Segmentation as Retention and Recognition: the R&R model

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

Abstract

We present the Retention and Recognition model (R&R), a probabilistic exemplar model that accounts for segmentation in *Artificial Language Learning* experiments. We show that R&R provides an excellent fit to human responses in three segmentation experiments with adults (Frank et al., 2010), outperforming existing models. Additionally, we analyze the results of the simulations and propose alternative explanations for the experimental findings.

Keywords: artificial language learning; segmentation; statistical learning; cognitive modelling

Introduction

A crucial step in the acquisition of a spoken language is to discover what the building blocks of a speech stream are. Children perform such segmentation by exploiting a variety of statistical and prosodic cues in the input. Understanding the unique ability of humans to acquire speech requires an understanding of the nature of this learning mechanism.

Artificial Language Learning (ALL henceforth) has, over the last 20 years, become a key paradigm to study the nature of learning biases in speech segmentation and rule generalization. In experiments in this paradigm, participants are exposed to artificial stimuli designed to incorporate particular aspects of speech and language, and they are subsequently tested on whether and under which conditions they discover the regularities in such artificial language.

A key result in this tradition is the demonstration that 8 month old infants are sensitive to transition probabilities between syllables, and can segment a speech stream based on these probabilities alone (Saffran, Aslin, and Newport (1996), Aslin, Saffran, and Newport (1998)). This ability to track statistics over concrete fragments of the input, known in the literature as *statistical learning*, has also been demonstrated in adults (Saffran, Newport, & Aslin, 1996).

However, these experiments do not reveal whether the underlying cognitive mechanism does operate over transitional probabilities or, instead, it performs computations of an entirely different nature but which can be *described* as transitional probabilities. In order to reveal the precise underpinnings of such cognitive mechanism, a useful methodology is computational modeling.

There exist several segmentation models in the literature, offering alternative accounts of the nature of this process. Thus, these models need to be compared and analyzed against empirical data to validate their predictions. Possibly the most comprehensive study for the evaluation of computational models in segmentation is presented in Frank et al. (2010). In that study, the authors evaluate a range of models based on their goodness of fit to three segmentation experiments that involve a great number of different conditions –thus providing a rich dataset for comparing the models.

In this paper we present one model for to account for segmentation experiments in ALL. Our model, called the *Retention & Recognition* model (henceforth R&R), is a novel processing model that explains segmentation based on the *retention* and *recognition* of subsequences of the input. Following Frank et al., we test our model against the experimental data from their study, and compare the goodness of fit of our model with those reported in previous studies.

The R&R Model

The model we propose, which we call the Retention-Recognition Model (R&R), takes a sequence of syllables $X = \langle x_o, x_1, x_2, \ldots, x_m \rangle$ as input, and considers all subsequences of length $l = 1, 2, \ldots, l_{max}$ as potential segments to be memorized.

The model maintains a memory M, which is a set of segment types and their associated counts. The memory is initially empty ($M_0 = \emptyset$) and it changes with update steps that either *add* an entry (with count 1) or *increase* the count of an existing entry:

ADD: $M_{t+1} \leftarrow M_t \cup \{\langle\langle x_j, \ldots, x_k \rangle, 1\rangle\}$

INCREMENT:
$M_{t+1} \leftarrow M_t - \{\langle\langle x_j, \ldots, x_k \rangle, c\rangle\} \cup \{\langle\langle x_j, \ldots, x_k \rangle, c+1\rangle\}$

For any candidate segment $s \in S$ (with segments processed in the order they are encountered in the stream), the model checks whether it is stored in memory and, if so, what the count of that segment in memory is (its 'subjective frequency'). The model may (with a probability p_1 that increases with that count) *recognize* it (i.e., match it with a segment in memory). If it succeeds, the count is incremented with 1. If it fails to recognize the segment, the model might (with a probability p_2 that decreases with the length of the segment) still *retain* it (i.e., add it to memory with initial count of 1 if it was not stored, or in the event that a previously stored segment was not recognized and is retained –very rare in practise– increase the count 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. R&R's flowchart is given in Figure 2.

The key components of the model are the equations for

computing the recognition probability (p_1) and retention probability (p_2). Recognition should become more probable the more often a segment has been recognized, but decrease with the number of segment types in memory ($|M|$). Hence, we define p_1 as follows, with B and D free parameters ($0 \leqslant B, D \leqslant 1$) that can be fitted to the data:

$$p_1(s,M) = (1 - B^{\text{COUNT}(s,M)}) \cdot D^{|M|} \quad (1)$$

If a segment is not recognized, the model considers *retaining* it with a probability that decreases with the length of the segment ($l(s)$), and which can be boosted if there are additional cues favoring this segment (e.g., a pause preceding it). Hence, we define p_2 as follows, with A and μ free parameters ($0 \leqslant A \leqslant 1; 0 \leqslant \mu$) that can be fitted to the data:

$$p_2(s) = A^{length(s) \cdot \mu} \quad , where\ \mu = \begin{cases} \mu_{wp} \text{ after a pause} \\ \mu_{np} \text{ otherwise} \end{cases} \quad (2)$$

The A parameter thus describes how quickly the retention probability decreases with the length of a segment. The probability is also affected by the presence of additional cues; in this paper, we consider only the pauses between sentences as additional cues. [1]

Putting everything together, the model can be described in pseudocode as in Figure 1. As can be seen, R&R is a simple model, but it gives a surprisingly accurate match with empirical data, as we will explore in the next sections, without even taking processes such as forgetting, priming, interference and generalization into account.

Related Models

There exist several models of segmentation in the literature. We do not have the space to address them all here, but we discuss how our model relates to those to which it has more similarities.

The *recognition* component of our model yields *rich-get-richer* dynamics (and thus consistently produces very skewed count distributions over segments in memory) similar to that of non-parametric Bayesian models, such as the Bayesian Lexical Model (BLM henceforth) in Goldwater, Griffiths, and Johnson (2009) (adapted for ALL in Frank et al. (2010)). The BLM implements such dynamics with 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; (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. Assumption (ii) does not allow for direct comparison, since R&R is not a generative model, and therefore it does not provide a probability for new types —rather, the incorporation of new types to the memory of the model depends on the retention probability, and it is based on a preference for shorter sequences (an intuition encoded also in assumption (iii) of the Bayesian model). As for assumption (i), the same principle is incorporated in the recognition process in R&R; however, in our model, the counts of the number of occurrences of a word is based on the subjective frequencies resulting from memorization, while in the BLM, these counts are based on absolute frequencies of the current hypothesis. This reflects a fundamental difference between the two approaches, which concerns their level of analysis (Marr, 1982). The Bayesian model is framed at Marr's computational level, and thus, it operates over the whole stimuli, since it does not incorporate perceptual or memory constraints (although some of the extensions in Frank et al. (2010) experiment with limitations on memory capacity, leading to a somewhat hybrid model; we return to this point later). In other words, the BLM is not proposed as a mechanistic explanation of the cognitive processes involved in the experiment; on the contrary, R&R is a processing model, which postulates that cognitive processes of retention and recognition, and psychological representations of exemplar segments are responsible for segmentation.

An existing model that is also pitched at Marr's processing level is PARSER (Perruchet & Vinter, 1998). PARSER is a symbolic model, built around basic principles of associative learning and chunking, that shares many similarities with R&R. Both PARSER and R&R are exemplar-based models that build a lexicon of segments (exemplars), and use this

```
Input: Stream X, and empty memory M_0 ← ∅.
Output: Memory M_{n+1}.
/* Compute candidate segments: */
S ← ⟨s_0, s_1, ..., s_n⟩
/* Process each segment: */
for i = 0 to n:
    /* Compute the recognition probability: */
    p_1 = p_1(s_i, M_i)
    /* Compute the retention probability: */
    p_2 = p_2(s_i, M_i)
    /* Draw two random numbers */
    r_1 ~ U(0,1)
    r_2 ~ U(0,1)
    /* Recognize, retain or ignore: */
    IF (r_1 < p_1)
      M_{i+1} ← increment(s_i, M_i)
    ELSE IF (r_2 < p_2)
      M_{i+1} ← add(s_i, M_i)
    ELSE
      M_{i+1} ← M_i
```

Figure 1: Pseudocode describing the R&R model.

[1] An earlier version of R&R (Alhama, Scha, and Zuidema (2016), Alhama and Zuidema (2016)) features a different probability for retention, with a binary switch over an attenuation parameter. This design was inspired by experimental studies in which the stimuli eventually contained 25ms pauses, a duration that is supposed to be perceived by humans only subliminally. The stimuli we plan to use for our simulations, based on Frank et al. (2010), differ significantly in the use of pauses, which have a duration of 500ms (and therefore should be clearly perceived). The retention probability we present here is more general, since the effect of pause length could be accounted for with different values of μ.

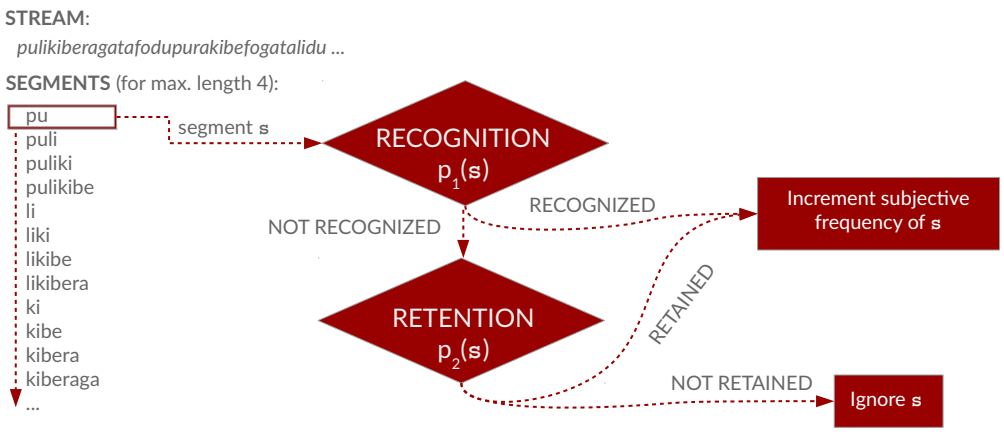


Figure 2: R&R:The Retention-Recognition Model

lexicon of already-memorized segments to decide on further segments to memorize. Each segment in the lexicon is stored together with a score that determines the impact of this segment in the next steps of the segmentation process. Thus, the models are similar in their procedure, but there are notable differences between them. One of them is the probabilistic nature of their components. 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 considers all possible subsequences of the stream (up to a maximum length), as inspired by research in Data-Oriented Parsing tradition (Scha (1990), Zuidema (2006)). Additionally, the model is inherently probabilistic in its basic processes of retention and recognition.

There exist other differences in the procedure of these approaches. To begin with, the process of retention in R&R penalizes longest segments, on the basis that they would require more working memory. However, PARSER is a *chunking* model, so it implements the opposite principle: whenever several segment candidates are possible, it selects those that are built of the longest units, creating in this way a bias for larger units. As for the process of recognition, it is implicitly implemented in PARSER when it maps the next segment to be read against the units in memory. This process involves a binary threshold: only units with weight above the threshold can be recognized as components of the segment (but those below the threshold are retained). In contrast, the interaction between recognition and retention in R&R is based on a graded probabilistic choice. Finally, an important difference between the models is that R&R does not implement any form of forgetting. Although we do not claim that humans are endowed with perfect memory, our results suggest that forgetting does not seem to play a key role in the timecourse of the experiments.

On the other extreme, at Marr's implementational level, we find TRACX (French et al., 2011; French & Cottrell, 2014), a connectionist proposal that is also based on the recognition of subsequences. TRACX is an autoencoder model that 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 chunk, and therefore can be recognized as a unit. This approach shares with R&R the intuition that words are consolidated in memory after repeated recognition; however, like PARSER, TRACX is a chunking model, that is, it is oriented to the integration of syllables in order to build larger fragments. In contrast, in R&R, words emerge in a process that actually penalizes larger fragments, as a consequence of consolidated memorization of statistically salient segments.

To sum up, R&R constitutes a new approach to modelling segmentation that offers a processing level explanation of the identification of words in a speech stream, which emerges as a result of the interplay between probabilistic memory processes. We now proceed to validate this model against empirical data.

Fitting R&R to Experimental Data

Experimental Results

Frank et al. investigate how distributional aspects of an artificial language have an effect on the performance of human adults in segmentation. Each of their three experiments involves a range of conditions that vary in one particular dimension: (i) sentence length, (ii) amount of exposure (number of tokens) and (iii) vocabulary size (number of word types).

The stimuli consists of an auditory sequence of sentences, each of which is created from a sample of artificial (unexisting) words. The sentences are separated with a silence gap

of 500 ms, while there is no acoustic nor prosodic cue indicating the separation between words within a sentence. After the participants have been exposed to a sample of sentences thus constructed, they participate in a 2-Alternative-Forced-Choice test (2AFC). The two alternatives in the test consist on one word from the artificial language (a correctly segmented sequence), and one "part-word" (a sequence resulting from incorrect segmentation).

To analyze the results, the mean number of correct choices is computed across participants in each condition. The curves formed by these datapoints (ordered by condition value) is taken as indication of how segmentation performance is affected by the varied dimension. These curves (which are shown in the continuous line in Figure 3) show that: (i) human adults have more difficulty in segmenting words when sentences are longer, presumably because they do not benefit from the extra cue provided by the silence gaps; (ii) when the amount of word tokens is varied, more occurrences of words facilitate the identification of such words, and (iii) the size of the vocabulary seems to cause lower performance in the experiment, with an almost-linear inverse relation.

Goodness of fit

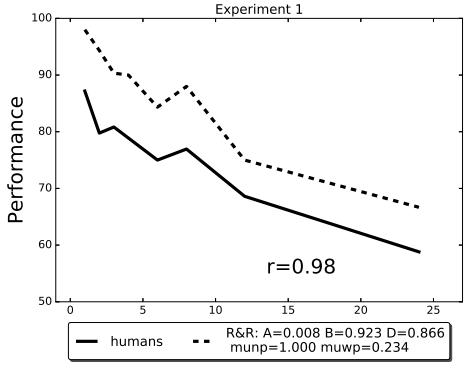
The study by Frank et al. evaluates a number of segmentation models in terms of their goodness of fit to the curve that describes the average performance of the human subjects. The evaluated models include the ones previously described (BLM, PARSER, and later, also TRACX, reported in French et al. (2011)), and four additional approaches, all of them consisting on normative models: Transitional Probabilities (TP), a Bayesian version of TP (by Frank et al.), Mutual Information (MI), and a version of MI model that identifies words when they exceed a threshold both on MI and raw frequency counts (MI Clustering, Swingley (2005)).

In order to compare the models, Frank and colleagues convert the output of each model to a metric that can be interpreted as behavioural predictions for the 2AFC task. To do so, they employ the Luce Rule (Luce, 1963). Given a pair of sequences s_1 and s_2 in test, the Luce Rule defines the probability of choosing s_1 as can be seen in Equation 3:

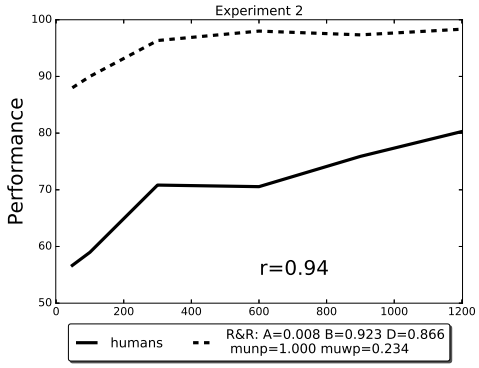
$$P(s_1) = \frac{SubjFreq(s_1)}{SubjFreq(s_1) + SubjFreq(s_2)} \quad (3)$$

Once the scores have been transformed to probabilities, the performance of the models is computed as the mean probability of choosing the correct item, averaged over participants and test trials. These datapoints are arranged in a curve in the same way as with human participants, and the correlation in the shape of these curves —measured with Pearson's r— is taken as an indication of good fit.

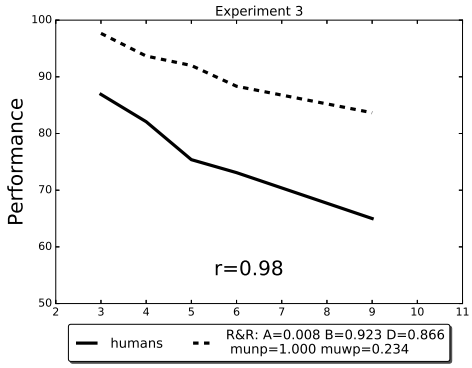
Likewise, we run simulations of the three experiments with R&R, transforming its output (the subjective frequencies) into test trials with the Luce Rule. We run a search over the parameter space, in order to find which parameters yield best correlation with human performance[2] [3]. The best results are shown in Table 1. As it can be seen, our model outper-



(a) Varying sentence length (experiment 1).



(b) Varying the number of tokens (experiment 2).



(c) Varying the vocabulary size (experiment 3).

Figure 3: Curve of performance for all conditions in the experiments in Frank et al. (2010).

[2]The only parameter that we keep fixed in our search is $\mu_{np} = 1.0$, since the interpretation of the relative importance of pauses is clearer if only one of the μ parameter is varied.

[3]We optimize our parameters on the same data we evaluate the model on, as seems to have been the case for the models we compare with. This brings the risk of overfitting, so in the discussion section we briefly discuss better ways of evaluating models.

Table 1: Comparison of model results to human performance. The reported metric is Pearson's r. *Experiment 2 was not reported in French et al. (2011). Therefore, the mean can be taken to be 0.63 (for a Pearson's r of 0.0 in experiment 2) or 0.945 (averaging only over experiments 1 and 3).

		Exp. 1: Sentence Length	**Exp. 2: Amount of tokens**	**Exp. 3: Word types**	**Mean**
1	Transitional Probabilities	0.84	0.43	-0.99	0.09
2	Mutual Information	0.83	-0.32	-0.99	-0.16
3	MI Clustering	0.11	-0.81	0.29	-0.13
4	PARSER	0.00	0.86	0.00	0.28
5	TRACX	0.92	—	0.97	—*
6	BLM	0.94	0.89	-0.98	0.28
7	Bayesian TPs 4% data	0.82	0.92	0.96	0.90
8	BLM 4% data	0.88	0.85	0.90	0.87
9	BLM Uniform forgetting (types)	0.95	0.92	0.73	0.86
10	BLM Prop. forgetting (types)	0.88	0.87	0.88	0.87
11	BLM Uniform forgetting (tokens)	0.86	0.82	0.97	0.88
12	**R&R**	**0.98**	**0.94**	**0.98**	**0.97**

forms all the other models in the three experiments, with a parameter setting that is common to the three experiments ($A = 0.008, B = 0.923, D = 0.866$, $\mu_{np} = 1.0$, $\mu_{wp} = 0.234$). The curves of the performance of both human adults and R&R can be see in Figure 3.

When it comes to experiment 1, one possible explanation for this result is that R&R is the only model that explicitly models the effect of the silence gaps. By increasing the length of sentences while keeping the number of types and tokens constant, the stimuli necessarily consists of fewer sentences when those are made longer; therefore, the number of silence gaps also decreases. For this reason, the performance of R&R declines with longer sentences, since it cannot obtain the same benefit from exploiting silence gaps. This explanation for the superior performance can be supported by looking at the values of the μ_{wp} parameter: the best fit of the model requires a low value for this parameter ($\mu_{wp} = 0.234$)), so in the presence of a pause it substantially boosts the otherwise very small ($A^{\mu_{np}} = 0.008$) retention probability.

In the second experiment, normative models based on point estimates (those based on TP and MI) do not offer a good fit with the data, since those metrics do not benefit from the accumulation of evidence offered by the increased number of tokens (contrary to humans). Frank et al. suggest that humans may be forgetting much of what they hear, which would explain the increased performance with the number of tokens. However, the extended versions of the BLM that incorporate some form of evidence limitation (with input data restricted to a random 4% sample) or forgetting exhibit mixed results (rows 8, 9, 10, 11 on table 1). Moreover, these extensions appear unrealistic from a cognitive perspective (e.g. one of the extensions forgets a random token when the memory capacity is full), and additionally, the resulting models are somewhat difficult to interpret, since after incorporating memory limitations, they are not computational level approaches anymore. PARSER offers a more intuitive account of forgetting, with modest correlation with human data; however, this model has zero correlation in the other experiments. So this pattern of results suggests that a rich-get-richer form of recognition combined with a process of retention as defined in R&R seems a more compelling explanation than a process of recognition with forgetting.

Also on experiment 3, the R&R model exhibits the best correlation with human data, followed closely by TRACX. Again, normative models show the opposite trend from humans (rows 1, 2, 3, 6 on table 1), since they do not have any memory limitations, and thus the effect of increasing vocabulary size only has an effect in the distributional properties of the stream, which result in less statistically coherent partwords. This is the case also for PARSER and the BLM. Frank et al. attribute this failure to the lack of forgetting in the models, but the same issues we have discussed above apply to this experiment. Therefore, the more convincing approaches are TRACX and R&R. But although TRACX naturally reproduces the human results without forgetting, it is difficult to interpret what is the component of the model that is responsible for its success in this experiment. Conversely, R&R explicitly incorporates a parameter that penalizes recognition based on the number of memorized types. In line with our intuitions, the corresponding parameter value for the best fit amounts to $D = 0.86$, which results in a relatively large penalization for recognition[4]. Therefore, in conditions of high number of types, humans have an increased difficulty in recognizing sequences, most likely originating from the process of matching the input segment to one of the many segments stored in memory.

[4]Even though the values that parameter D can take range from 0.0 to 1.0, the number of types stored by R&R grow very rapidly in our model due to the memorization segments of any length. For this reason, small values are impracticable, since the probability of recognizing a segment quickly drops close to zero.

Discussion

With our model, R&R, we provide a theory of the process of segmentation based on the interaction of two cognitive mechanisms of memorization. We believe that one of the best features of our model is its transparency: pitched at the processing level, and with a very simple formalization that involves clearly identified components, R&R allows for straightforward interpretation of the results. Even though, for reasons of space, we have not been able to report a thorough analysis of the behaviour of the model under different parameter settings, we have shown a glimpse on how these parameters allow for the identification of the relative importance of each component.

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. Even though we consider that the evaluation data and procedure initiated by Frank et al. is one of the most thorough in the ALL modelling literature, in Alhama et al. (2015) we argue that averaging the responses over stimuli classes is likely to mask important differences between otherwise seemingly equivalent models. The work reported in this paper is a necessary first step to confirm that R&R is comparable to other models, but for future work it is important to move to evaluating models based on response distributions over individual test items (albeit our first attempts to evaluate our model with this procedure are inconclusive), and replace the Luce choice rule and correlation metric with a more cognitively realistic response model.

Finally, segmentation is a fundamental ability for language learners, but any segmentation model must at some point be related to other cognitive mechanisms that operate in natural and artificial language learning. In Alhama and Zuidema (2016) we show that the subjective frequencies computed by R&R have the necessary distributional properties to explain some of the main results in rule learning in ALL. Future work may explore how the model relates to other linguistic processes (e.g. word learning), so that we can eventually achieve a complete understanding of how segmentation relates to the complete picture of language learning.

Acknowledgments

This work was developed with Remko Scha, who sadly passed away before the finalization of this paper. We are grateful to Andreea Geambasu, Clara Levelt, Michelle Spierings, Carel ten Cate and Padraic Monaghan for their feedback. This research was funded by a grant from the Netherlands Organisation for Scientific Research (NWO), Division of Humanities, to Levelt, ten Cate and Zuidema (360-70-450).

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.*

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.*

Alhama, R. G., & Zuidema, W. (2016). Generalization in Artificial Language Learning: Modelling the propensity to generalize. In *Proceedings of the 7th workshop on Cognitive Aspects of Computational Language Learning* (pp. 64–72). Berlin: Association for Computational Linguistics.

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.

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

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. In *Proceedings of the 36th annual conference of the cognitive science society.*

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.

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

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

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.

Scha, R. (1990). Taaltheorie en taaltechnologie; competence en performance. In R. de Kort & G. Leerdam (Eds.), *Computertoepassingen in de neerlandistiek* (pp. 7–22). Almere, the Netherlands: LVVN. (English translation at http://iaaa.nl/rs/LeerdamE.html.)

Swingley, D. (2005). Statistical clustering and the contents of infant vocabulary. *Cognitive Psychology*, *50*(1), 86–132.

Zuidema, W. (2006). What are the productive units of natural language grammar?: a DOP approach to the automatic identification of constructions. In *Proceedings of the Tenth Conference on Computational Natural Language Learning* (pp. 29–36).