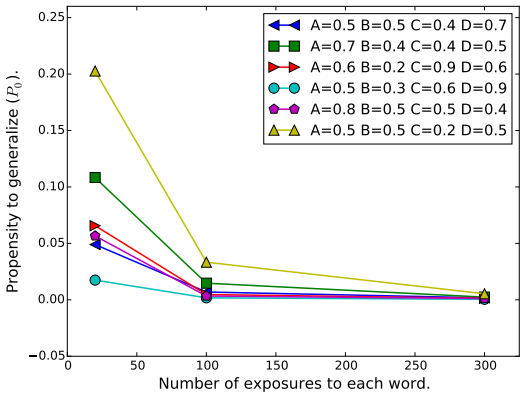
Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53–B64.

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

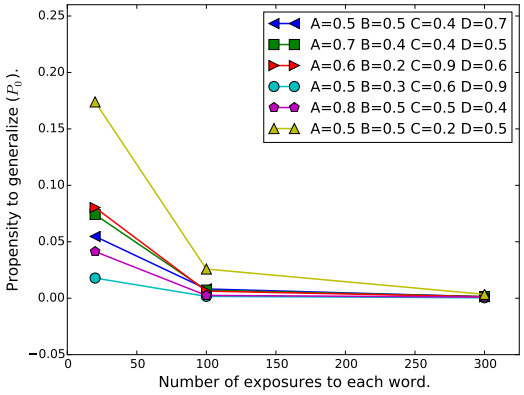
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621.

Toro, J. M., & Trobalón, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception & Psychophysics*, *67*(5), 867–875.



(a) Exposure without pauses.



(b) Exposure with pauses.

Figure 6: Propensity to generalize, for several parameter settings (average of 100 runs). Our model shows a clear decrease for all parameter settings we tried, consistent with the empirical data (compare with figure 1).

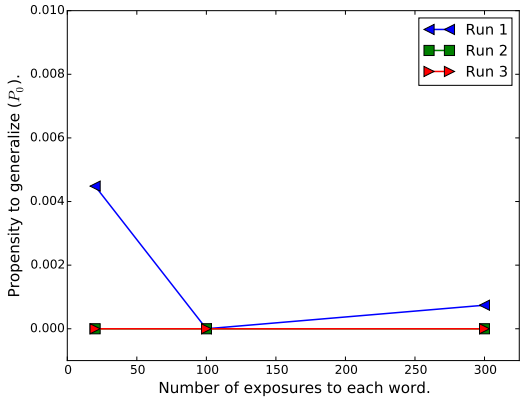


Figure 7: Propensity to generalize for the frequencies computed with the model described in Goldwater et al. (2006).