MAIR Cognitive Processing 4

Models of human language acquisition

Frans Adriaans

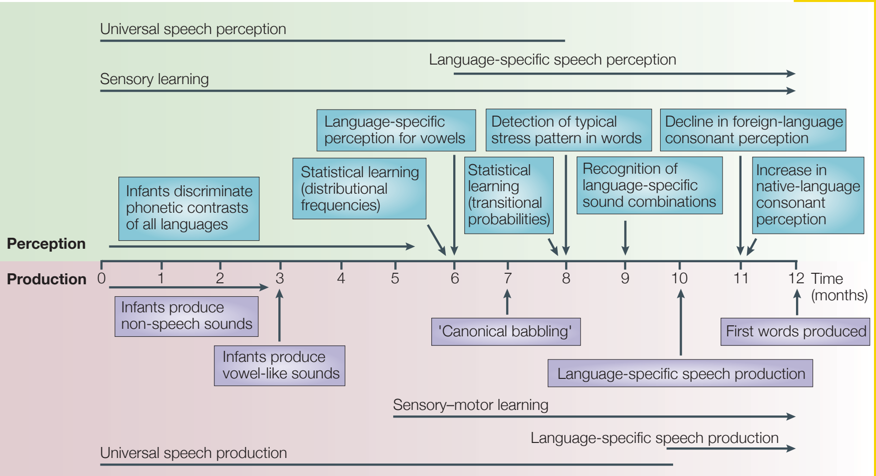
Utrecht University

f.w.adriaans@uu.nl

Today

- Models of human language acquisition
- Methodology basics: training, testing, and evaluation
- Experimentation
  - Simulations on big data sets
  - Artificial language learning experiments (with humans)

Timeline of early language acquisition
(Kuhl, 2004)



Early language acquisition

- Infants between the age of 6 and 12 months learn:
  - Vowels (6 months, Kuhl et al., 1992)
  - Consonants (10 months, Werker and Tees, 1984)
  - Stress patterns (8 months, Jusczyk, Houston, & Newsome, 1999)
  - Phonotactics (9 months, Jusczyk, Luce, & Charles-Luce, 1994)
- Around the same time, the first words are starting to emerge (Bergelson & Swingley, 2012; Fenson et al., 1994)

Infants' phonological acquisition

Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk (1993)

- Infants must learn what sounds and sound combinations are permissible in their native language
- Learning about sounds:
  - English: /θ/ , */x/
  - Dutch: /x/ , */θ/
- Learning about sound combinations: *(phonotactics)*
  - Words in Dutch can begin with /kn/, English words cannot

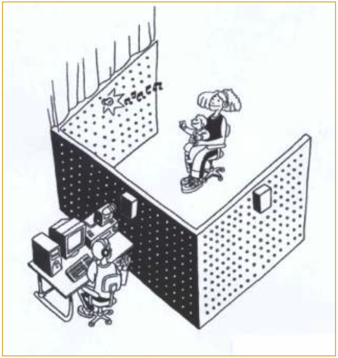
When do infants begin to learn phonotactics?

- Experiment: 24 American infants, 9 months old
- Stimuli: lists of English words + lists of Dutch words
- Phonetic and phonotactic differences:
  - Phonetic realization of /r/
  - word-final /d/ in English, not Dutch
  - /kn/, /zw/ word beginnings in Dutch, not English
  - etc.

Example of word list

English	Dutch
vacate	structuur
avoid	waardig
lengthen	geslacht
brutal	oprecht
jostle	nerveus
trustworthy	efferent
admission	revolutie
thistle	hersteld
exotic	uitsteeksel
lavish	woestijn
abundant	obstructie
jury	eggen
fluctuate	anderzins
usage	verwant
impact	lading
Mean duration = 28.05 s	28.28 s

Headturn preference task



- Dutch from speakers on one side, English from other side
  - Blinking green light in center to get infant's attention
  - Blinking red light to get infant's attention to side speaker
  - List starts playing upon head turn towards speaker
  - Ends when infant looks away for 2 seconds

Results

- American-English infants listened longer to English than to Dutch word lists
- Results indicate that 9-month-old infants are familiar with the phonetic and phonotactic structure of English
- Younger infants (aged 6 months) did not have a preference for English words
- Infants learn basic phonological patterns between 6 and 9 months

The Big Question

- What is the algorithm that drives human language learning?

Computational models of early language acquisition

- Explicit accounts
  - Exact characterizations of the relation between the input and infants' linguistic knowledge
  - Insights into learnability, formal conditions
- Evaluation on real natural language data
  - Models trained on corpora of infant-directed speech
  - Input data realistic in terms of quality

    (although child-directed corpora can have shortcomings in terms of size and transcription quality)

Word segmentation

The Buckeye Corpus

well i work in the accounting department i'm an accounting assistant

so i pretty much um it's not stressful at uh i it's really easy it's not challenging

hardly at all which i yknow it's actually my family's business so um i started working there to help them out but yknow as i'm been working there for a while i want to move up yknow so

Word segmentation

The Buckeye Corpus

welliworkintheaccountingdepartmentimanaccountingassistant

soiprettymuchumitsnotstressfulatuhiitsreallyeasyitsnotchallenging

hardlyatallwhichiyknowitsactuallymyfamilysbusinesssoumistarted workingtheretohelpthemoutbutyknowasimbeenworkingthereforawhile iwanttomoveupyknowso

Word segmentation

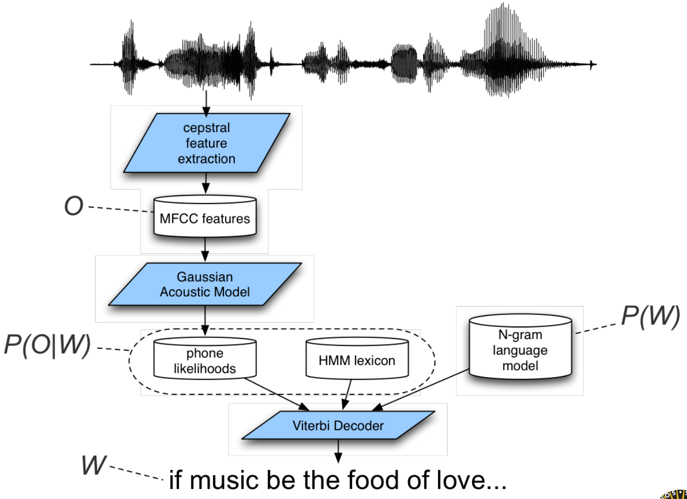
The Buckeye Corpus

w ah aa w er k ih n n ih ah k aw iy ih p aa r tq m ih n aa m ah ah k aw n iy ng ih s ih s t eh n tq

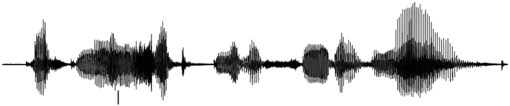
s ow ay p er ih dx iy m ah ch ah m ih t s n aa tq sh r ah s f el ah dx ah ay ih z r ih l iy iy z iy ih t s n aa tq ch ae l ah n jh ih ng

hh aa r l iy eh dx ao w ih ch ay y ih ow ih t s ae k sh l iy m ay f ae m l iy z b ih z n eh s s ow ah m ah s aa r dx ih w er k ih n n eh r dx eh hh eh l p dh eh m aw tq b ah tq y ih n ow ae z ay m b ih n w ah r k ih n n eh r f er w w aa l ay w ah n t t uw m uw v ah p y ih nx ow s ow

Modeling vs. engineering



Modeling vs. engineering



if music be the food of love...

Predicting word boundaries

/ðəlæmpfɛl/

$\Rightarrow$

/ðə.læmp.fɛl/

How do infants detect word boundaries in continuous speech?

(Saffran et al., 1996)

- Statistical cue: transitional probabilities (TP)
  - Given syllable x, what is the probability that it will be followed by syllable y?
- For example:

  pre-tty-ba-by

- $P(pre \rightarrow tty) > P(tty \rightarrow ba)$
- Low probabilities indicate word boundaries

Predicting word boundaries

- Statistical learning (Saffran, Aslin, & Newport, 1996)
  - Infants track co-occurrence probabilities in speech and use these for segmentation.
- A low bigram probability indicates a word boundary
- Let's assume that $P(f|p)$ is low:

/ðəlæm**pf**ɛl/

$\Rightarrow$

/ðəlæm**p.f**ɛl/

Approach

- *Hypothesis:* Statistical learning supports word segmentation.
- The proposal can be formalized by implementing it in a computer program.
    - *N*-gram model
    - The resulting model can be applied to a realistic data set.
    - The performance of the model can be evaluated.

Back to *N*-grams

- *N*-grams formalize the notion of *prediction*
- Word prediction:
  - *I'd like to make a collect...*
- Letter prediction:
  - *weathe...*
- Phoneme prediction:
  - [skwɑ...]

N-grams

- N-grams are probabilistic models that predict the next item from the previous $N - 1$ items.
- Also known in NLP as *language models (LM)*
- Think of N-gram models as small windows that slide over a sentence or word.
- Only N items are visible at a time.

Bigram models

- Bigrams give the smallest possible window.
- Predicting an item based on the previous item:

  $P(call|collect)$

- This is calculated as follows:

  $P(call|collect) = \frac{C(collect\ \ call)}{C(collect)}$

- The same logic applies to $P(f|p)$:

  $P(f|p) = \frac{C(pf)}{C(p)}$

Decomposition into bigrams

- A bigram language model is created by decomposing a corpus into occurrences of two adjacent items:

  I'd like to make a collect call

  $\Rightarrow$

  $[I'd\ like], [like\ to], [to\ make],$

  $[make\ a], [a\ collect], [collect\ call]$

- Or:

  /ðəlæmpfɛl/

  $\Rightarrow$

  /ðə/, /əl/, /læ/, /æm/, /mp/, /pf/, /fɛ/, /ɛl/

Training the model

- The model is trained on unsegmented utterances in training set.
- Training = statistical learning
  - Calculating bigram probabilities
- **Unsupervised learning**

Testing the model

- The model is tested on its ability to predict word boundaries in the test set.
- Output of the model consists of a hypothesized segmentation of the test set.
- We need to define a segmentation algorithm that decides when to insert a boundary.

Segmentation algorithm

- Utterance is parsed using context.
  - *wxyz*-window
- Insert boundary whenever the probability of *xy* is lower than those of *wx* and *yz*

Training and test sets (1)

- Models should be tested on novel data (i.e., data that was not used to train the model)
- This avoids overfitting to the training sample.
- Data needs to be split up into separate training and test sets.
- Common divisions: 80% training set, 20% test set; or 90% training set, 10% test set

Training and test sets (2)

- 10-fold cross-validation further increases the reliability of modeling results
  - Partition the data into 10 subsets.
  - 10 different simulations, each using one subset as test set, and the remaining nine subsets as training set.
- The mean score of 10 runs gives a more reliable estimate of model's performance than a single test set.
- Variance tells you how stable the results are.

Evaluation (1)

- For each bigram, the model predicts either presence or absence of a boundary.
- This prediction is either correct or incorrect.
- Responses fall into 4 categories:

Category	Boundary?	Correct?
True Positive *(hit)*	yes	yes
False Positive *(false alarm)*	yes	no
True Negative *(corr.reject)*	no	yes
False Negative *(miss)*	no	no

Evaluation (2)

- Categorize the following segmentation decisions:
  - The Model: `a.bc.d`
  - The Corpus: `abc.d`

  - The Model: `abc.d`
  - The Corpus: `a.bc.d`

Evaluation (3)

- Hit rate (H):

$$H = \frac{TruePositives}{TruePositives + FalseNegatives}$$

- False alarm rate (F):

$$F = \frac{FalsePositives}{FalsePositives + TrueNegatives}$$

Simulations on Spoken Dutch Corpus (600k words)

- Performance:
  - `Hit rate: 0.6109`
  - `False alarm rate: 0.2242`
- How good is this result?

Simulations on Spoken Dutch Corpus (600k words)

- Example:

Orthography:	*Maar in ieder geval in die film heeft ie wat langer haar.*
Translation:	'But in any case in this film his hair is somewhat longer.'
Transcription:	ma ɪn i fɑl ɪn di fɪlm heft i wɑt lɑŋə har
Continuous:	maɪnifɑlɪndifɪlmheftiwɑtlɑŋəhar
O/E:	ma ɪni fɑl ɪn difɪl mhef tiwɑt lɑŋ əh ar
TP (*N*-gram):	ma ɪni fɑl ɪndi fɪlm he ft iwɑt lɑ ŋə har
StaGe:	ma ɪnifɑlɪndifɪlm hef ti wɑt lɑŋə har

- Results: (Adriaans & Kager, 2010)

Model	Hit rate	False alarm rate	d'
O/E	0.5943	0.2143	1.0301
TP (*N*-gram)	**0.6109**	0.2242	1.0399
StaGe	0.4135	**0.0913**	**1.1142**

Relevance of computational modeling

- Formulating and implementing explicit hypotheses about the learning mechanism
- Testing with natural language corpora
- Comparison of different models gives an indication of how successful a particular learning mechanism is.
- Are the learning mechanisms implemented by computational models cognitively plausible?

Experimentation: artificial language learning (ALL)

- To study learning mechanisms, you need to observe learning in action, using novel stimuli
  - Artificial languages = mini languages with particular structural constraints
- Human participants are familiarized with the AL

  ... and then tested on their knowledge of the AL
- Control over the structural properties of the language allows researcher to zoom in on a particular aspect of learning

Artificial language

(Saffran et al., 1996)

- Four 3-syllable nonsense words

  tupiro, *golabu*, *bidaku*, *padoti*

- Presented as a continuous speech stream

  ...bidakupadotigolabubidaku...

- Word structure is predicted by TPs

  $TP_{within} = 1.0$ da $\rightarrow$ ku

  $TP_{between} = 0.33$ ku $\rightarrow$ pa

Experiment

(Saffran et al., 1996)

- 8-month-old infants
- Familiarization to continuous speech (2 min)
- Test phase:
  - Words (e.g., bidaku)
  - Nonwords (e.g., dakupa)
- Infants listened longer to nonwords.

Statistical learning mechanism

- Infants segmented "words" from the continuous speech stream
- This finding suggests that statistical learning may be an important mechanism to bootstrap into lexical acquisition.
- Statistical learning also tested in various other domains (e.g., phonotactics, artificial syntax)

Infant vs. adult participants

- ALL experiments have been found to produce similar results in infants and adults
- Many experiments addressing the nature of learning mechanisms use adult participants
  - Easier and less time-consuming to test
  - ‘Tryout’ for infant experiment
- Exception: age-related questions
  - e.g., at what age do infants prefer statistical cues over stress cues in segmentation (e.g., Thiessen & Saffran, 2003)

Phonotactic learning from continuous speech

(Adriaans & Kager, 2017)

- Phonotactic learning: learning of novel consonant patterns
- Artificial languages where vowel slots are filled at random
  - No recurring words in the speech stream
- Two training languages, one set of test items
  - Ensures that responses are not solely driven by native language preferences
- 40 adult Dutch participants

Design of the experiment

ABC language		BCA language	
Consonant frames (C_1_C_2_C_3_)	Vowel fillers (*random*)	Consonant frames (C_2_C_3_C_1_)	Vowel fillers (*random*)
p_d_g_	[_a,_e,_o,_i,_u,_y]	d_g_p_	[_a,_e,_o,_i,_u,_y]
p_z_k_		z_k_p_	
t_b_x_		b_x_t_	
t_z_g_		z_g_t_	
s_b_k_		b_k_s_	
s_d_x_		d_x_s_	

Note. A = voiceless obstruents, B = voiced obstruents, C = dorsal obstruents.

- TP(*within*) = 0.5; TP(*between*) = 0.33
  - ABC condition: ABC.ABC.ABC.ABC.ABC.ABC...
  - BCA condition: A.BCA.BCA.BCA.BCA.BCA.BC...
- Test trials: pairs of novel items (ABC vs. BCA)

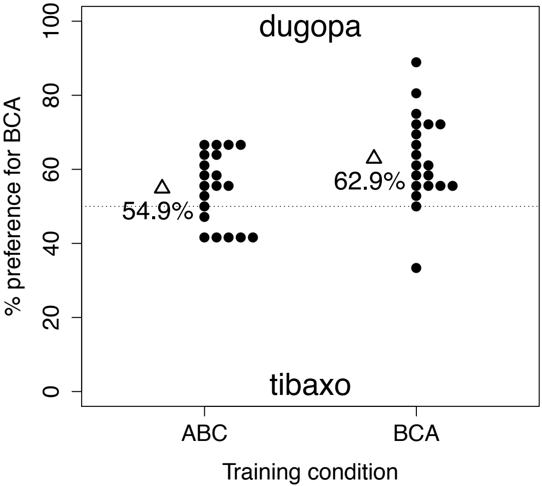
Let's give it a try...

2-alternative forced-choice task

- Which of the following two words sounds more like the language you just heard?

/tibaxo/ - /dugopa/

Experiment results



- ABC condition: *ns*, BCA condition: ***
- Significant learning effect ($\beta = 0.3628, p = 0.0168$)

Want to learn more?

- Reading: (required for exam!)
  - Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, 62(3), 311-331.
- Experimental procedures with infants: `https://www.youtube.com/watch?v=EFlxiflDk_o`
- **Cognitive Modeling** (block 2)
- **Experimentation in Psychology and Linguistics** (block 3)