Memorization of sequence-segments by humans and non-human animals: the Retention-Recognition Model

Raquel G. Alhama (rgalhama@uva.nl)
Remko Scha (scha@uva.nl)
Willem Zuidema (zuidema@uva.nl)
Institute for Logic, Language and Computation, Science Park 107
Amsterdam, 1098XG, The Netherlands

version July, 2015

Abstract

The existence of a statistical learning mechanism underlying the cognitive capacity to learn a "language" from an auditory input stream is well established. Computational models of segmentation formalize and test the specific theoretical assumptions about the mechanistic nature of this process. We present the Retention and Recognition model (R&R), a probabilistic model based on the cognitive processes of retention and recognition. We show that R&R outperforms other models in explaining for a range of experimental results: for a 2AFC task with human adults reported in Frank, Goldwater, Griffiths, and Tenenbaum (2010), as well as with data of human adults from a variant experiment from Peña, Bonatti, Nespor, and Mehler (2002), and with the responses of a segmentation experiment with rats (Toro & Trobalón, 2005). Our model also offers a new prediction on the response distribution over test items, which we will confirm revisiting these experimental results.

Keywords: artificial language learning; rule learning; statistical learning; animal cognition; cognitive modelling

1 Introduction

In artificial language learning (ALL) experiments, infants, human adults and nonhuman animals are exposed to samples from an artificial language, and tested on their ability to discover the pattern that characterizes that language. Over the last two decades, ALL has come to play a key role in many debates about the nature of the cognitive mechanisms underlying language and music, and questions about whether these mechanisms are unique to humans, language and/or music. The discovery that young infants are sensitive to transitional probabilities in speech streams (Saffran, Newport, & Aslin, 1996), which likely plays a role in discovering the words of their native language, has led to wide-spread acceptance of a 'statistical learning' mechanism. However, many researchers have argued that there is a second, distinct 'rule learning' mechanism at work when subjects process a stream containing patterns defined by 'algebraic rules' (e.g., Marcus, Vijayan, Rao, and Vishton (1999)).

In a cleverly designed experiment, Peña, Bonatti, Nespor and Mehler (2002) obtain results that, in their interpretation, support the existence of (at least) these two distinct mechanisms and show that they can be triggered by subtle cues in the input. The Peña et al. work was subject of a vigorous debate that focused on possible confounds and the question whether the data really demonstrates the existence of a separate rule learning mechanism (Onnis, Monaghan, Richmond, and Chater (2005); Endress and Bonatti (2006); Perruchet, Tyler, Galland, and Peereman (2004), etc.).

In our opinion, the debate about whether 'rule learning' and 'statistical learning' are best described as separate mechanisms with an overarching control structure that selects between them, or as different processing modes of a single mechanism, is premature, since at this moment we do not yet understand the most fundamental aspects of the process of segmentation. We therefore focus on what we feel is the most urgent issue: identifying which units of an unsegmented speech stream are extracted and how can we quantify the strength of their memory trace. We argue that a real understanding of the process of segmentation that explains the pattern of results reported in many papers on the topic requires a precise, integrated computational model that makes correct predictions on the behav-

ior of subjects for a variety of artificial languages. Such a model must, in our view, be pitched at what Marr (1982) has termed the *processing level*, because a model at the *rational level* could not address the differences that were observed between different species; and because we do not know enough yet to venture into modelling at the *neural level*. A model with these characteristics has, to the best of our knowledge, not yet been proposed.

Before presenting such processing model, we first summarize the experimental record. We test our model — the *Retention-Recognition Model* (R&R) — against results from a variety of experiments: two conditions from the Toro and Trobalón (2005) studies with rats, a variant of the baseline experiment from the Peña et al. (2002) studies, and the three internet-based experiments with human adults reported in (Frank et al., 2010). In the next section, we will first describe the main experimental findings; the following sections then define the basic model and evaluate it against empirical data. In section 5 we discuss related work, including models proposed for some of the same datasets and a model (PARSER, by Perruchet and Vinter (1998)) that has some similarities with ours but also important differences.

2 Experimental Record

Peña et al. (2002) report results from an Artificial Language Learning experiment in which French-speaking adults are exposed to a stream of nonsense words, and then tested to ascertain whether they have detected the rules underlying the internal structure of the words. The "words" in these experiments are syllable triples of the form AXC, where A and C reliably predict each other while for a given AXC pattern, X is instantiated in 3 different ways. For instance, 'puliki' and 'puraki' and 'pufoki' constitute one "family" of words. In the familiarization phase, subjects heard a stream of words constructed by randomly picking words from 3 such families. In some of the experiments, subliminal pauses were inserted between subsequent words in the stream.

In the test phase of the experiments, subjects were tested on whether they recognized these words, but also on the recognition of *partwords* (triples that occurred in the speech stream but that cross word-boundaries, thus having the structure C_nA_mX or XC_nA_m), and the "recognition" of *rulewords* (triples AYC that conform to an attested A_C pattern, but with a middle syllable Y that did not occur in this position in the stream). Some examples of the stream, words, partwords and rulewords are given in table 1.

In the original paper, all tests involve a forced choice task, where subjects are presented with pairs of triples and are asked which of the two was more likely to be part of the artificial language they heard in the familiarization phase. Tested after 10 minutes of exposure, the subjects show a significant preference for words over partwords, but they have no preference when they compare rulewords and partwords. If the exposure time is increased to 30 minutes, they prefer partwords to rulewords.In a third experiment, micropauses of 25 ms are added between words; now, only 2 minutes of exposure results in a preference for rulewords.

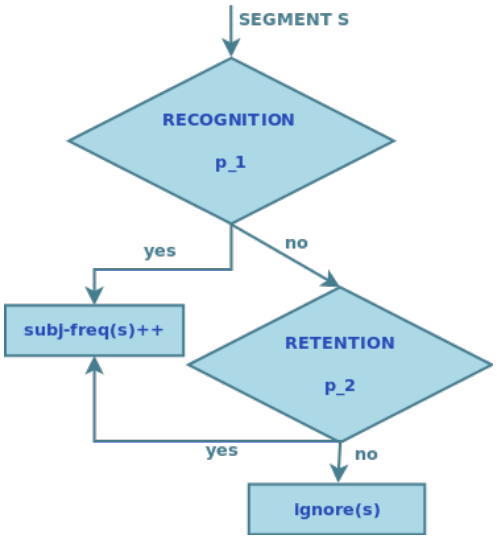
We present in this paper a variant of the baseline experiment (10 minutes exposure, and a test involving words and partwords), in which we substitute the forced choice task with an alternative test. In this set up, participants will have to answer a 'yes/no' question about a sequence being a word of the artificial language; each of this questions will be presented together with a confidence rate about the answer. As explained in section 4.2, this alternative type of test reveals interesting properties in the responses per test item.

stream	pulikiberagatafodupurakibefogatalidu ...
words (AXC)	puliki, beraga, tafodu, puraki, befoga, ...
part-words (CAX, XCA)	kibera, ragata, gatafo, fodupu, dupura, ...
rule-words (AYC)	pubeki, beduga, takidu, ...

Table 1: Summary of the stimuli used in the experiments by Peña et al. (2002)

Toro and Trobalón (2005, Experiment 3A) report similar experiments with rats. The animals are exposed to a 20 minute speech stream (with or without pauses) made out of the same triples as used by Peña et al. (2002). Although the rats could segment a simpler speech stream on the basis of co-occurrence frequencies, after the Peña et al. stream (without micropauses) their response rates do not differentiate between words and partwords; only with the insertion of micropauses they show a higher response rate for words. With or without micropauses, the reponses to rulewords are not significantly different from the responses to partwords. Toro and Trobalón interpret this as evidence for lack of generalization —rats do generalize, but less readily than humans. But since partwords were actually present in the familiarization stream and rulewords weren't, the data is consistent with a model that assumes degrees of generalization. Unfortunately, the control experiment with non-attested non-rule-obeying syllable combinations was not performed.

Frank et al. (2010) present an extensive study of seg-



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

Figure 1: R&R:The Retention-Recognition Model

mentation in human adults. They focus on how different properties of the stimuli can influence the performance of the participants. They vary the number of words in the sentences that compose the stimuli, as well as the total number of different words in the language and the amount of repetitions of each word. These results confirm that the length of a sentence and the number of words increase the difficulty of the task, while the amount of repetitions boosts the performance of the subjects.

3 R&R: the Retention-Recognition Model

3.1 Model description

In designing our model we assume that any process of learning and generalization, whatever form it may take, must operate on items extracted from the input stream which are committed to memory. The process of Artificial Language Learning as viewed by Peña et al. and its follow-up studies may thus be divided into two steps: (1) segmentation of the input stream and memorization of segments, and (2) generalizing the store of memorized fragments. In the present paper we propose a detailed, quantitative model of the first of these steps, allowing us to account for existing experimental results concerning segmentation and memorization. We present a simple processing model, R&R, that describes the memorization process during the familiarization phase of the experiments.

The R&R model is represented in figure 1. It operates on segments (subsequences of the stream, with no a priori constraints on the length of the subsequence considered), and maintains a memory of such segments with an associated (subjective) count. Given an (initially empty) memory, and any segment from the input stream, the model may (with probability p_1) *recognize* it (i.e., match it with a segment in memory). If it succeeds, the subjective frequency of the segment is incremented with 1. If it fails to recognize the segment, the model might (with probability p_2) still *retain* it (i.e., add it to memory with initial subjective frequency 1 if it was not stored, or increase the subjective frequency by 1 as a form of 'late recognition'). In this way, the model builds a memory of segments that have different degrees of familiarity depending on their distribution in the stream.

The model involves free parameters (A,B,C,D) that may be fitted to empirical data. The B parameter describes how the recognition probability *increases* with the subjective count ('activation') of the segment in memory. The D parameter describes the *decrease* in recognition probability with the number of different items in memory. Note that the current model equates the activation of a segment with its subjective frequency, but the term "activation" suggests future extensions of the model, for instance to take recency effects into account. Note further that to account quantitatively for the recognition of rule-words in the experiments of Peña et al. (2002), the model should be extended with a capacity to assign non-zero activations to unseen triples which share statistically significant properties with the stored exemplars, but that is beyond the scope of this paper.

The A parameter describes how quickly the retention probability decreases with the length of segment. The factor C^{π} attenuates this probability unless the segment appears right after a micropause. In this way we model the increased likelihood of a perceptual boundary after a micropause, which makes pause-delimited segments more salient.

3.2 Qualitative behaviour of the model

R&R exhibits *rich-get-richer* dynamics: as the subjective frequency of a sequence grows, the probability for this

sequence to be recognized on its next occurrence in the stream also grows, and therefore its subjective frequency is likely to increase again. A sequence, however, cannot be recognized before it has been retained. The stochasticity of the retention will cause some sequences to be retained later than others, so not all sequences will benefit equally from a high recognition probability.

With this interplay between the stochasticity of the retention and the (also stochastic) rich-get-richer dynamics of the recognition, even sequences that are identical in terms of absolute frequency may end up with substantially different subjective frequency. This effect can be observed in figure 2, which shows the subjective frequency of the model for the baseline experiment in Peña et al. (2002). As can be seen, the model presents different behaviour under different parameters, but all of them produce a somewhat skewed distribution of subjective frequencies.

The parameters of the model can regulate the degree of skew (in general, low probabilities yield greater skew) and other aspects of the behaviour of the model: as shown in figure 2, R&R can yield distributions of subjective frequencies in which the words of a stream are clearly distinguished from the partwords, as well as others in which some of the partwords will have higher subjective frequencies than some of the words.

The qualitative behavior of R&R predicts therefore that the responses of subjects in the segmentation experiments will reflect the skewed distribution of the memorized sequences. The next section analyzes whether this predictions is found in the empirical results.

4 Empirical predictions and empirical data

4.1 Prediction of observed skew in response distribution in rats

We have implemented the R&R model in Python and studied its behavior under a variety of parameter settings. We exposed the model to a stream of syllables that we created by following the description of the familiarization stream in Peña et al. (2002). In that stream, *words* appear all with exactly the same frequency (e.g. *puliki*, *beraga*, *tafodu*, etc. appear 100 times each in the 10 minute condition). *Partwords* have a much lower frequency (approximately $1/2$ of the word frequency)[1].

[1]The exact frequency depends on the random process by which *words* are sampled; in our simulations we have assumed that Peña et al. repeatedly play the complete sequence of words in randomized order. We also tried other processes consistent with the description they give, and obtained very similar results

In line with the general observations made above, R&R generates skewed distributions when presented with this familiarization stream (see figure 2). Such a skew has, to the best of our knowledge, not been reported yet in the analysis of experimental ALL results on adults, which all report averages over responses in a forced choice setting. Nor has it been reported in papers on experiments with prelinguistic infants and animals, which do measure reponses to individual test items but all *report* averages over stimulus classes.

To investigate whether the experimental results reflect our prediction of skew, we need another type of analysis. We first apply such an analysis to the original data of (Toro & Trobalón, 2005), which the authors kindly shared. The authors measure leverpressing responses of the rats when exposed to the test triples to assess their recognition of words, partwords and rulewords. In figures 3 and 4 we plot, with small solid circles, the responses of rats after familiarization to the stream without and with pauses respectively. We again plot the data for partwords and words ordered by response frequency.

Interestingly, thanks to plotting the individual responses per item, we can now observe that the responses clearly follow a very skewed distribution. This new observation of these experimental results confirm the prediction that we derived from the qualitative behavior of our model: that the responses of subjects for the test items are not uniform within the class of stimuli.

To evaluate how well the model can fit these data quantitatively, we make the additional assumption that the measured response rates are directly proportional to the subjective frequencies of the triples in the memories of the rats and then search for parameter settings that produce the best fit (measured with squared error) to the 'median' rat[2] (shown as blue lines in the graphs). We fit parameters B and D, and a value[3] that combines A and C, to the data without pauses; we then used the data with pauses solely to differentiate between the contributions of A and C. The pink lines in the graphs give the prediction of the model with the thus fitted parameters, and demonstrate a surprisingly good fit.

We also plot the average responses of the rats and average subjective frequencies of the model in figures 5 and 6. In the experiments without pauses (figure 5), the rats seem, counterintuitively and in contrast to the model's

[2]Defined as the rat with the median lever pressing response to the words and partwords with the highest and second highest response rates.

[3]As all considered segments have length 3 and there is no information to differentiate between the contributions of A and C, we estimate the value of A^3C instead. We then assume these values as given, and employ the corresponding data from the experiment with micropauses to estimate A and C.

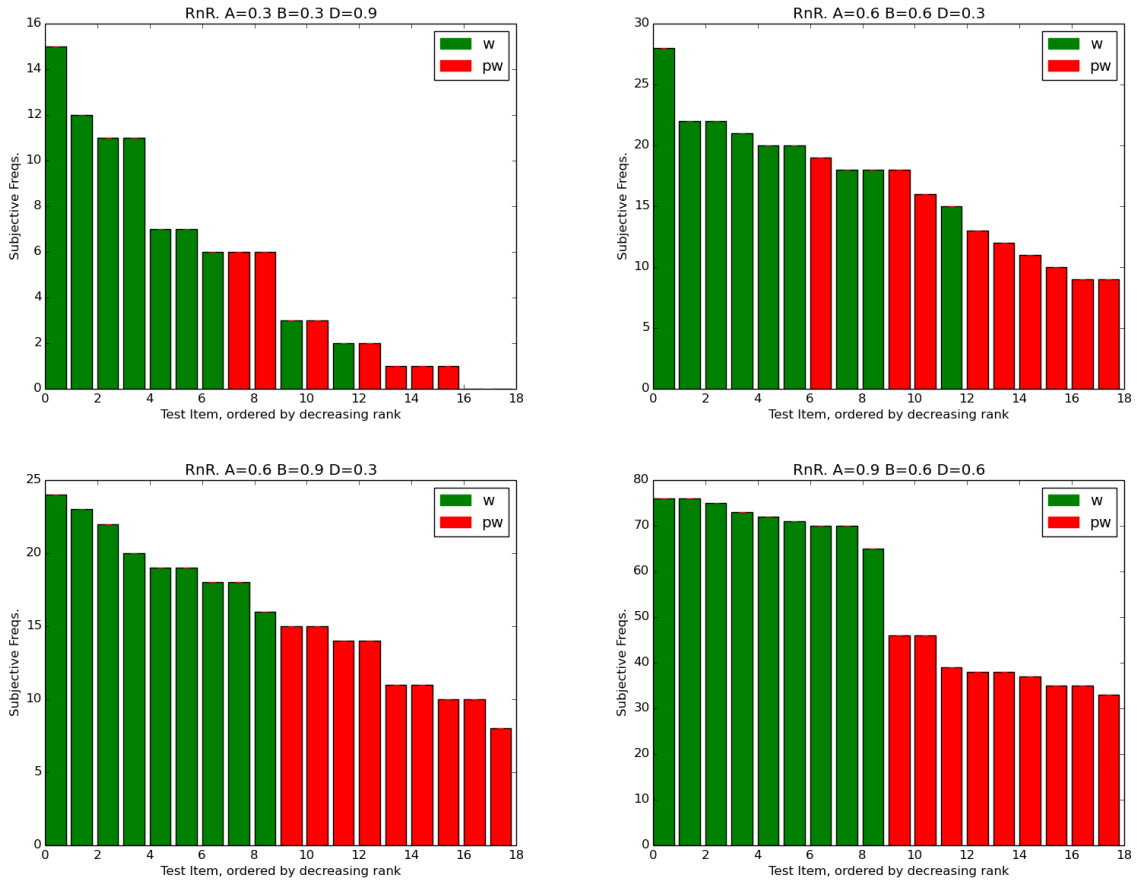


Figure 2: Subjective Frequencies in 4 runs of the R&R model, with different parameters, when familiarized with the AXC language of Peña et al. (2002).

prediction, to prefer partwords over words. The difference is not statistically significant, however (t(7)=-1.47, p=.184, as reported by (Toro & Trobalón, 2005)).

4.2 Prediction of observed skew in response distribution in humans

We have been able to confirm the prediction of skew in the response distribution of rats because we obtained access to the original data, which consisted of responses per item. But when it comes to humans, we encounter some complications: adults are typically tested in 2AFC tasks, which do not allow for a study of the distribution of responses for single items; as for infants, although the type of responses that are recorded (typically, listening times) would allow to investigate the preference for single items, the reported data consists only on averages for classes of sequences, and we could not obtain access to the original data for any of the relevant studies, despite repeated requests.

For these reasons, we have run an experiment with human adults to investigate if the skew of response distributions is consistent with that predicted by R&R. For this, we used stimuli following the structure proposed in Peña et al. (2002).

Methods

Participants

13 participants, master students of the University of Amsterdam, participated in the study as part of one of their courses.

Stimuli

We presented an 11 minute speech stream of synthetic speech syllables generated with eSpeak. We used two conditions that only differed in the randomization of the position of a syllable in a word, and the randomization of the order of appearance of those words. For one group,

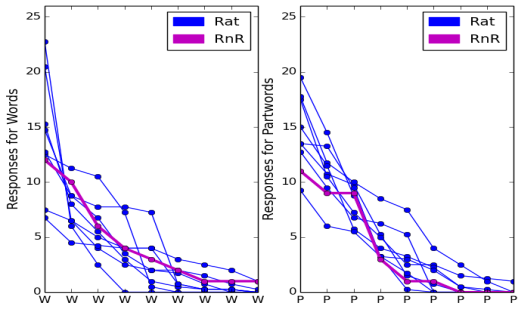


Figure 3: Responses of rats (blue) and subjective frequencies of the model (pink), without pauses. W indicates words; P indicates partwords; both ordered by response frequency. Parameter setting of the model: A=0.5; B=0.6; C=0.7; D=0.85.

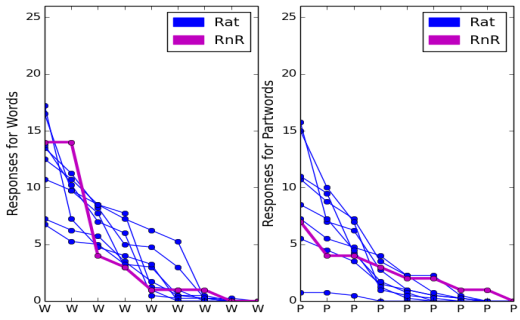


Figure 4: Responses of rats (blue) and subjective frequencies of the model (pink), with pauses. W indicates words; P indicates partwords; both ordered by response frequency. Parameter setting of the model: A=0.5; B=0.6; C=0.7; D=0.85.

the words were: *jaduki, jamaki, jataki, lidufo, limafo, litafo, sudube, sumabe, sutabe*; for the other, the words were: *jabeta, jaduta, jakita, mabefo, madufo, makifo, subeli, suduli, sukili*. Each word was presented 100 times, and their order of appearance was random with the constrain that one word cannot follow another of the same family (i.e., that starts and ends with the same syllable).

The test items consist of the nine words of the familiarization stream and nine partwords, also present in the familiarization stream, consisting of two syllables of one word and one syllable of the next, or of one syllable of one word and two syllables of the next. These eighteen items appear two times in the test set, and their order of appearance is randomized (but constant across participants), with the constraint that the same sequence

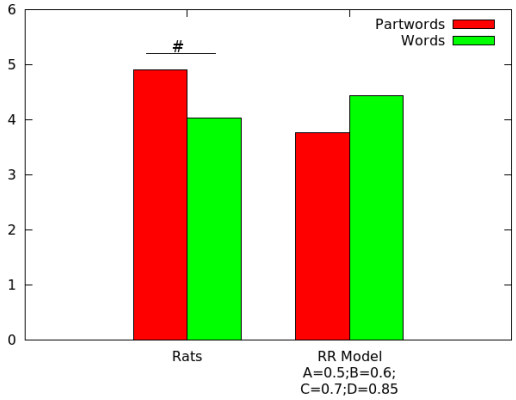


Figure 5: Average responses of rats and model, for experiment without pauses.

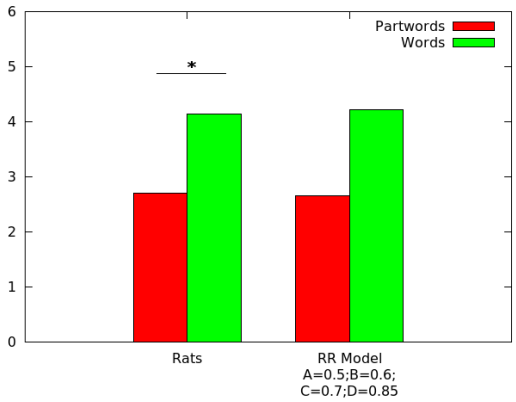


Figure 6: Average responses of rats and model, for experiment with pauses.

doesn't appear consecutively.

Procedure

The participants were randomly assigned to one of the two conditions. The stimuli were presented with the use of a web form. They were instructed to listen to the whole familiarization stream, for which they would have to answer questions afterwards. In the test phase, each test item was presented acoustically, followed with the question 'Is this sequence part of the language you have heard?', to be answered with yes/no. Afterwards, they were asked to rate their confidence in the previous answer, in a scale from 1 to 7 (where 1 is minimum confidence and 7 is maximum).

Results

The average accuracy of the participants is 59.25%. This number is below that of Peña et al. (2002) (73.3%); this difference may reflect the fact that test items are

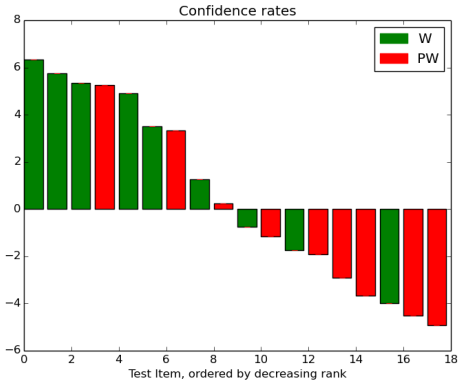


Figure 7: Confidence rates, averaged per ranked item.

presented in isolation, in contrast with the 2AFC task that Peña et al. (2002) used, where two items are presented at the same time and therefore the participant has more information (e.g., for a word that might have been accepted at chance level, the presence of its paired partword in the test can provide an extra hint for accepting the word). Nevertheless, the difference between words and partwords is significant (T-test over scale responses: t = 2.8722, df = 21.971, p-value = 0.008859).

We use the scale response of the confidence rate, multiplied with -1 if the answer to the yes/no question was negative. For each participant, we order their responses, maintaining the separation between words and partwords. Then we align the responses by their class (word or partword) and rank (position in the ordered list or responses for a particular class) and we average across participants. The assumption behind this procedure is that words are indistinguishable in terms of their frequency, but yet the most salient word for one participant need not be the most salient word for the other participant. In other words, we anonymize the particular item, while maintaining their confidence rate, rank and class.

The results are shown in figure 7, combining the two conditions. Thanks to analysing the responses per item, we can observe that, as we expected, the data of human adults also bear out the prediction of skew. The responses given to items of the same class show a great degree of assymetry, to the extent that among the sequences with highest acceptability there are also a few partwords. We therefore confirm that the skewed responses are not restricted to nonhuman animals such as rats, but it also a characteristic behavior of human adults.

4.3 Fitting R&R to forced choice data

We have seen that the predictions of R&R in terms of skew are visible in the experimental data of rats and humans. However, both experiments involve a small number of subjects, so we turn our attention now to a more comprehensive study (Frank et al., 2010), in order to use a sufficiently big number of datapoints to give a quantitative measure of goodness of fit of our model.

Frank et al. assess the performance of human adults in segmenting an artificial language. Each of their three experiments involves a range of conditions that make the task more difficult in one particular dimension: (1)sentence length, (2)amount of exposure or (3)number of word types).

Their study also involves an evaluation of how existing models of segmentation fit human performance. The models were evaluated in terms of their goodness of fit to the curve that describes the average performance of the human subjects in all the conditions of a certain experiment. From all the models evaluated, the Bayesian Lexical model (Goldwater, Griffiths, & Johnson, 2006, 2009) achieved the best performance in experiments (1) and (2). Later, French, Addyman, and Mareschal (2011) show that their model (TRACX, a connectionist model based on recognition of previously encountered segments) provide even higher performance in experiments (1) and (3), but it is not evaluated on experiment (2).

This evaluation scheme is thus based on reproducing accuracies under different conditions of difficulty, averaged over all stimuli classes and averaged over all participants, with parameters of the models optimized on the entire dataset. In Alhama, Scha, and Zuidema (2015) we argue that this is too weak an evaluation to distinguish between models, but we adapt it here nevertheless to demonstrate that our model is easily comparable with other models when evaluated in the same way.

In order to use the same evaluation procedure with R&R, we need to make two small adaptations to the model. Our design of the model was inspired by the results presented in Peña et al. (2002), where the pauses in the stimuli, when present, have a length of 25ms, and this duration is supposed to be perceived by humans only subliminally. The stimuli used in Frank et al. (2010) differ significantly in the use of pauses: they have a duration of 500ms, and they are used as a separation of sentences instead of words. We adapt the formula for Retention, using an exponential parameter regulating the effect of the pauses (1). [4]

[4] As explained in footnote 3, we cannot distinguish the contributions of the two parameters A and C, so we opted to change the formula in this way. We keep the value of μ_{np} at 1.0 in this simulation so that the

$$p_2(s) = A^{length(s)\cdot\mu} \quad (1)$$

$$\mu = \begin{cases} \mu_{wp} \text{ after a pause} \\ \mu_{np} \text{ otherwise} \end{cases}$$

The other adaptation is the use of the Luce Rule. The R&R model assigns a score to each sequence: the subjective frequency. To derive prediction about behavior in a 2AFC test, we follow Frank et al. (2010) and transform scores into probabilities by applying the Luce choice rule (Luce, 1963) (2):

$$P(s_1) = \frac{SubjFreq(s_1)}{SubjFreq(s_1) + SubjFreq(s_2)} \quad (2)$$

Table 2 summarizes the goodness of fit between the models and the experimental data, using Pearson's r as a metric to compare the curves that describe how the performance evolves when the difficulty of the task is increased. As it can be seen, for the parameter setting that yields better fit in the three experiments ($A = 0.005, B = 0.948, D = 0.827, mu_{np} = 1.0, mu_{wp}$=0.288), our model outperforms all the other models in experiment 1. One possible explanation for this result is that R&R is the only model that incorporates a treatment for the effect of pauses: increasing the length of the sentence entails fewer pauses in the input stream, and therefore R&R finds longer sentence more difficult, in the same way as humans do.

For the second experiment, R&R provides the best results, together with the Bayesian TP model with restricted input and the Bayesian Lexical Model with uniform forgetting of types. R&R does not at the moment implement any form of forgetting (although it could be easily incorporated), and it does not need to force a limitation in the input data. Instead, we obtain the effect of imperfect storage of all the input data by modelling the process of retention and recognition.

Also on experiment 3, the R&R model shows the best correlation with human data. In this experiment, the performance of humans decreases when the number word types increases. It is worth noticing that R&R and TRACX reproduce this phenomenon with very good fit without the need of adapting the models for the task. In contrast, the normative models benefit from the fact that the partwords become less frequent when the number of types increases, exhibiting higher performance when that of humans decreases. This effect is counteracted in variant implementations that add a limitation of the input data or some form of forgetting: with this solution, the models invert their trend and present a much better correlation with human performance. It is fair to notice that TRACX and R&R do not need to be adapted for this experiment; on the contrary, it is an intrinsic property of their design that not all the segments of the input stream will be memorized.

The curves of the performance of both human adults and R&R can be see in figure 8.

This study shows that our model can fit 2AFC data on human adults with a correlation that is at least on par with that of other models.

5 Related work

Many models of segmentation of artificial grammar learning have been proposed. Among the models that can be classified as at Marr's rational level, the most successful approach seems to be the Bayesian model presented in Goldwater et al. (2009) and evaluated in Frank et al. (2010). Using Bayesian inference, the model considers segmentation hypotheses that are consistent with the input stream, and computes their posterior probability incorporating a prior distribution based in a Dirichlet process. The main assumptions of this process are: i) the probability of a word in the i_{th} position is proportional to the number of occurrences of this word in previous positions, and ii) the relative probability for a new word type in the i_{th} position is inversely correlated with the total number of word tokens, and iii) a new word type is more probable if it is shorter. Some of these assumptions are also embodied in R&R: assumption (i) is also contained in the recognition function in R&R, but based on subjective frequency rather than absolute frequency; and (iii) is directly encoded in the retention probability of R&R.

On the other extreme, at Marr's implementational level, some connectionist models have been proposed (Cleeremans & McClelland, 1991; Servan-Schreiber, Cleeremans, & McClelland, 1991); the most recent, TRACX (French et al., 2011; French & Cottrell, 2014) provides comparable results to the Bayesian model in modelling human's performance in a range of conditions (Frank et al., 2010). TRACX is a neural network that uses the architecture of autoencoders: it learns a representation for the input data. The error of the output layer is computed by comparing it with the input, and it serves as an indication of the degree of recognition of the input. The model processes the input stream sequentially, maintaining a context window. After successful recognition of a segment, the internal representation learned by the network is used as the context for the next segment to be presented. In this way, contiguous segments that are successfully recognized are gradually represented as a single

resulting model remains comparable.

[5]This experiment was not reported in (French et al., 2011), so we assume a Pearson R of 0.0, which means that there is no correlation.

	Exp. 1: Sentence Length	Exp. 2: Amount of tokens	Exp. 3: Word types	Avg.
Transitional Probabilities	0.84	0.43	-0.99	0.09
Mutual Information	0.83	-0.32	-0.99	-0.16
Bayesian Lexical model	0.94	0.89	-0.98	0.28
MI Clustering	0.11	-0.81	0.29	-0.13
PARSER	0.00	0.86	0.00	0.28
TRACX	0.92	0.0[5]	0.97	0.63
Bayesian TPs 4% data	0.82	0.92	0.96	0.9
Bayesian Lexical model 4% data	0.88	0.85	0.90	0.87
Bayesian Lexical model Uniform forgetting (types)	0.95	0.92	0.73	0.86
Bayesian Lexical model Prop. forgetting (types)	0.88	0.87	0.88	0.87
Bayesian Lexical model Uniform forgetting (tokens)	0.86	0.82	0.97	0.88
Retention & Recognition	**0.98**	**0.97**	**0.98**	**0.97**

Table 2: Comparison of model results to human performance from Frank et al. (2010) experiments. The reported metric is Pearson's r.

chunk, and therefore can be recognized as a unit. This approach shares with R&R the intuition that words in the input stream obtain that status after being recognized over and over.

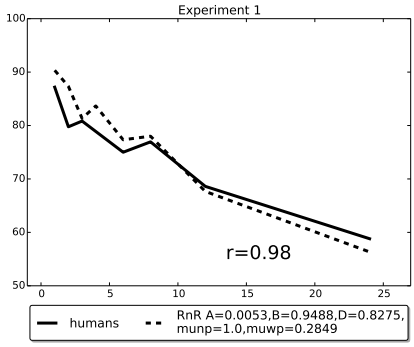
Unlike these models, R&R is clearly situated at Marr's processing level. In that category, we are only aware of the existence of PARSER, a model proposed by (Perruchet & Vinter, 1998). PARSER is a symbolic, exemplar-based model that shares many similarities with R&R. We now briefly present PARSER and discuss the similarities and differences with our model.

PARSER is built around basic principles of associative learning and chunking. Starting with a few primitives (typically, the syllables of the stream), it will incrementally build a lexicon of segments, each one with an associated weight, that will have an effect on determining which segments will be memorized next. The size of the next segment to be perceived is determined randomly; however, the units that compose this segment will be either primitives or already-memorized segments that have a weight higher than a certain threshold. As an example, if the size of the next segment to be perceived is 2, it might be composed of two primitives (syllables), two segments (larger than the syllables) or one of each. The algorithm chooses the combination that allows the largest units, from left to right. For every new segment that is perceived, its weight in memory is incremented (or it's added with an initial weight), but the smaller units that compose it are decremented; this is what the authors interpret as *interference*. Additionally, at each timestep, all the units in memory have their weights decreased.

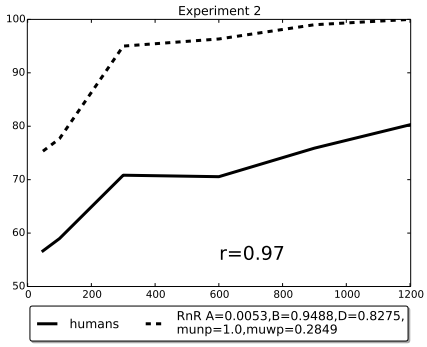
PARSER and R&R are both exemplar-based models that build a lexicon of segments (exemplars), and use this lexicon of already-memorized segments to decide on further segments to memorize. Each segment in the lexicon is stored together with a score (the *memory weight* in PARSER or the *subjective frequency* in R&R) that will determine the impact of this segment in the next steps of the segmentation process.

The models are similar in their procedure, but there are notable differences among them. One of them is their probabilistic nature. For PARSER, the stochasticity is limited to the random selection of the size of the next segment to read from the stream. In contrast, R&R is probabilistic in both of its basic processes (retention and recognition), but it does not apply a random process for selecting the size of the segments. Instead, it considers all possible segments starting from the next syllable (until a maximum length).

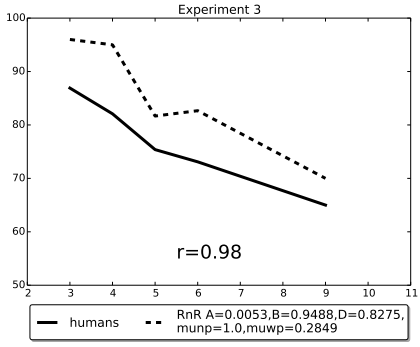
The process of retention in R&R penalizes longest segments, on the basis that they would require more working memory. PARSER implements the opposite intuition: whenever several segment candidates are possible, it selects those that are built of the longest units. Recognition is also modelled differently. PARSER implements some form of recognition when it maps the next segment to be



(a) Curve of performance over different conditions (varying sentence length) of experiment 1.



(b) Curve of performance over different conditions (varying the number of tokens) of experiment 2.



(c) Curve of performance over different conditions (varying the vocabulary size) of experiment 3.

Figure 8: Curve of performance of all experiment in Frank et al. 2010.

read against the units in memory. This is modelled as a threshold: only units with weight above the threshold can be *recognized* as part of the new segment to be read. In R&R, recognition is also based on the score (the subjective frequency) of the segments in memory, but it provides a probabilistic account rather than a binary choice; furthermore, the number of types already stored in memory also plays a role, decreasing the probability as more types are stored.

R&R does not at the moment implement any form of forgetting. Although we initially planned to add a form of decay in the recognition probability, we see in the experimental data that some of the less frequent sequences have higher responses than some of the more frequent sequences. While this is not necessarily contradictory with some form of forgetting, it is still unclear to us whether including forgetting is actually necessary to explain the experimental data.

We have implemented PARSER following the explanation in their paper, and we have used the same parameters that the authors report as the authors in their study of Peña et al. (2002) (Perruchet et al., 2004). Figure 9 shows the average responses of 14 runs of the model, for each test item, ordered by decreasing weight. The distribution of weights shows the skew that we have also presented for R&R, but all words have higher weights than all partwords. Thanks to the free parameters of R&R, we can provide a behaviour similar to PARSER but also one that gives more prominence to some of the partwords in the stream, like in the experimental results from figure 7.

The scores for partwords are also the reason the two models differ when evaluated against the 2AFC data presented in Frank et al. (2010). As can be see in table 2, R&R shows almost perfect correlation with humans in all experiments, but PARSER yields a Pearson's r of 0.0 in two out of the three experiments. As discussed in the study, the reason for this failure is that PARSER neglects the scores of partwords, which are very close to zero in all the conditions of the three experiments.

6 Conclusions

ALL has proven to be very useful for finding out what cues are exploited when subjects are faced with the task of learning an unknown language. Researchers have postulated theories about the mechanisms underlying the learning process. In this work we focus on one of the first problem that learners face: the identification of words in a speech stream.

With our model, R&R, we provide a theory that considers the process of segmentation as the interaction of two

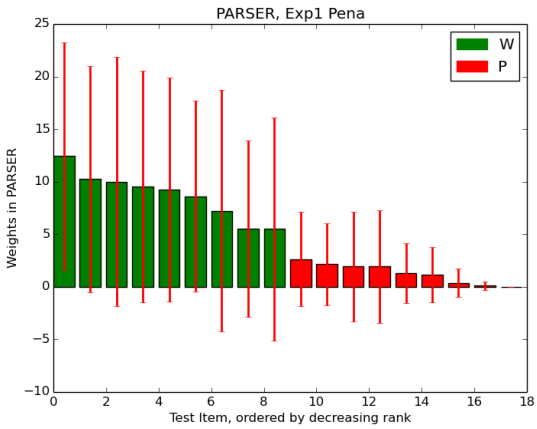


Figure 9: Responses of PARSER for the words in experiment 1 in Peña et al. (2002)

cognitive mechanisms: retention and recognition. Pitched at the processing level, and with a very simple formalization, our model offers a way to understand the pattern of experimental results that we find in the literature. Nevertheless, it is fair to point out that we have not described how our model would explain one of the relevant aspects of the experimental results we have reviewed; namely, the conditions under which subjects accept *rulewords* (sequences that do not appear in the familiarization stream but are consistent with the rules of formation of words). Our ongoing modelling research on the process of generalization to new valid exemplars is out of the scope of this paper, but it is nonetheless described as a process that operates on the segments memorized by R&R.

Models do not only help us reason about the cognitive processes underlying existing experiments, but also allow us to make predictions for experimental results. R&R predicts that the memorized segments of the familiarization stream show a strong skew in the distribution of subjective frequencies; an observation that, to our knowledge, has never been reported before.

To confirm this prediction, we have revisited the experimental results of Peña et al. (2002), on human adults, and Toro and Trobalón (2005), on rats; focusing on the responses per test item: by replicating the experiment with a different test type in the former, and by providing a more fine-grained analysis in the latter. We find that the data shows a clearly skewed distribution of responses, confirming our prediction. Furthermore, we show that R&R also provides a good quantitative fit to the experimental data of Toro and Trobalón (2005).

In order to contrast R&R with other computational models that have been proposed, we have followed the evaluation procedure presented in the work by Frank et al. (2010). This extensive study provides a quantitative comparison of many models of segmentation, based on their goodness of fit to several datapoints in 2AFC experiments. We have followed the same procedure to include our model in the study, and we have shown that R&R produces a better correlation with experimental results than the other models. Nevertheless, we raise awareness of the need of a type of evaluation that takes into account responses per individual item rather than average performance.

We further discuss how our model contrasts with some of the more relevant models in the literature. We conclude that, while some of the ideas embodied in R&R are already present in other approaches, our model constitutes a simple yet powerful characterization of the mechanisms underlying speech segmentation that shows a better correlation with the experimental data, and that has already allowed us to provide a new observation of the existing data, proving therefore to be a promising tool for revealing the properties of this basic process of language learning.

7 Acknowledgments

We thank Juan M. Toro for sharing the data of his experiments with us. We are grateful to Carel ten Cate, Clara Levelt, Michelle Spierings, Andreea Geambasu and Padraic Monaghan for their feedback.

References

Alhama, R. G., Scha, R., & Zuidema, W. (2015). How should we evaluate models of segmentation in artificial language learning? In *Proceedings of 13th international conference on cognitive modeling*.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*(3), 235.

Endress, A., & Bonatti, L. (2006). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*(2), 247–299.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*.

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

French, R. M., & Cottrell, G. W. (2014). Tracx 2.0: A memory-based, biologically-plausible model of sequence segmentation and chunk extraction..

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).

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

Luce, R. D. (1963). Detection and recognition. In *Handbook of mathematical psychology*. New York: Wiley.

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

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information*. New York: W. H. Freeman.

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

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.

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

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

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.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. L. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning*, *7*(2-3), 161–193.

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