How should we evaluate models of segmentation in artificial language learning?

Raquel G. Alhama (rgalhama@uva.nl)

Remko Scha (scha@uva.nl)

Willem Zuidema (w.h.zuidema@uva.nl)
Institute of Language, Logic and Computation, Science Park 107
Amsterdam, 1098XG, The Netherlands

Keywords: Statistical Learning; Cognitive Modelling

Introduction: Statistical Learning

One of the challenges that infants have to solve when learning their native language is to identify the words in a continuous speech stream. Some of the experiments in Artificial Grammar Learning (Saffran, Newport, and Aslin (1996); Saffran, Aslin, and Newport (1996); Aslin, Saffran, and Newport (1998) and many more) investigate this ability. In these experiments, subjects are exposed to an artificial speech stream that contains certain regularities. Infants are typically tested in a preferential looking paradigm; adults, in contrast, in a 2-alternative Forced Choice Tests (2AFC) in which they have to choose between a word and another sequence (typically a *partword*, a sequence resulting from misplacing boundaries).

One of the key findings of AGL is that both infants and adults are sensitive to transitional probabilities and other statistical cues, and can use them to segment the input stream. Several computational models have been proposed to explain such findings. We will review how these models are evaluated and argue that we need a different type of experimental data for model evaluation than is typically used and reported. We present some preliminary results and a model consistent with the data.

Models of Segmentation

Many different types models of segmentation have been proposed, that differ in the representational framework used (including symbolic, statistical, connectionist and exemplar-based representations, and combinations thereof) and in the level of description chosen. We focus here on three representative models: (i) Goldwater, Griffiths, and Johnson (2009) present a Bayesian rational model, which assumes an ideal learner and computes the most probable set of segments that could have produced the observed stream. (ii) PARSER (Perruchet & Vinter, 1998) is a symbolic, exemplar-based model that incrementally breaks the input stream into segments and memorizes them; when the weight of a segment in memory is strong enough, it influences how the subsequent part of the stream is segmented. The model also incorporates forgetting and interference. (iii) In the connectionist paradigm, TRACX (French et al., 2011) present a recognition-based neural network that learns to represent the input. The resemblance of the output representations with the input sequence indicates how well the sequence is recognized by the model.

Which of these models fits the experimental data best? That question turns out to be difficult to answer, as data from 2AFC experiments – the vast majority of experiments with adults in the AGL paradigm – make it difficult to choose between models, despite important differences in the cognitive processes they assume. This is because in 2AFC, only the relative preference of one stimulus over another one is measured, and typically only a single average accuracy is reported. All the models we considered have enough free parameters to reproduce any desired average accuracy.

In existing work on model evaluation, several authors have therefore proposed to focus on the analyses of the internal representations of the model (consisting on segments of the stream, as well as some score in the form of memory strength or probability), or on comparing the performance of the model in a 2AFC setting over a range of conditions. Perruchet and Vinter (1998) provide an example of the former. They define two criteria: the *loose criterion* states that the internal memory of the model contains the words with the highest weights, but also other sequences; to fulfill the *strict criterion*, the memory must contain the words with the highest weights, but other 'legal' sequences are possible (ex: two words concatenated). We will show that the assumptions that *all* the words should have the highest weights, and that the non-legal sequences should be forgotten, is not consistent with empirical data.

Frank, Goldwater, Griffiths, and Tenenbaum (2010), on the other hand, provide an example of the latter. They evaluate a number of different models by comparing the performances in a 2AFC task with those of humans for a range of conditions (e.g., for different numbers of words). Although this constitutes a major improvement over comparing only to a single datapoint, we still find that models which embody fundamentally different assumptions can easily provide similar performance. This is the case, for instance, with the Bayesian model of (Goldwater et al., 2009) and TRACX (French et al., 2011).

A call for a different type of experiments

We suggest a different experimental setup that we believe should complement the extensive body of research with 2AFC tests, and we advance some preliminary results.

In our experiment, we replicated the familiarization phase of experiment 1 in Pena, Bonatti, Nespor, and Mehler (2002).

In the test phase, each trial consisted of two questions about a single sequence (either a word or a partword). The first question was "Is this sequence a word of the language you have heard?", and it allows for a yes/no answer. The following question was a confidence rate about the previous answer, from 1 (not confident) to 7 (very confident).

Figure 1 shows the average responses for each test item. We do not claim that these responses provide us with direct access to the strength of the mental representations of the subjects, but we believe they are more revealing than the 2AFC responses.

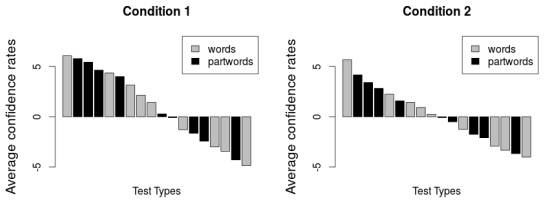

Figure 1: Average confidence rates for each test stimulus type, in decreasing order. Confidence rates for negative answers have negative values. Conditions only differ in the randomization of the syllables.

An important observation of these responses is that some partwords are rated higher than some of the words. This invalidates the criteria proposed by Perruchet and Vinter (1998), and justifies the need of models that store segments creating the skewed distributions observed. In the next section we present a model that can generate this kind of output.

The Retention & Recognition Model

Our model, the Retention&Recognition model (RnR) (Alhama, Scha, & Zuidema, 2014) can be considered a probabilistic chunking model. It incrementally breaks the stream into segments which may be stored into its internal memory, along with a weight that we call 'subjective frequency'. Given an initially empty memory, for any segment from the input stream, the model will attempt to recognize it with probability P_{rec} (eq. 1). If it succeeds, the subjective frequency of the segment is incremented with 1. If it fails, the model may still retain it in memory, with probability P_{ret} (eq. 2). In this case, it will either add it for the first time with initial subjective frequency one, or increment its subjective frequency with 1.

$$P_{rec}(segment) = (1 - B^{activation(segment)}) \cdot D^{\#types} \quad (1)$$

$$P_{ret}(segment) = A^{length(segment)} \cdot C^{\pi} \quad (2)$$

The model involves free parameters (A, B, C, D) that may be fitted to empirical data. The retention probability is inversely correlated with the length of the segment. The factor C^{π} attenuates the retention probability unless the segment appears right after a micropause. The recognition probability uses an activation function that depends on the accumulated subjective frequency of the subsequence. The number of word types adds difficulty to the task, resulting in a decreased recognition probability.

The interaction between retention and recognition can generate a range of results similar to those seen in the experiment. The recognition formula provides the dynamics of *rich get richer* (also present in some nonparametric bayesian models): once a sequence starts being recognized, it will be easier and easier to recognize, leading to a big subjective frequency. However, the retention and the probabilistic nature of the model are responsible for the fact that not all sequences are first incorporated into memory at the same time. This yields skewed distributions, similar to the distribution observed in our experiment.

Conclusions

In order to choose between models of segmentation we need to contrast them with data from experimental paradigms that complement the existing 2AFC results with data that shows other properties. We have proposed an alternative paradigm for experiments, with the hope that it will inspire many more experiments along this line. We have also illustrated how it provides empirical support for the misrepresentation of partwords in some models. Finally, we have described RnR, a probabilistic chunking model that can reproduce the patterns revealed by the data.

References

Alhama, R. G., Scha, R., & Zuidema, W. (2014). Rule learning in humans and animals. In *Proceedings of the international conference on the evolution of language* (p. 371-372).

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

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.

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

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

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.