How should we evaluate models of segmentation in Artificial Grammar Learning?

Raquel G. Alhama, Remko Scha, Willem Zuidema

Institute for Logic, Language and Computation. University of Amsterdam.

We analyze Artificial Grammar Learning models of segmentation and we find that models that embody different theories make similar predictions on traditional evaluation methods. We conclude that the experimental results should be reported as a response distribution over individual test items, and evaluation methods should be based on the goodness of fit to this distribution.

Data

Artificial Grammar Learning

Frank et al. 2010 (segmentation on adults):

FAMILIARIZATION (auditory stream):

kidadatulatikidapalapi datulatikidapalapikida ...

Sentence length

CONDITIONS:
Exp. 1: sentence length (1,2,3,4,6,8,12,24)
Exp. 3: word types (3, 4, 5, 6, 9)

TEST:

Which one sounds more like a word in the language you've just heard?
Choose one: *latiki* | *palapi*

RESULTS:
Human performance (% correct)

SEMINAL RESULTS ON WORD SEGMENTATION:

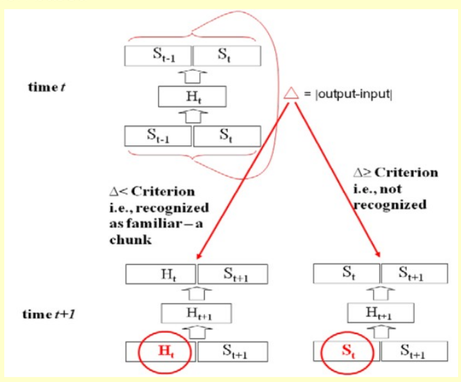
- Saffran et al. 1996: adults can segment a speech stream on the basis of statistical cues
- Saffran et al. 1996: 8 m.o. can segment a speech stream on the basis of statistical cues
- Aslin et al. 1998: 8 m. o. can segment a speech stream on the basis of transitional probabilities

Models

TRACX

French, R. M.; Addyman, C. & Mareschal, D. (2011)

- Neural network: autoassociatior
- Recognition based

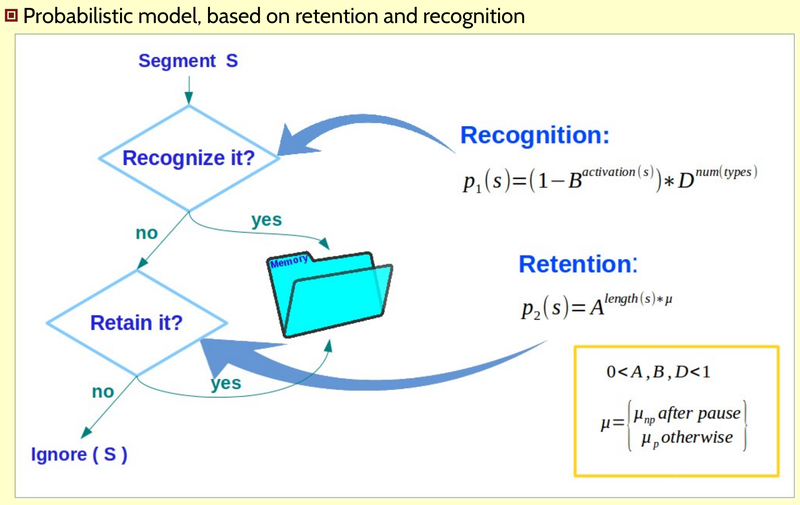


When an item is recognized, its representation in the hidden layer becomes the left part of the next input.

Retention & Recognition

Alhama, R. G.; Scha, R. & Zuidema, W. (2014)

- Probabilistic model, based on retention and recognition



Bayesian Lexical Model

Goldwater, S.; Griffiths, T. L. & Johnson, M. (2009)

P(H|D) = P(D|H) * P(H)

- Hypotheses: sequences of word tokens
- Likelihood: always either 0 or 1 because every sequence of word tokens is either entirely consistent or entirely inconsistent with the input data.
- Prior: assumes that syllables have been generated with a Dirichlet process:

$$P(w_i=w|w_1\dots w_{i-1})=\frac{n_{i-1}(w)+\alpha P_0}{i-1+\alpha}$$

  - a word is more probable if it has occurred many times already (n)
  - relative probability of a novel word:

$$P_0(w=x_1\dots x_m)=\prod P(x_j)$$

- Posterior: the model need only consider consistent segmentations of the input, and of these, the one with the highest prior probability is the optimal hypothesis.

Evaluation Criteria

PEARSON R OVER PERFORMANCE CURVE (Frank et al. 2010)
Models apply Luce Choice Rule (in 2AFC) over their scores S:
P(a) = S(a) / (S(a) + S(b))
The performance of the model is compared to that of humans using the Pearson R with the curve defined by the average performance in every condition of an experiment.

ALTERNATIVES:

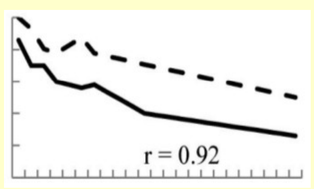
LOOSE CRITERION (Perruchet et al. 1998)
The internal memory contains the words with the highest weights, but also other sequences.

STRICT CRITERION (Perruchet et al. 1998)
The internal memory contains the words with the highest weights, but other legal sequences are possible.

PROPORTION BETTER (French et al. 2011)
Relative score for sequences a and b:
Prop. Better = (S(a) – S(b)) / (S(a) + S(b))

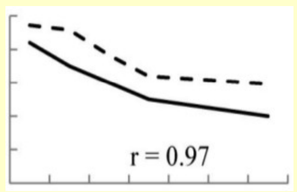
TRACX

Pearson R



Frank et al. EXP1

Pearson R

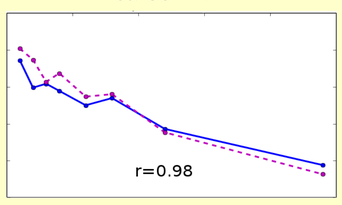


Frank et al. EXP3

Retention & Recognition

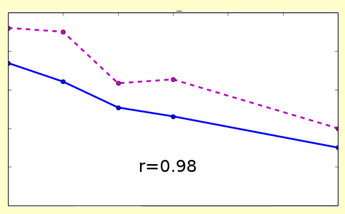
(with parameters: A=0.005, B=0.948, D=0.827, munp=1.0, muwp=0.284)

Pearson R



Frank et al. EXP1

Pearson R

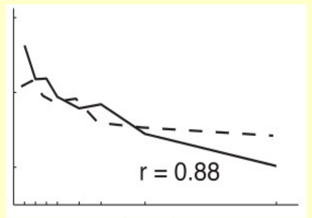


Frank et al. EXP3

Bayesian Lexical Model

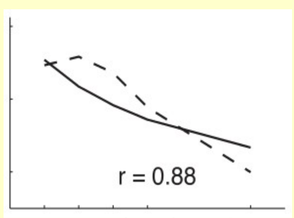
(extended with forgetting proportional to types)

Pearson R



Frank et al. EXP1

Pearson R



Frank et al. EXP3

Conclusion

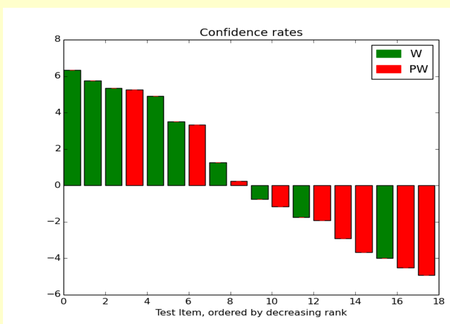
All these models approximate human performance under this evaluation. How do we choose one?

→ We should look at data of responses per item

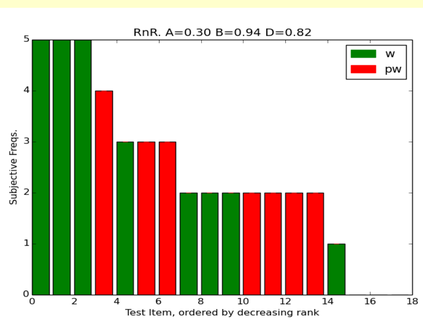
→ Don't average over stimuli types and participants!

We propose:

- Using experimental types that provide graded responses per item. Ex:
  - Familiarization: listening to an unsegmented speech stream.
  - Test: ask, for a single sequence (word or partword):
    - Is this sequence a word of the language you have heard?
      Yes / No
    - How confident are you?
      [Not confident] 1-2-3-4-5-6-7 [Very confident]
- Reporting response distribution over individual test items
  Recipe:
  - for each subject and each class of stimuli, order the responses by rank
  - average over individual test items maintaining their order
- Evaluating models on their goodness of fit to this response distribution



Some partwords are rated higher than some words!



So we need models that provide this kind of profile.

References:

- Alhama, R. G.; Scha, R. & Zuidema, W. Rule Learning in Humans and Animals. *Proceedings of the International Conference on the Evolution of Language,* 2014
- Aslin, R. N.; Saffran, J. R. & Newport, E. L. Computation of conditional probability statistics by 8-month-old infants. *Psychological science,* 1998
- Frank, M. C.; Goldwater, S.; Griffiths, T. L. & Tenenbaum, J. B. Modeling human performance in statistical word segmentation. *Cognition*, 2010
- French, R. M.; Addyman, C. & Mareschal, D. TRACX: A Recognition-Based Connectionist Framework for Sequence Segmentation and Chunk Extraction *Psychological Review*, 2011
- Goldwater, S.; Griffiths, T. L. & Johnson, M. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 2009,
- Perruchet, P. & Vinter, A. PARSER: A model for word segmentation Journal of *Memory and Language*, 1999
- Saffran, J. R.; Aslin, R. N. & Newport, E. L. Statistical learning by 8-month-old infants. *Science*, 1996
- Saffran, J. R.; Newport, E. L. & Aslin, R. N. Word segmentation: The role of distributional cues. *Journal of memory and language*, 1996

Institute for Logic, Language and Computation

Universiteit van Amsterdam