Generalization in Artificial Language Learning: Modelling the Propensity to Generalize

Alhama, R.G.; Zuidema, W.

DOI
10.18653/v1/W16-19

Publication date
2016

Document Version
Final published version

Published in
The 54th Annual Meeting of the Association for Computational Linguistics: proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning

Generalization in Artificial Language Learning: Modelling the Propensity to Generalize

Raquel G. Alhama, Willem Zuidema
Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands
{rgalhama, w.h.zuidema}@uva.nl

Abstract

Experiments in Artificial Language Learning have revealed much about the cognitive mechanisms underlying sequence and language learning in human adults, in infants and in non-human animals. This paper focuses on their ability to generalize to novel grammatical instances (i.e., instances consistent with a familiarization pattern). Notably, the propensity to generalize 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 statistical models (Peña et al. (2002); Endress and Bonatti (2007)). In this paper, we propose to model generalization as a three-step process, and we demonstrate that the use of statistical models for the first two steps, contrary to widespread intuitions in the ALL-field, can explain the observed decrease of the propensity to generalize with exposure time.

1 Introduction

In the last twenty years, experiments in Artificial Language Learning (ALL) 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 of the input, such as the transitional probabilities between adjacent syllables (Saffran et al. (1996a); Aslin et al. (1998)). This ability also seems to be present in human adults (Saffran et al., 1996b), and to some extent in nonhuman animals like cotton-top tamarins (Hauser et al., 2001) and rats (Toro and Trobalón, 2005).

Even though this statistical mechanism is well attested for segmentation, it has been claimed that it does not suffice for generalization to novel stimuli or *rule learning*[1]. Ignited by a study by Marcus et al. (1999), which postulated the existence of an additional *rule-based* mechanism for generalization, a vigorous debate emerged around the question of whether the evidence from ALL-experiments supports the existence of a specialized mechanism for generalization (Peña et al. (2002); Onnis et al. (2005); Endress&Bonatti (2007); Frost&Monaghan (2016); Endress&Bonatti (2016)), echoing earlier debates about the supposed dichotomy between rules and statistics (Chomsky, 1957; Rumelhart and McClelland, 1986; Pinker and Prince, 1988; Pereira, 2000).

From a Natural Language Processing perspective, the dichotomy between rules and statistics is unhelpful. In this paper, we therefore propose a different conceptualization of the steps involved in generalization in ALL. In the following sections, we will first review some of the experimental data that has been interpreted as evidence for an additional generalization mechanism (Peña et al. (2002); Endress&Bonatti (2007); Frost&Monaghan (2016)). We then reframe the interpretation of those results with our 3-step approach, a proposal of the main steps that are required for generalization, 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. We model the first step with the *Retention&Recognition* model (Alhama et al., 2016). We propose that a rational charac-

[1]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 the generalization process have the form of algebraic rules.

Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning, pages 64–72,
Berlin, Germany, August 11, 2016.

terization of the second step can be accomplished with the use of *smoothing* techniques (which we further demonstrate with the use of the Simple Good-Turing method, (Good&Turing (1953); Gale (1995)). We then argue that the modelling results shown in these two steps already account for the key aspects of the experimental data; and importantly, it removes the need to postulate an additional, separate generalization mechanism.

2 Experimental Record

Peña et al. (2002) conduct a series of Artificial Language Learning 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 A_iXC_i such that the A_i syllable always co-occurs with the C_i syllable (as indicated by the subindex i). This forms a consistent pattern (a "rule") consisting in a non-adjacent dependency between A_i and C_i, with a middle syllable X that varies. The order of the words in the stream is randomized, with the constraint that words do not appear consecutively if they either: (i) belong to the same "family" (i.e., they have the same A_i and C_i syllables), or (ii) they have the same middle syllable X.

stream	pulikiberagatafodupuraki..
words A_iXC_i	puliki, beraga, tafodu, ...
part-words C_jA_iX, XC_iA_j	kibera, ragata, gatafo, ...
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.

After the familiarization phase, the participants respond a two-alternative forced choice test. The two-alternatives involve a word vs. a *part-word*, or a word vs. a *rule-word*, and the participants are asked to judge which item seemed to them more like a word of the imaginary language they had listened to. A part-word is an ill-segmented sequence of the form XC_iA_j or C_iA_jX; a choice for a part-word over a word is assumed to indicate that the word was not correctly extracted from the stream. A rule-word is a rule-obeying sequence that involves a "novel" middle syllable Y (meaning that Y did not appear in the stream as an X, although it did appear as an A or C). Rule-words are therefore a particular generalization from words. Table 1 shows examples of these type of test items.

In their baseline experiment, the authors expose the participants to a 10 minute stream of A_iXC_i 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. In a second experiment the same setup is used, 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, in a third experiment, 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

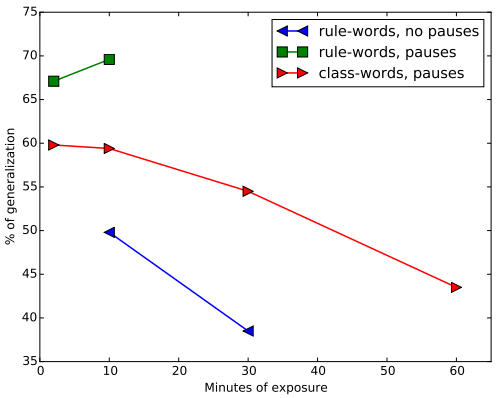


Figure 1: Percentage of choices for rule-words and class-words, in the experiments reported in Peña et al. (2002) and Endress&Bonatti (2007), for different exposure times to the familiarization stream.

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 A_iZC_i, where A_i and C_i 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 ZC_iA_j or C_iA_jZ.

The pattern of results is complex, but we can identify the following key findings: (i) generalization for a stream without pauses is only manifested 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 phenomenon, which we call *time effect*, is precisely the aspect we want to explain with our model. (Note, in figure 1, that in the case of rule-words and pauses, the amount of generalization increases a tiny bit with exposure time, contrary to the time effect. We cannot test whether this is a significant difference, since we do not have access to the data. Endress&Bonatti, however, provided convincing statistical analysis supporting a significant inverse correlation between exposure time and generalization to class-words).

3 Understanding the generalization mechanism: a 3-step approach

Peña et al. interpret their findings as support for the theory that there are at least two mechanisms, which get activated in the human brain based on different cues in the input. Endress and Bonatti adopt that conclusion (and name it the *More-than-One-Mechanism* hypothesis, or *MoM*), and moreover claim that this additional mechanism cannot be based on statistical computations. The authors predict that statistical learning would benefit from increasing the amount of exposure:

> *"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).*" (Endress and Bonatti, 2007)

We think this argument is based on a wrong premise: stronger representations do not necessarily entail greater generalization. On the contrary, we argue that even very basic models of statistical smoothing make the opposite prediction. To demonstrate this in a model that can be compared to empirical data, we propose to think about the process of generalization in ALL as involving the following steps (illustrated also in figure 2):

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

Crucially, we believe that step (ii) has been neglected in ALL models of generalization. This step accounts for the fact that generalization is

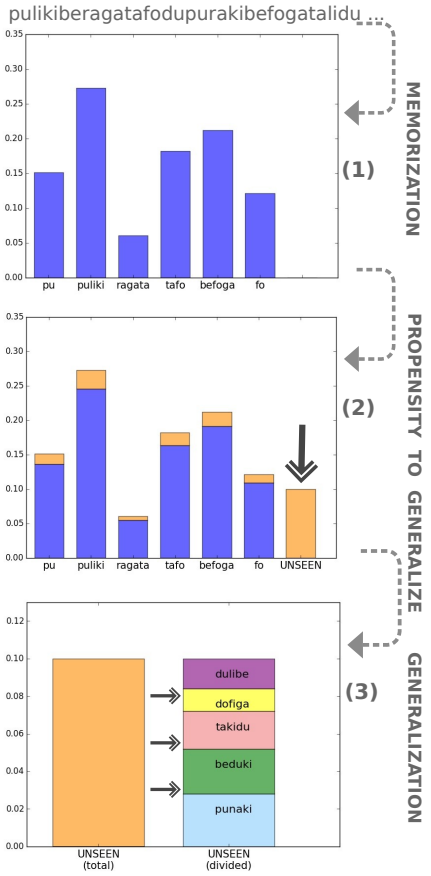


Figure 2: Three step approach to generalization: (1) memorization of segments, (2) compute probability of new items, and (3) distribute probability between possible new items.

not only based on the particular structure underlying the stimuli, but also depends on the statistical 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 the impact of exposure time on generalization depends on the model used for step (ii). Next, we show that a cognitive model of step (i) and a rational statistical model of step (ii) already account for the *time effect*.

4 Memorization of segments: the *Retention and Recognition* model

For step (i) of our approach, several existing models maybe used, including models based on recurrent neural networks (Seidenberg and Elman, 1999), autoencoders (French et al., 2011; French and Cottrell, 2014), exemplar-based processing (Perruchet and Vinter, 1998) and non-parametric Bayesian inference (Goldwater et al., 2006). We have decided to implement the Retention&Recognition (R&R) model, proposed in (Alhama et al., 2016). R&R is a probabilistic exemplar-based model that has been shown to fit experimental data from a range of ALL experiments on segmentation, and, importantly, produces very skewed frequency distributions that fit well with our intuition about step (ii).

Starting from an initially empty memory, R&R processes subsequences (segments) of the speech stream, and decides probabilistically whether those segments will be stored in its internal memory. The output of the model is a memory of segments, each one with a count of how many times the model has decided to store it in memory. The authors refer to these counts as *subjective frequencies*.

In each iteration, R&R is presented with one segment from the input stream. Each segment may be composed of any number of syllables (until an arbitrarily set maximum). For instance, for a stream starting with *talidupuraki...*, the model would be presented, in order, with the segments *ta, tali, talidu, talidupu, li, lidu, lidupu, lidupura*, etc. (assuming a maximum length of four syllables).

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). 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. 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 segment is processed.

The recognition probability $p_1$for segment s is defined as follows (eq. 1):

$$p_1(s) = (1 - B^{activation(s)}) \cdot D^{\#types} \qquad (1)$$

$$0 \leqslant B, D \leqslant 1$$

where B and D are parameters to be set with the empirical data. The recognition probability depends on the $activation$ of the segment, which equals the subjective frequency. As it can be deduced from eq. 1, segments with greater subjec-

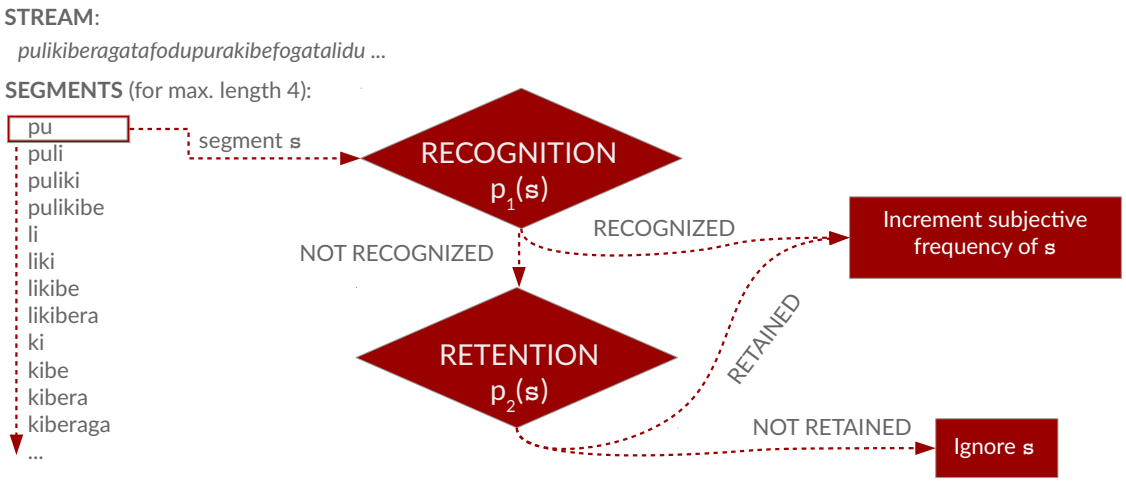


Figure 3: The Retention&Recognition model. Diagram based on Alhama et al. (2016).

tive frequency are easier to recognize. However, the number of different segment types in memory ($\#types$) makes the recognition task more difficult.

The retention probability p_2 is defined in eq. 2:

$$p_2(s) = A^{length(s)} \cdot C^{\pi} \qquad (2)$$

$$0 \leqslant A, C \leqslant 1; \quad \pi = \begin{cases} 0 \text{ after a pause} \\ 1 \text{ otherwise} \end{cases}$$

A and C are parameters to be set with empirical data, and π takes the value 0 when the segment being processed occurs right after a pause, and 1 otherwise. The retention probability is greater for shorter segments (as can be deduced from the *length(s)* exponent applied to an A parameter that ranges between 0 and 1). The C parameter, which is again between 0 and 1, attenuates this probability unless a pause precedes the segment. This has the effect of boosting the retention of segments that appear after a pause.

The four parameters involved in the model (A, B, C, D) set the contribution of each of its components, and allow for the adaptation of the model to different tasks or species. Alhama et al. did not report the optimal parameter setting for the experiments we are concerned with here, but they assert that the main qualitative features of the model (such as the *rich-get-richer* dynamics of the recognition function) are independent of the parameters.

Among these qualitative features, one that is particularly relevant for our study is the *skew* that can be observed in the subjective frequencies computed by the model. This feature, which can be observed in figure 4, is presented in the original paper as being in consonance with empirical data. Here, we show that this property can also be validated in a different way: when R&R is part of a pipeline of models (like the 3-step approach), the skew turns out to be a necessary property for the success of the next model in the sequence. We come back to this point in section 7.

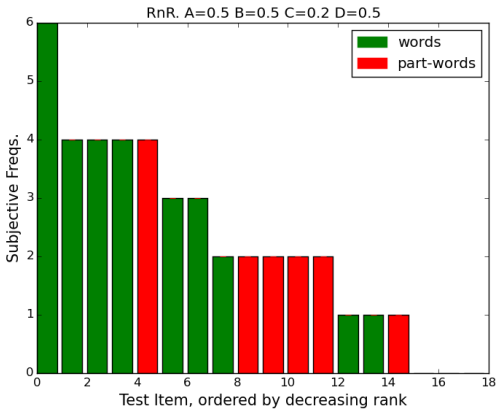


Figure 4: Subjective frequencies computed by the R&R model (A=0.5, B=0.5, C=0.2, D=0.5), for an exposure of 10 minutes (without pauses) to the stimuli used by Peña et al.

5 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. This is a well known problem in Natural Language Processing, and the techniques to deal with it are known as *smoothing*. Here, we explore the use of the Simple Good Turing (Gale and Sampson, 1995) smoothing method as a computational level characterization of the propensity to generalize.

Simple Good-Turing (SGT), a computationally efficient implementation of the Good-Turing method (Good, 1953), is a technique to estimate the frequency of unseen types, based on the frequency of already observed types. The method works as follows: we take 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 sequences that have a certain subjective frequency r. The values N_r are then *smoothed*, that is re-estimated with a continuous downward-sloping line in log space. The smoothed values $S(N_r)$ are used to reestimate the frequencies according to (3):

$$r^* = (r+1)\frac{S(N_{r+1})}{S(N_r)} \tag{3}$$

The probabilities for frequency classes are then computed based on these reestimated frequencies:

$$p_r = \frac{r^*}{N} \tag{4}$$

where N is the total of the unnormalized estimates[2] .

Finally, the probability for unseen events is computed based on the (estimated) [3] probability of types of frequency one, with the following equation:

$$P_0 = \frac{S(N_1)}{N} \tag{5}$$

This probability P_0 corresponds to what we have called "propensity to generalize".

[2]It should be noticed that the reestimated probabilities need to be renormalized to sum up to 1, by multiplying with the estimated total probability of seen types $1 - P_0$ and dividing by the sum of unnormalized probabilites.

[3]SGT incorporates a rule for switching between N_r and $S(N_r)$ such that smoothed values $S(N_r)$ are only used when they yield significantly different results from N_r (when the difference is greater than 1.96 times the standard deviation).

As can be deduced from the equations, SGT 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.

6 Results

6.1 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 test items (the nine words and nine possible 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.

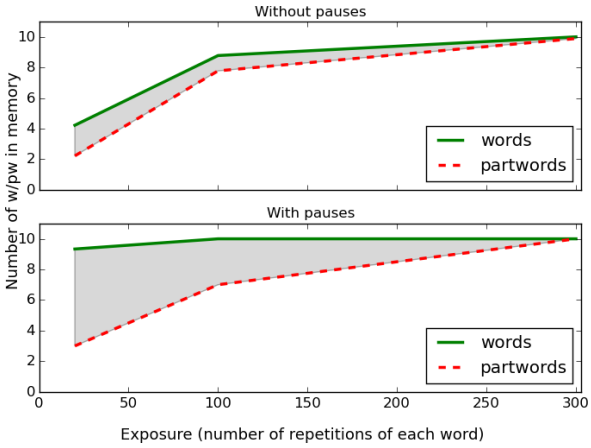


Figure 5: Average number of memorized words and part-words after familiarization with the stimuli in Peña et al., for 10 runs of the R&R model with an arbitrary parameter setting (A=0.5 B=0.5 C=0.2 D=0.5).

The graph also shows that, when the micropauses are present, words are readily identified after much less exposure, yielding clearer differences 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?

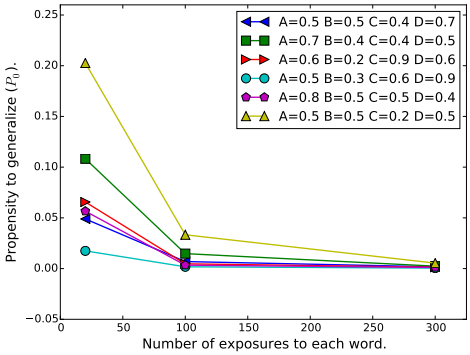
6.2 Prediction of observed decrease in the propensity to generalize

Next, we apply the Simple Good-Turing method [4] to subjective frequencies computed by the R&R model. As shown in figure 6, we find that the propensity to generalize (P_0) 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.

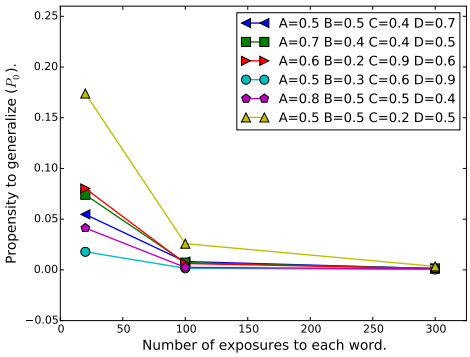
The results of these simulations point to a straightforward explanation of the experimental finding of a reduced preference for the generalized sequences: longer exposures repeat the same set of words (and partwords), and consequently, participants may conclude that there are no other sequences in that 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 slightly smaller for the micropause condition. The reason for that is that R&R identifies words faster when micropauses are present, and therefore, the subjective frequencies tend to be greater. This might appear unexpected, but it is in fact not contradicting the empirical results: as shown in figure 5, the difference between words and partwords is much bigger in the condition with micropauses, so this effect is likely to override the small probability difference (as would be confirmed by a model of step (iii)). It should be noted that, as reported in Frost&Monaghan (2016), micropauses are not essential for all type of generalizations (as is evidenced with the fact that rule*-words are generalized in the no-pause condition). Like those authors, we see as the role of the micropauses to enhance the salience of initial and final syllables (A and C) to compensate for the odd construction of the test items (rule-words and class-words), which include a middle syllable that occupied a different position in the familiarization stream.

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



(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).

7 Discussion

The experiments we have focused on 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 distribu-

tion of subjective frequencies that is crucial for step (ii); 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 in discussion of ALL, 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 about the probability of next observing any unattested item, and the responses for generalized sequences must be affected by it. Moreover, this step is based on statistical information, proving that — contrary to the MoM hypothesis — a statistical mechanism can account for the negative correlation with exposure time.

It should be noted that our conclusion concerns the qualitative nature of the learning mechanism that is responsible for the experimental findings. It has been postulated that such findings evidence the presence of *multiple* mechanisms (Endress and Bonatti, 2016). In our view, the notion of 'mechanism' is only meaningful as a high-level construct that may help researchers in narrowing down the scope of the computations that are being studied, among all the computations that take place in the brain at a given time. After all, there is no natural obvious way to isolate the computations that would constitute a single 'mechanism', from an implementational point of view. Therefore, our 3-step approach should be taken as sketching the aspects that any model of generalization should account for, and our modelling efforts show that the experimental results are expected given the statistical properties of the input.

One issue to discuss is the influence of the use of the R&R model in computing the propensity to generalize. The Simple Good-Turing method is designed to exploit the fact that words in natural language follow a Zipfian distribution —that is, languages consist of a few highly frequent words and a long tail of unfrequent words. This is a key property of natural language that is normally violated in ALL experiments, since most of the artificial languages used are based on a uniform distribution of words (but see Kurumada et al. 2013). But it would be implausible to assume that subjects extract the exact distribution for an unknown artificial language to which they have been only briefly exposed. R&R models the transition from absolute to subjective frequencies, resulting in a distribution of subjective frequencies that shows a great degree of skew, and much more so than alternative models of segmentation in ALL. Thanks to this fact, the frequency distribution over which the SGT method operates (the subjective distribution) is more similar to that of natural language, and the pattern of results found for the propensity to generalize crucially depends on this type of distribution.

Finally, we have thus accomplished our goal qualitatively. We capture the downward tendency of the propensity to generalize, but a model for step (iii), a longstanding question in linguistics and cognitive science, is required to also achieve a quantitative fit. Developing a model of step (iii) is left as future work, but our approach already allowed us to propose concrete models of the first two steps, and 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

Raquel G. Alhama, Remko Scha, and Willem Zuidema. 2016. Memorization of sequence-segments by humans and non-human animals: the retention-recognition model. *ILLC Prepublications*, PP-2016-08.

Richard N Aslin, Jenny R Saffran, and Elissa L Newport. 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9(4):321–324.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

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

A.D. Endress and L.L. Bonatti. 2016. Words, rules, and mechanisms of language acquisition. *WIREs Cognitive Science*. (in press).

Robert M French and Garrison W Cottrell. 2014. Tracx 2.0: A memory-based, biologically-plausible model of sequence segmentation and chunk extraction. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Robert M. French, Caspar Addyman, and Denis Mareschal. 2011. Tracx: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, 118(4):614.

Rebecca LA Frost and Padraic Monaghan. 2016. Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, 147:70–74.

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

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the annual meeting of the association for computational linguistics*, volume 44, pages 673–680.

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

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

G.F. Marcus, S. Vijayan, S.B. Rao, and P.M. Vishton. 1999. Rule learning by seven-month-old infants. *Science*, 283(5398):77–80.

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

Fernando Pereira. 2000. Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society*, 358(1769):1239–1253.

Pierre Perruchet and Annie Vinter. 1998. Parser: A model for word segmentation. *Journal of Memory and Language*, 39(2):246–263.

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

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193.

D.E. Rumelhart and J.L. McClelland. 1986. On learning past tenses of English verbs. In D.E. Rumelhart and J.L. McClelland, editors, *Parallel Distributed Processing, Vol. 2*, pages 318–362. MIT Press, Cambridge, MA.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996a. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

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

Mark S Seidenberg and Jeffrey L Elman. 1999. Networks are not 'hidden rules'. *Trends in Cognitive Sciences*, 3(8):288–289.

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