Can AI help the study of language development?

Emmanuel Dupoux

Ecole des Hautes Etudes en Sciences Sociales

0. Introduction

- 2 deep scientific puzzles
- 4 traditional approaches
- The reverse engineering approach

Two deep scientific puzzles

1. Logical problem (bootstrapping)
   - learnability: from finite input to infinite competence

- The input to the learner is finite (and small)
- The adult competence is (almost) infinite

→ ***how?***

Two deep scientific puzzles

1. Logical problem (bootstrapping)
   – learnability: from finite input to infinite competence

- The input to the learner is finite (and small)
- The adult competence is (almost) infinite

→ ***how?***

The longest sentence in French (856 words, Proust, A la recherche du temps perdu, Vol 4)

Sans honneur que précaire, sans liberté que provisoire, [..] et de façon qu'à eux-mêmes il ne leur paraisse pas un vice.

Two deep scientific puzzles

1. Logical problem (bootstrapping)
   – learnability: from finite input to infinite competence

- The input to the learner is finite (and small)
- The adult competence is (almost) infinite

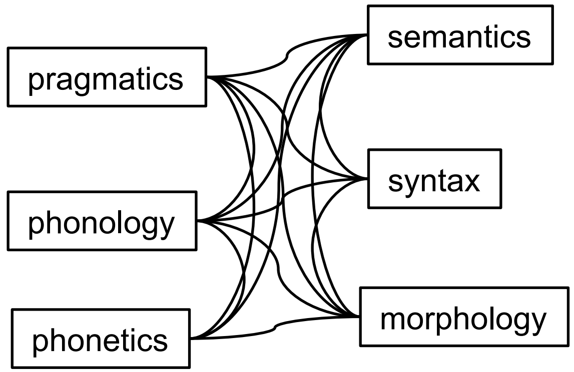
→ ***how?***

The longest sentence in French (856 words, Proust, A la recherche du temps perdu, Vol 4)
Sans honneur que précaire, sans liberté que provisoire, [..] et de façon qu'à eux-mêmes il ne leur paraisse pas un vice.

A longer sentence:
Proust wrote « Sans honneur que précaire, sans liberté que provisoire, [..] et de façon qu'à eux-mêmes il ne leur paraisse pas un vice. »

Two deep scientific puzzles

1. Logical problem (bootstrapping)
   – learnability: from finite input to infinite competence
   – co-dependency: chicken vs eggs



- Infants have a Language Acquisition Device (Chomsky, 1965)
(an innate machine for learning any language)

-However, learning one component requires many others
(e.g. learning the sounds requires the words and vice versa)

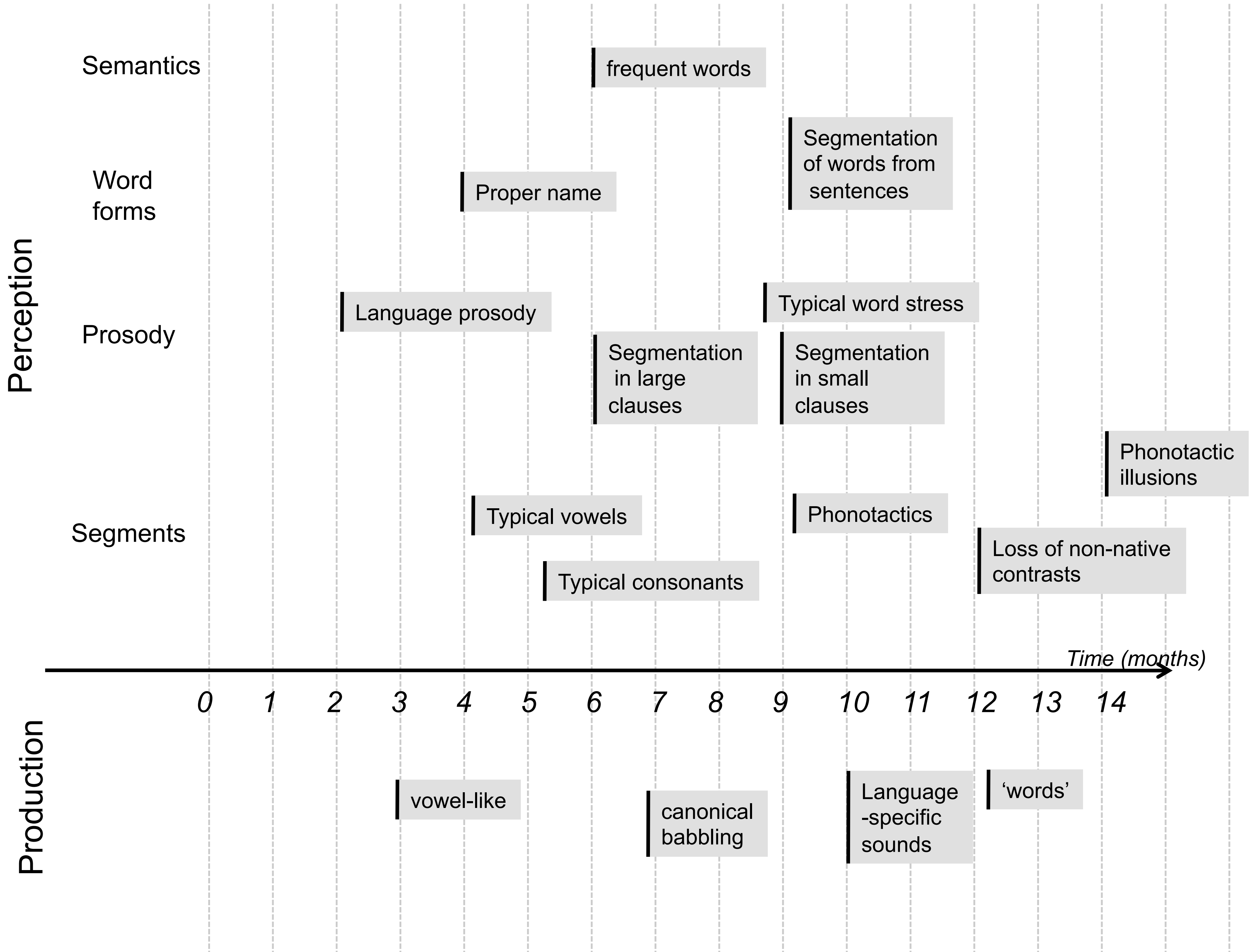
→ ***how?***

Two deep scientific puzzles

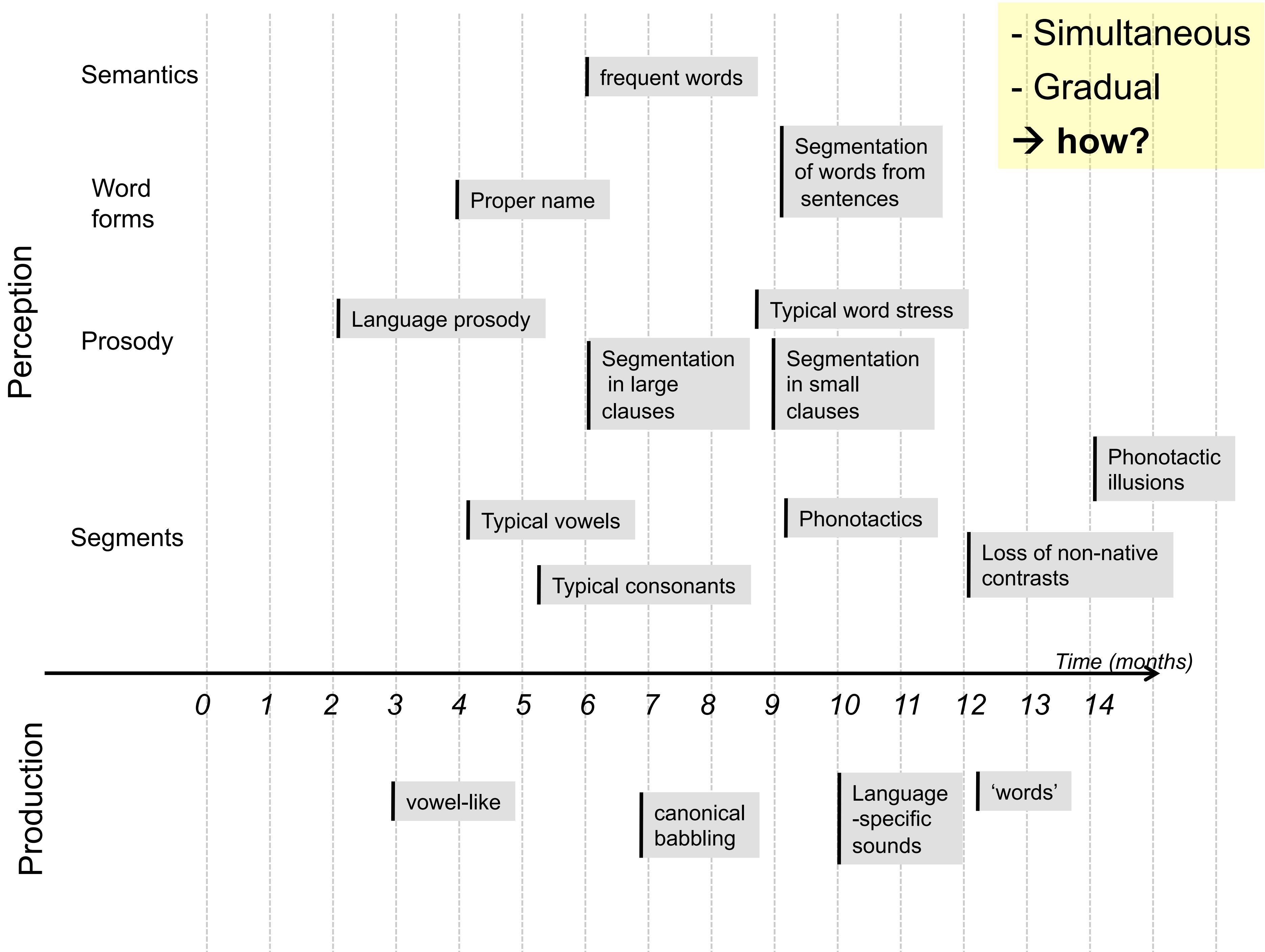
1. Logical problem
   (bootstrapping)
   - learnability: from finite input to infinite competence
   - co-dependency: chicken vs eggs

2. Explanatory problem
   - learning trajectories: simultaneous and gradual
   - resilience: nonlinear relationships between inputs and outcomes

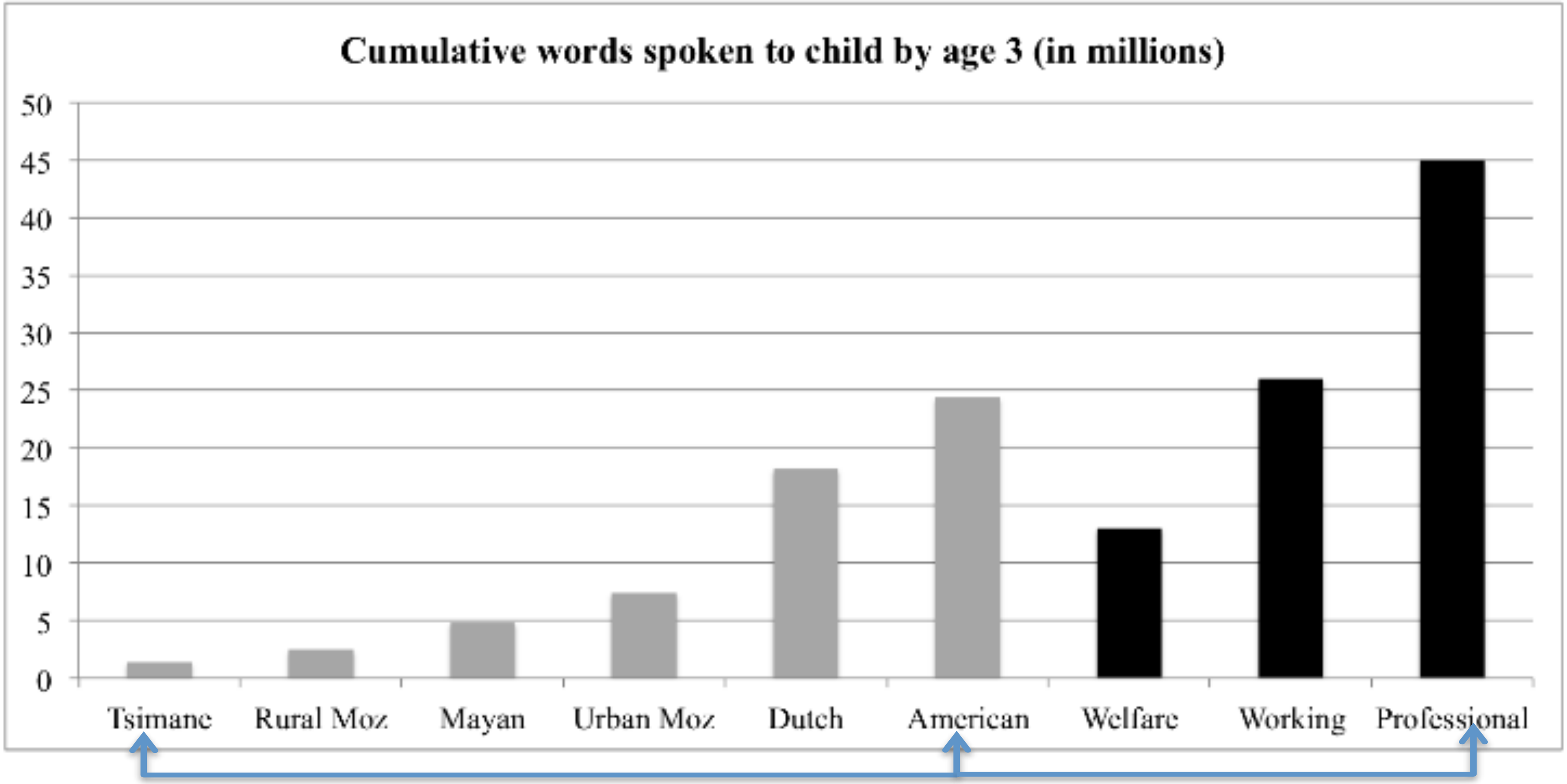
Learning trajectories



Learning trajectories



Resilience



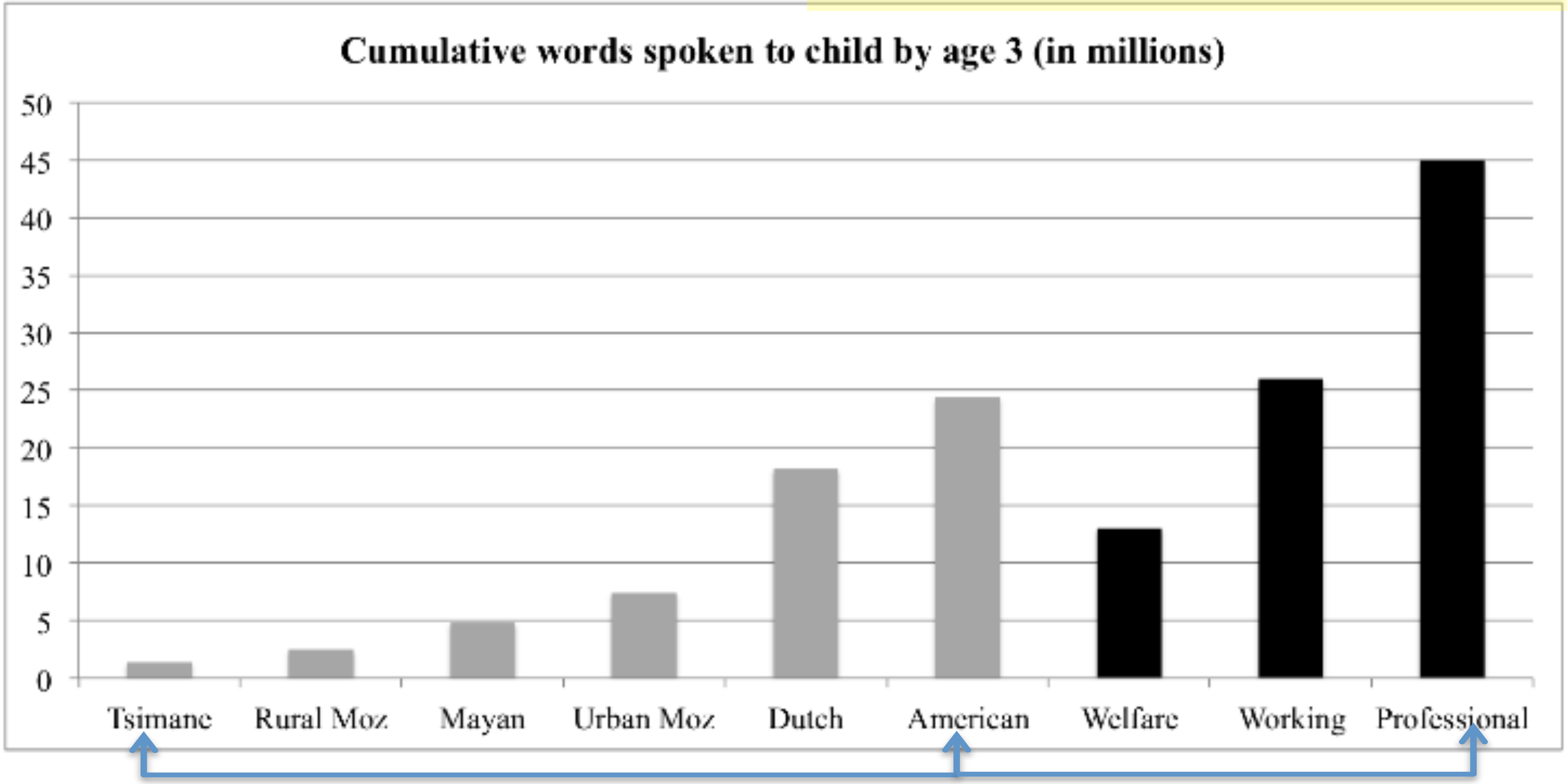
66h/y 800h/y 1200h/y

Cristia et al, 2017

Resilience

- large differences in amout of child directed input (up to 2000%)
- much smaller differences in differences in outcome (language landmarks: stable)

→ **how?**



66h/y

800h/y

1200h/y

Cristia et al, 2017

Four traditional approaches

1. Psycholinguistics (conceptual)
2. Psycholinguistics (experimental)
3. Formal linguistics
4. Developmental AI

1. Psycholinguistics

- Conceptual frameworks
  - Bootstrapping problem
    - semantic bootstrapping (Pinker, 1984)
    - syntactic bootstrapping (Gleitman, 1990)
    - prosodic bootstrapping (Morgan & Demuth, 1996)

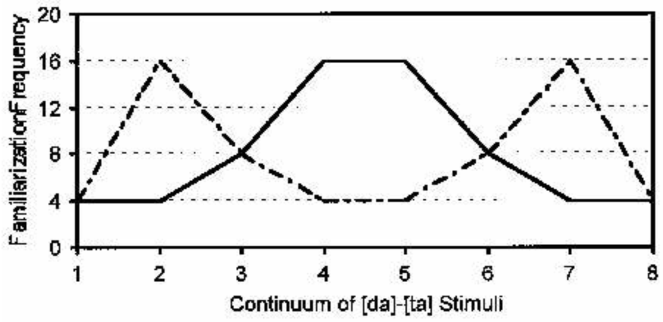
→ *do they work? can they be implemented?*

  - Explanatiory problem
    - Knowledge driven LAD (Lidz & Gagliardi 2015)
    - WRAPSA (Juczyk, 1997)
    - PRIMIR (Werker & Curtin, 2005)
    - Competition Model (Bates &MacWhinney, 1987)
    - Usage Based Theory (Tomasello, 2003)

→ *can they be refuted? distinguished?*

2. Psycholinguistics (experimental)

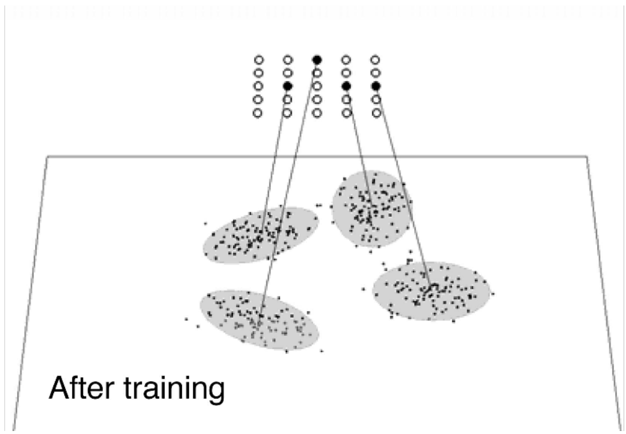
- Artificial language learning
  - distributional learning
    - (Maye, Werker & Gerken, 2002)



Maye and Gerken (2002)

→ *does it scale up to realistic input?*

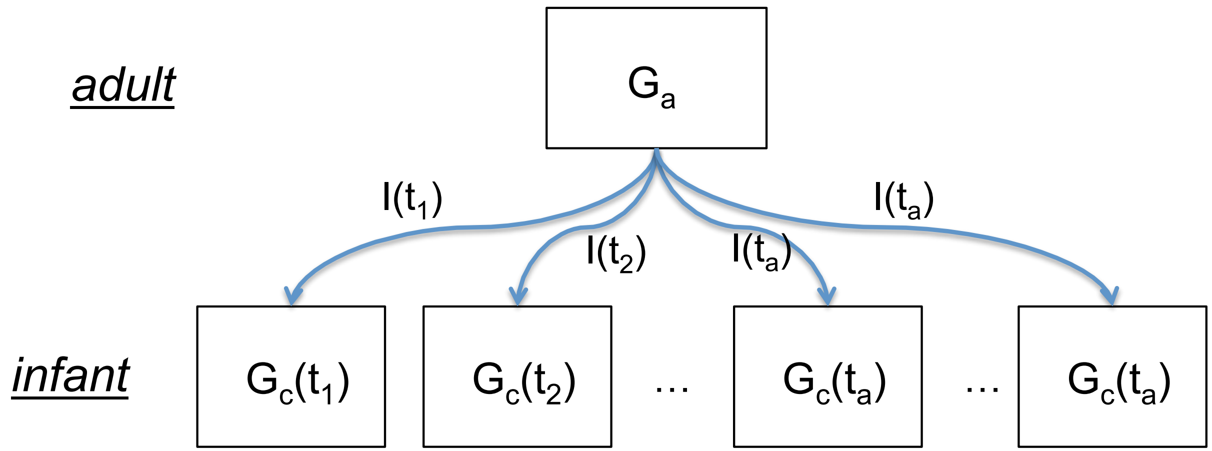
  - rule learning (ABB vs ABC, Markus, et al.)



Vallabha, et al (2007

→ *does this help language learning?*

3. Formal learning theories/linguistics

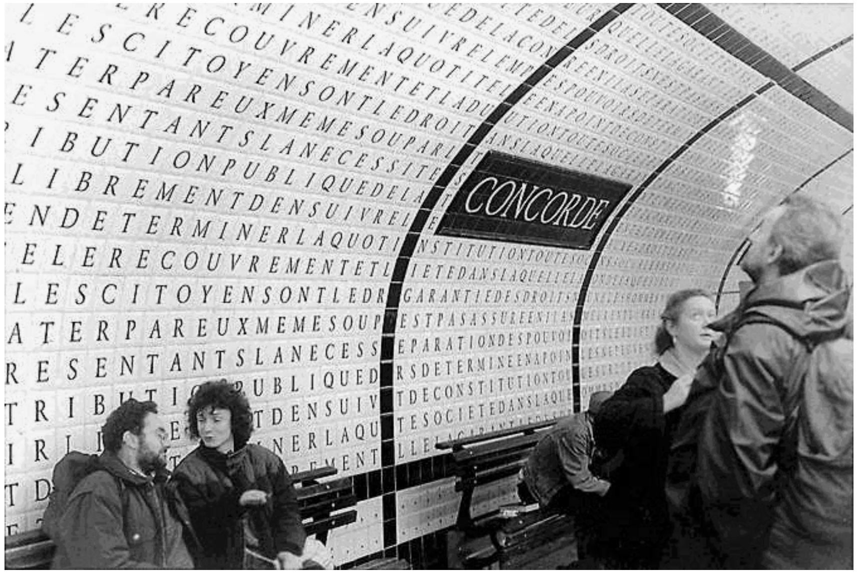


- Learnability in the limit: Gold (1967)
- Phonological grammar: Tesar & Smolensky (1998), Dresher & Kaye (1990); etc
- Syntax: see Clark & Lappin (2011)

→ *are the hypotheses valid in real life?*

4. Developmental Artificial Intelligence

- language learning=learning a compact representation for the input (Kelley, 1967, de Marcken, 1996)
  - e.g. word segmentation
- language learning=learning to translate between surface input to underlying concepts (Siklossy, 1968; Siskind, 1996)
  - e.g. word learning
- language learning=learning to communicate (Bruner 1975)
  - e.g. word emergence



https://www.davidphenry.com/Paris/paris090_fr.htm

word segmentation

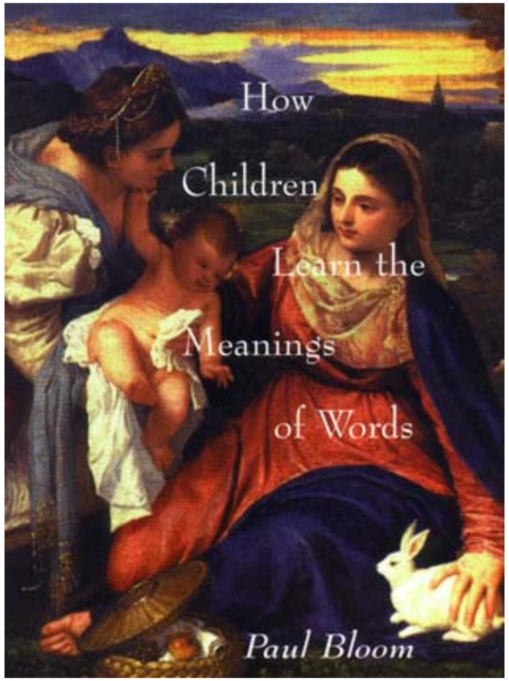
- <u>Minimal description length</u>
  minimize the size of the lexicon plus corpus description (Brent & Cartwright,1996)

SEGMENTATION	REPRESENTATION		LENGTH
	LEXICON	DERIVATION	(Objective)
do you see [thekitty] see [thekitty] do you like [thekitty]	1 do 2 [thekitty] 3 you 4 like 5 see	1 3 5 [2] 5 [2] 1 3 4 [2]	25+10=35
do you see the kitty see the kitty do you like the kitty	1 do 2 the 3 you 4 like 5 see 6 kitty	1 3 5 2 6 5 2 6 1 3 4 2 6	26+13=39
do [yousee] the kitty see the kitty do you like the kitty	1 do 2 the 3 you 4 like 5 see 6 kitty [7] [yousee]	1 [7] 2 6 5 2 6 1 3 4 2 6	33+12=45

- <u>Non Parametric Bayesian (Chinese Restaurant process)</u>
  maximize the probability that the corpus is generated by a lexicon (Goldwater, 2007; Johnson, Griffith Goldwater, 2007)

4. Developmental Artificial Intelligence

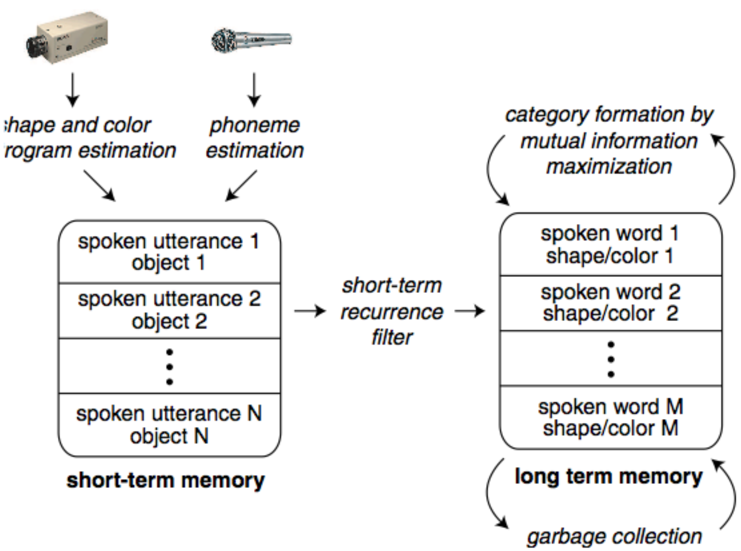
- language learning=learning a compact representation for the input (Kelley, 1967, de Marcken, 1996)
  - e.g. word segmentation
- language learning=learning to translate between surface input to underlying concepts (Siklossy, 1968; Siskind, 1996)
  - e.g. word learning
- language learning=learning to communicate (Bruner 1975)
  - e.g. word emergence



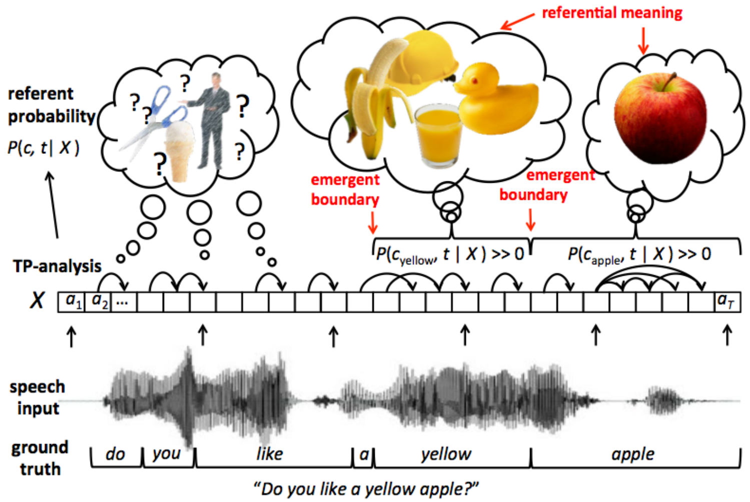
Bloom (2000), MIT Press

word learning

<u>Cross situational learning</u>
learning the correspondance between words and meaning across many examples



Roy & Pentland, 2002



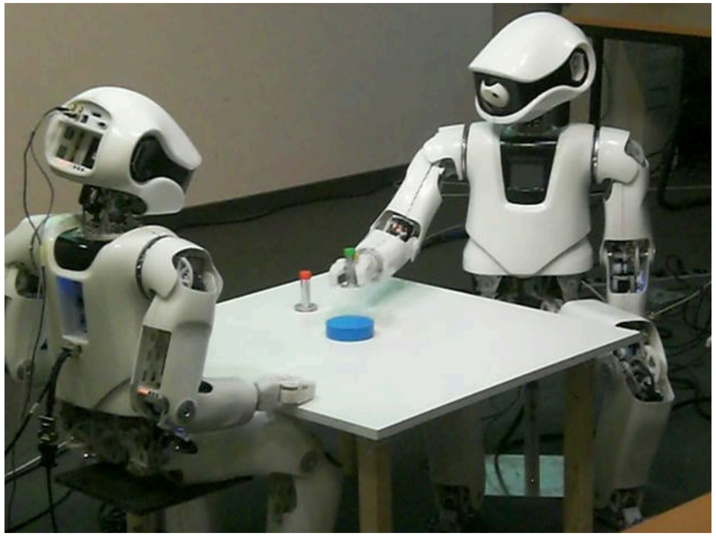
Rasanen & Rasilo, 2005

see also Siskind 1996; Kwiatkowski et al 2012

4. Developmental Artificial Intelligence

- language learning=learning a compact representation for the input (Kelley, 1967, de Marcken, 1996)
  - e.g. word segmentation
- language learning=learning to translate between surface input to underlying concepts (Siklossy, 1968; Siskind, 1996)
  - e.g. word learning
- language learning=learning to communicate (Bruner 1975)
  - e.g. word emergence

word emergence



https://ikw.uni-osnabrueck.de/~neurokybernetik/projects/alear.html

<u>Grounded communication</u>
language emerges as a communication protocol to help solving a particular task

Talking heads (Steels et al 2001)

Mordatch & Abbeel 2017

see also Foerster et al., 2016; Sukhbaatar et al., 2016; Lazaridou et al., 2016; Havrilov & Titov, 2017

4. Developmental Artificial Intelligence

- language learning=learning a compact representation for the input (Kelley, 1967, de Marcken, 1996)
  - e.g. word segmentation
- language learning=learning to translate between surface input to underlying concepts (Siklossy, 1968; Siskind, 1996)
  - e.g. word learning
- language learning=learning to communicate (Bruner 1975)
  - e.g. word emergence

→ *are the hypotheses and results compatible with infant data? Do they scale with real data?*

In brief

	Effective Model	Realistic Data	Human/Model Comparison
Conceptual Frameworks	No (verbal)	Yes	No (verbal)
Artficial Language Learning	Yes (but not scalable)	No	Yes
Formal Linguistics	Existence proof	Idealized	In the limit
Developmental AI	Yes	Simplified	Qualitative / In the limit

	Effective Model	Realistic Data	Human/Model Comparison
Conceptual Frameworks	No (verbal)	Yes	No (verbal)
Artficial Language Learning	Yes (but not scalable)	No	Yes
Formal Linguistics	Existence proof	Idealized	In the limit
Developmental AI	Yes	Simplified	Qualitative / In the limit
Reverse Engineering	Yes	Yes	Yes

→*the reverse engineering approach*

(or, new AI to the rescue)

Roadmap

Reverse engineering: *construct a scalable model that discover phonetic categories like infants do using real data.*

I. Why real data?

II. Scalable Models

III. Testing predictions

Dupoux, Cognition, 2018

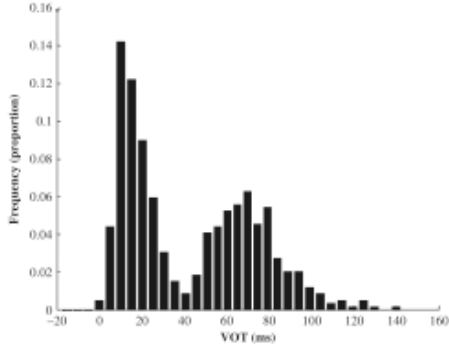
I. Why using real data?

or: why simplification is not always a good idea

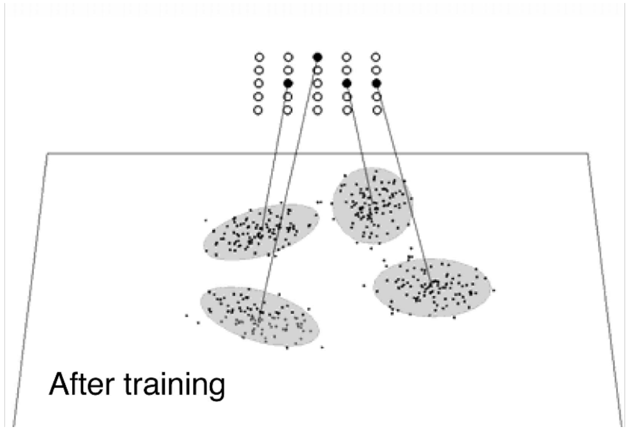
1. Variability is part of the problem

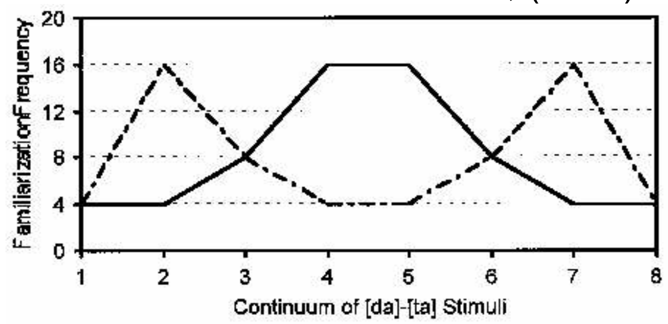
- Simplification is important in science: splitting complicated problems into simpler one
- But… simplifying changes the learning problem

- e.g. Phoneme learning

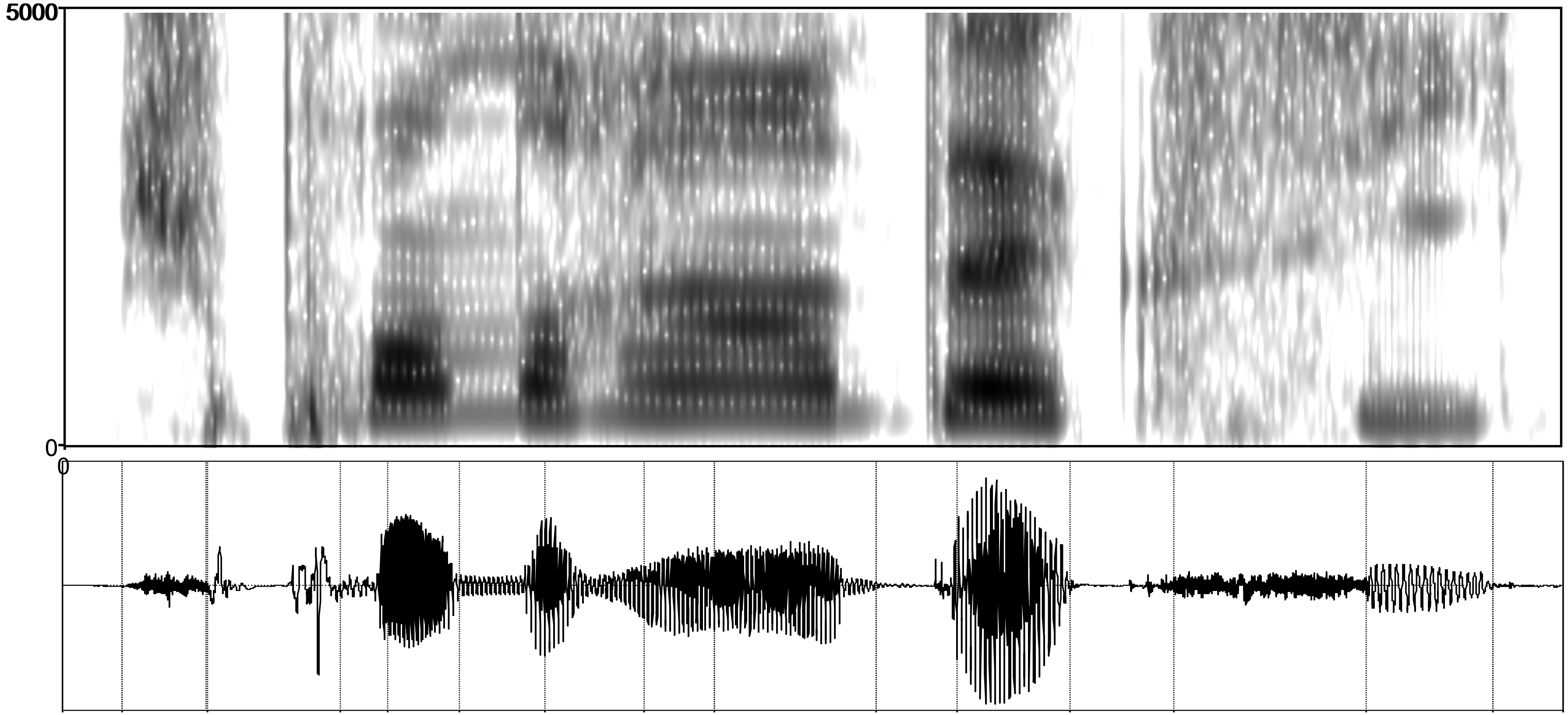


Lisker & Ambramson (1964)
Allen & Miller, (1999)

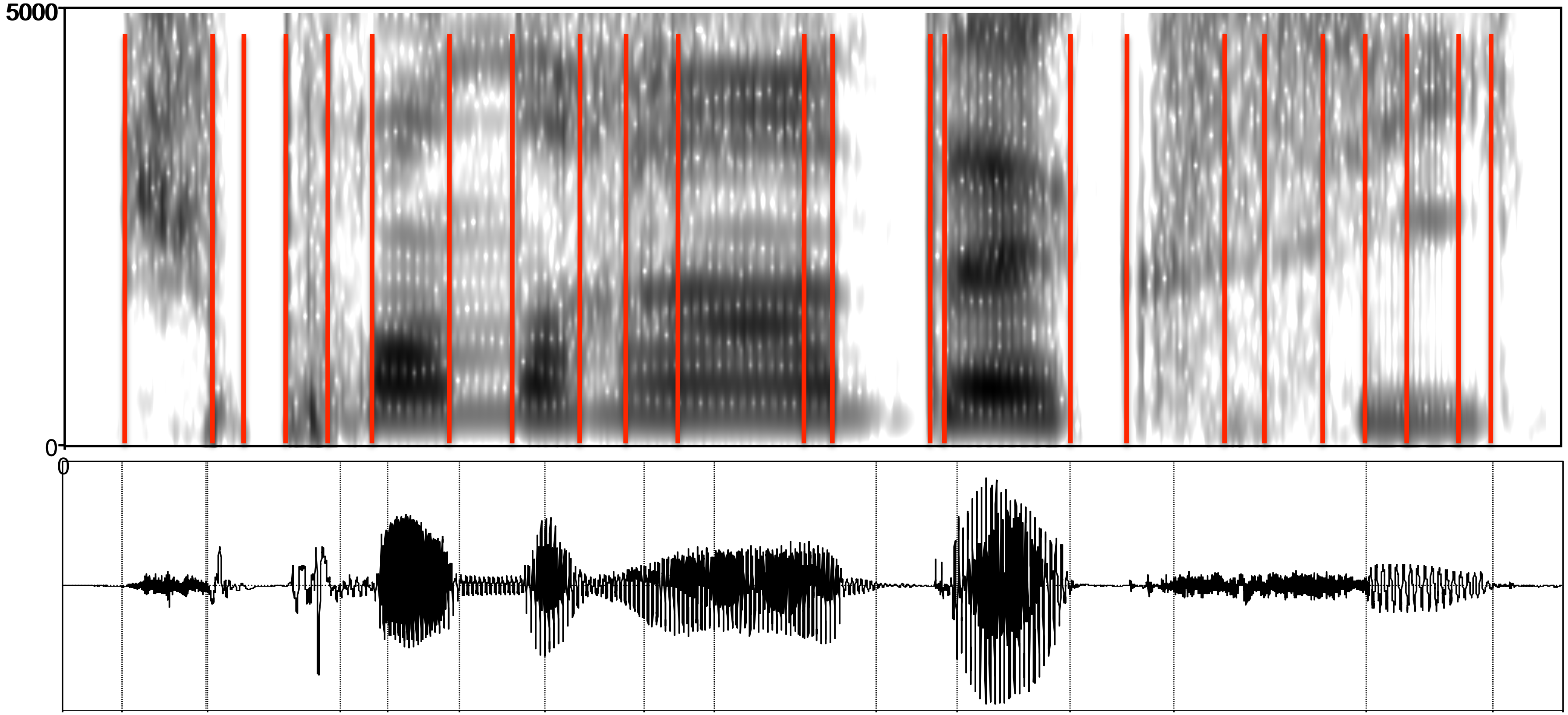




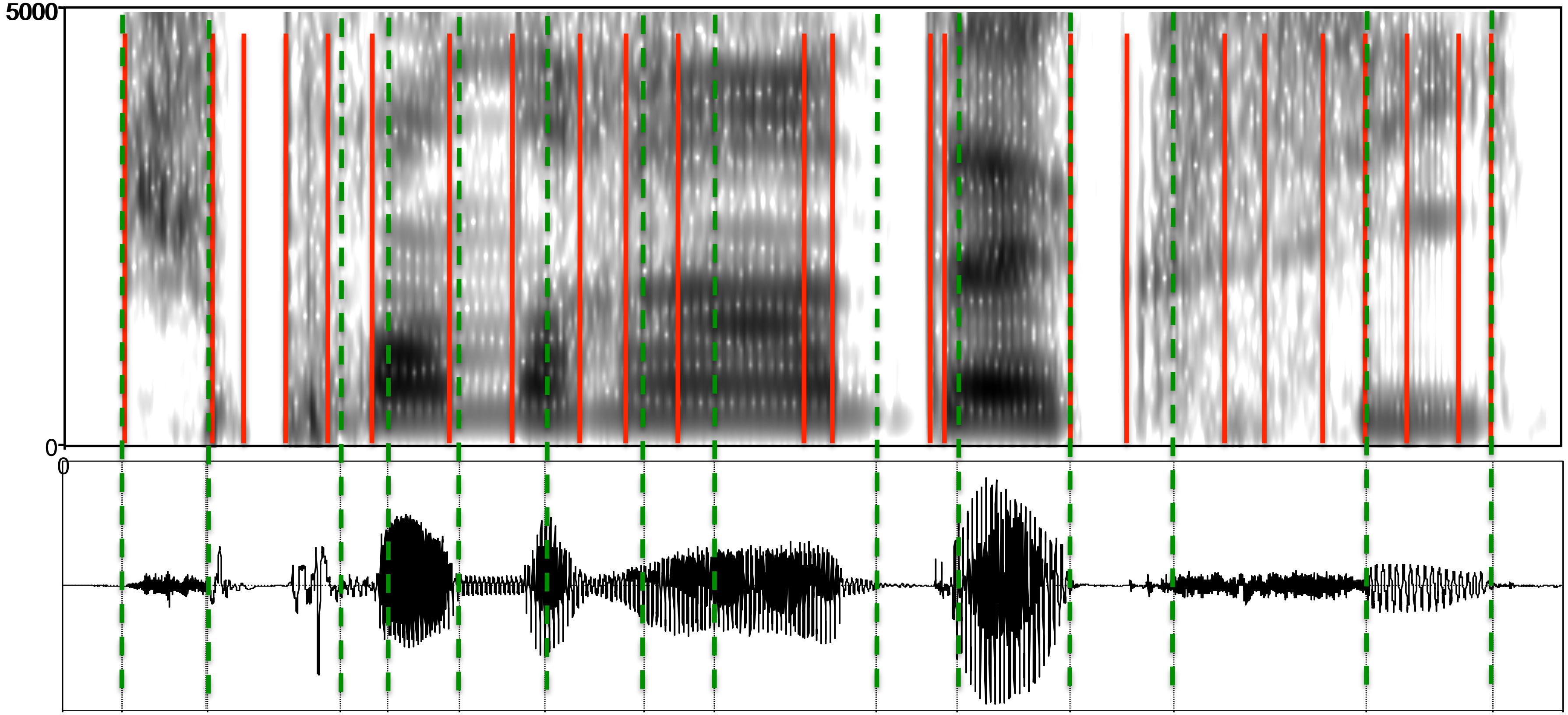
study	stimuli	Nb	measures	algo
de Boer & Kuhl, (2003)	3 CVC words, 10 speakers	3	F1, F2	EM(N known)
Vallabha et al. (2007)	fake from mono&bisyll nonwords, 20 Eng and 10 Jap speakers	4	F1, F2, duration	OME, TOME
McMurray et al. (2009)	english stop-V syllables	2	VOT	GMM+ MLE + competitive learning
Lake et al (2009)	fake da vs ta	2	"VOT"	GMM+OME
Lake et al (2009)	fake vowels	3	F1 & F2	GMM+OME
Toscano & McMurray (2010)	fake stops	2	VOT and v duration	GMM+OME
Kouki et al. (2010)	continuous speech, 12 speakers	5	MFCC	SOM



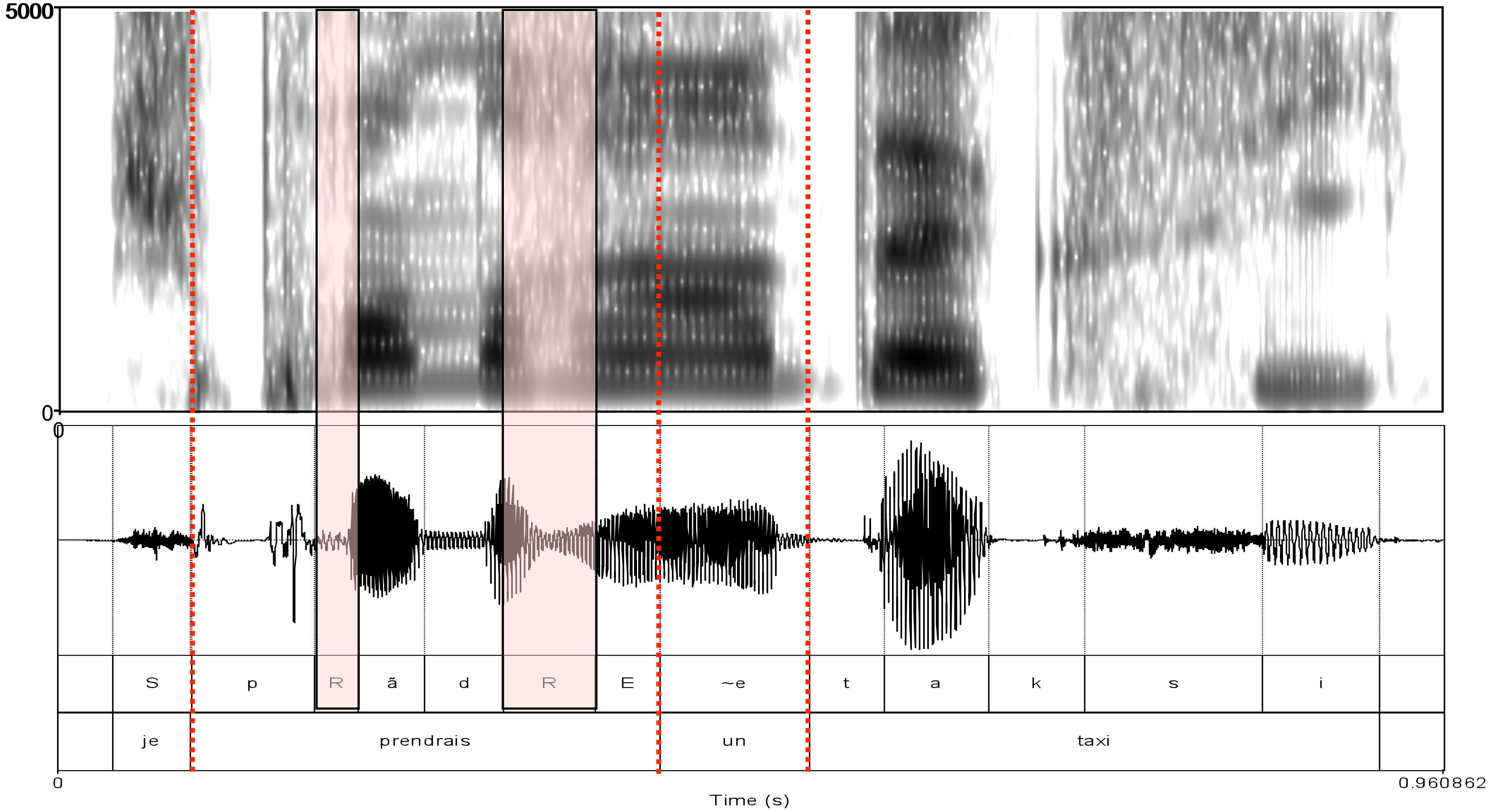
→ *Does this scale up?*



→ *Does this scale up?*

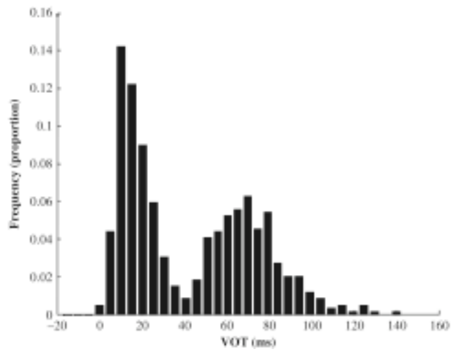


→ *Does this scale up?*



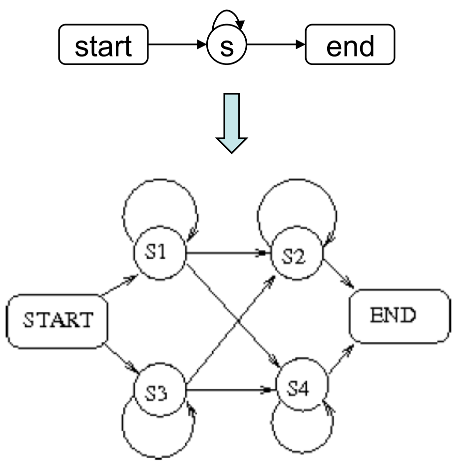
→*Does this scale up?*

not really; phonemes are not well separated, discrete entities



Lisker & Ambramson (1964)
Allen & Miller, (1999)

Phoneme learning with real speech



State seq	Allophones
11,28,32	[V]-**t**+[e|a|o]
15,17,2	[g|k]-**[u|o]**+[⋆]
3,17,2	[k|t|g|d]-**a**+[k|t|g|d]
31,5,13,5	[V]-**[s|sj|sy]**+[V]
17,2,31,11	[g|t|k|d]-**[a|o]**+[t|k]
3,30,22,34	[⋆]-**a**
6 24 8 15 22	[⋆]-o
22 35 11 28 32	[N|i|u|o]-**[t|d]**+[e|o|i]
4 17 24 2 31	[s|sy|z]-**o**+[t|d], [t|d]-**o**+[s|sy|z]

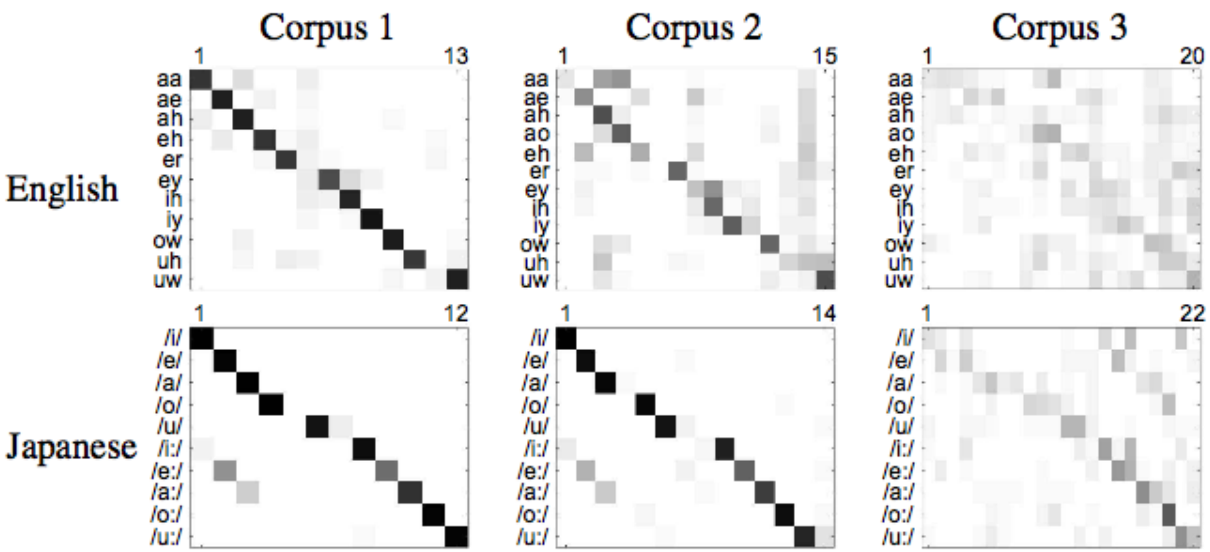
what is learned is pseudo phones:
→too small
→too context dependant
→too talker dependant

Varadarajan, Kudanpur & Dupoux. (2008)

- e.g. Phoneme learning with the help of the lexicon

More realistic corpus

Word forms	phonemic (dictionary)	phonetic (human annotated)	phonetic (human annotated)
Consonants	gold	gold	gold
Vowels	Resampled F1 and F2	Resampled F1 and F2	Measured F1 and F2



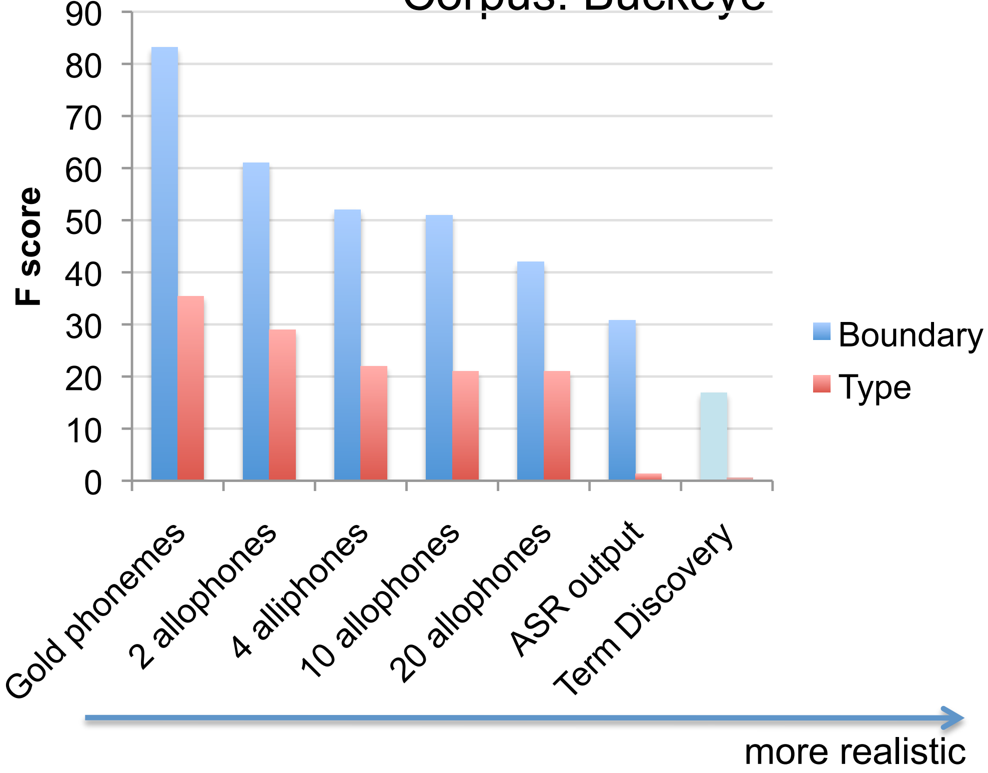
Worse results

Antetomaso, Miyazawa, Feldman, Elsner, Hitczenko, & Mazuka (2016)

- e.g. word learning & segmentation
  - from symbolic input: *Findingwordsincontinuousspeech*.
    - local probabilities: Saffran et al
    - lexical based: Brent et al Goldwater et al.

    -> state of the art: ~ 80% correct (in English)
  - from speech:
    - 'fake data':
      - ASR contextual allophones
      - ASR output
    - real data
      - Term Discovery (Jansen)

Model: Unigram Non Parametric Bayes
Corpus: Buckeye



From Fourtassi & Dupoux (2014); Ludusan et al. (2014)

→ *using simplified data changes the nature of the learning problem*

2. Other forms of simplification

- Mode of presentation: the way in which infants are presented with language samples.
  – pedagogic curriculum: from simple to complex
  – neutral curriculum: random sample
  – adversarial curriculum: designed to make infants fail
  → *mode of presentation matters for algorithms (Gold, 1967; Angluin 1988)*
  → *Are parents pedagogic in all cultures?*
- Data selection: linguistic vs non linguistic channels
  – many algorithms run on ‘cleaned’ data (and fail on raw data)
  → *but what counts as speech depend on the language (eg, sign vs oral; clicks; creaky voice, etc)*
  → *some nonspeech hurt (noise), other help (context)*

In brief

- Simplification is useful in science, but
  - learnability is extremely dependent on input
  - changing the input means addressing a different learning problem
- Therefore, to answer the two puzzles, we have to use ***realistic corpora***
- Now it is possible to do so (personal big data):

home recording (LENA device)

dense multimedia recording (Roy 2009)

life logging

ACLEW (ANR-NSF)
BabyCloud

II. What kind of algorithms?

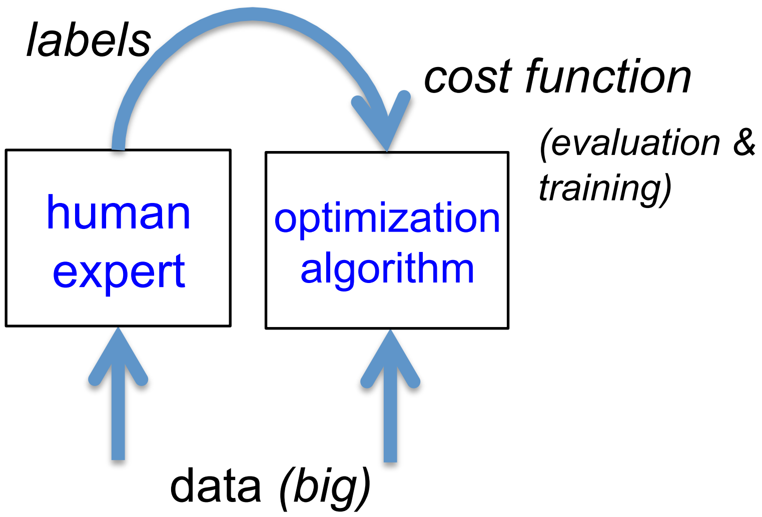
Popular AI algorithms needs a lot of (supervised) data

To be relevant, machine learning has to go data efficient and unsupervised

Standard Machine Learning

human supervision:

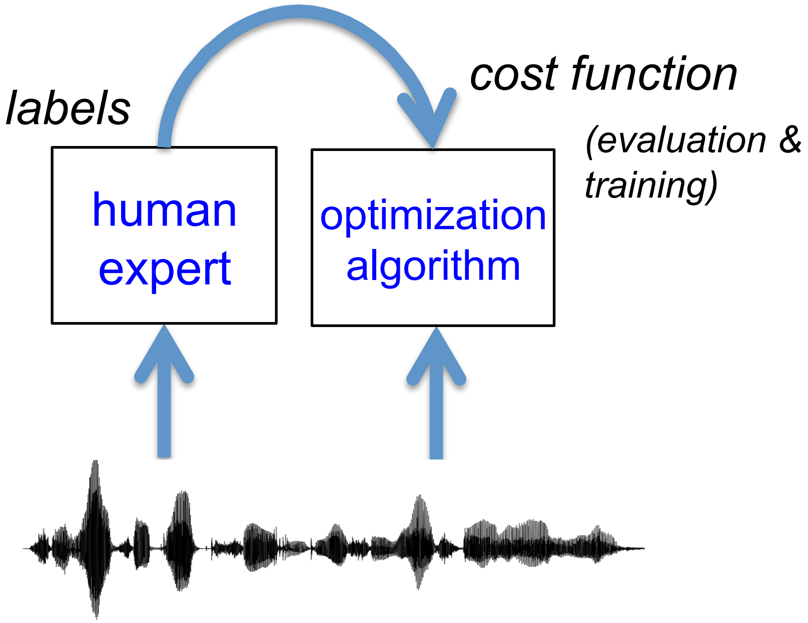
- *strong (unambiguous)*
- *dense (high bitrate)*
- *mono directional*



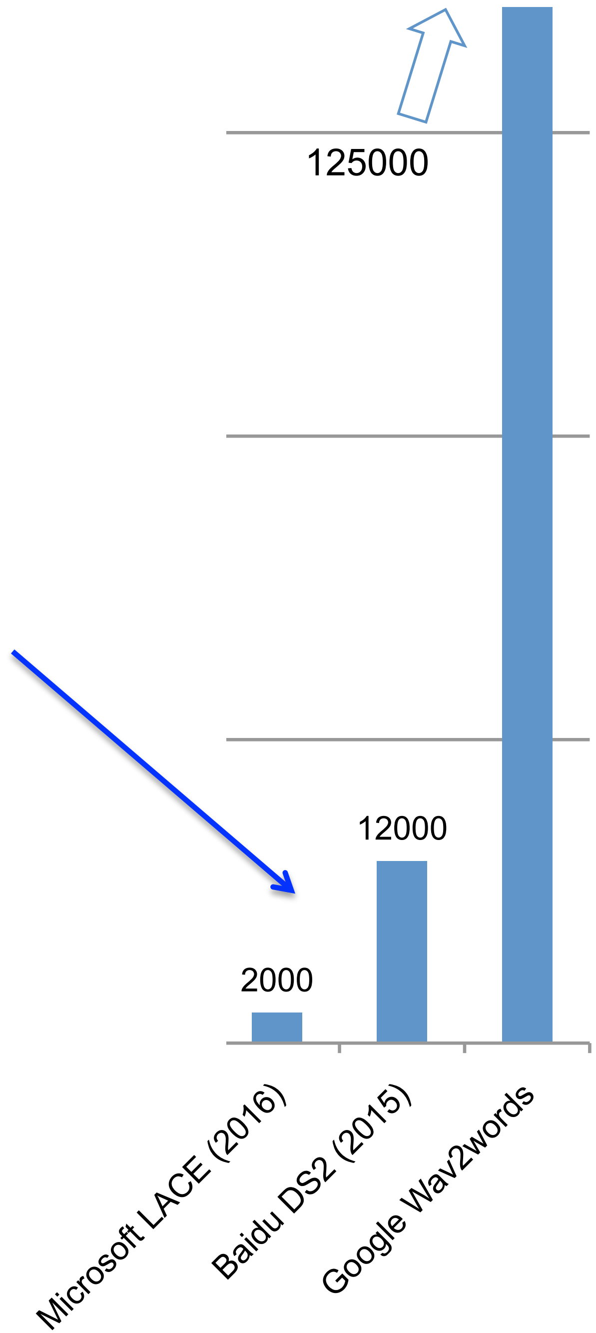
The data addiction problem

<u>End-to-end ASR</u>

« She had your dark suit in greasy wash water all year »

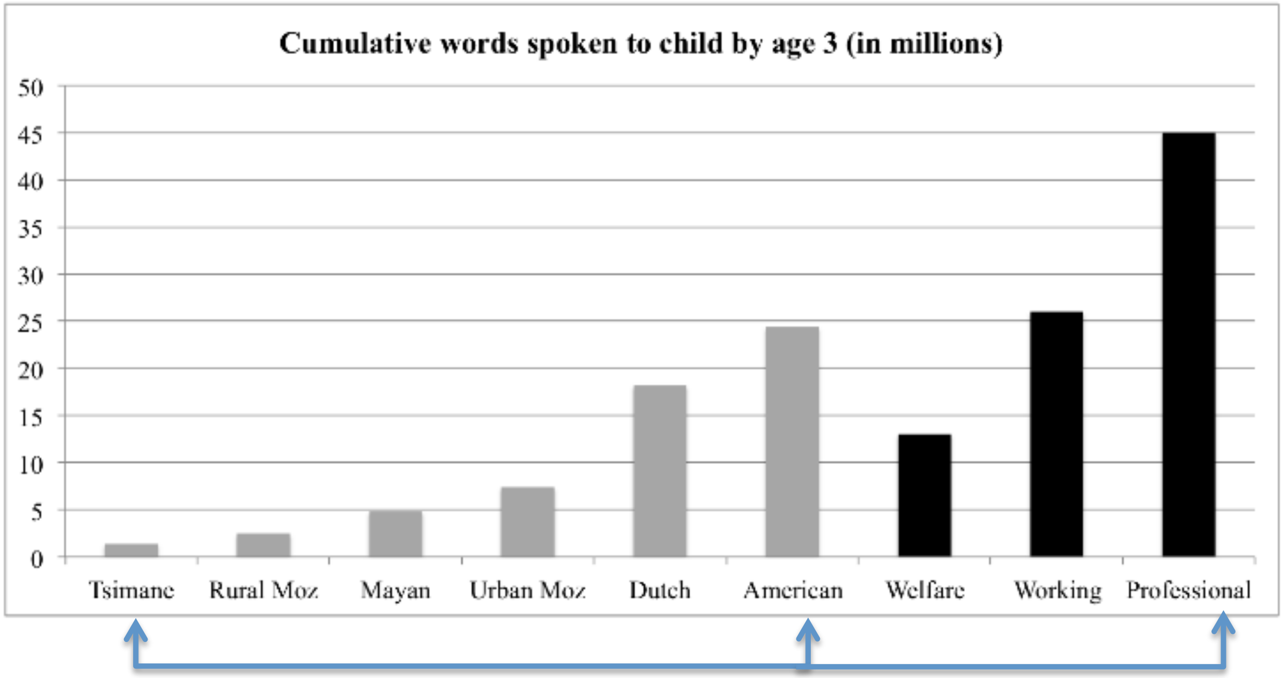


+ 1000000000 words of text for language modeling! (10000 books)



The data addiction problem

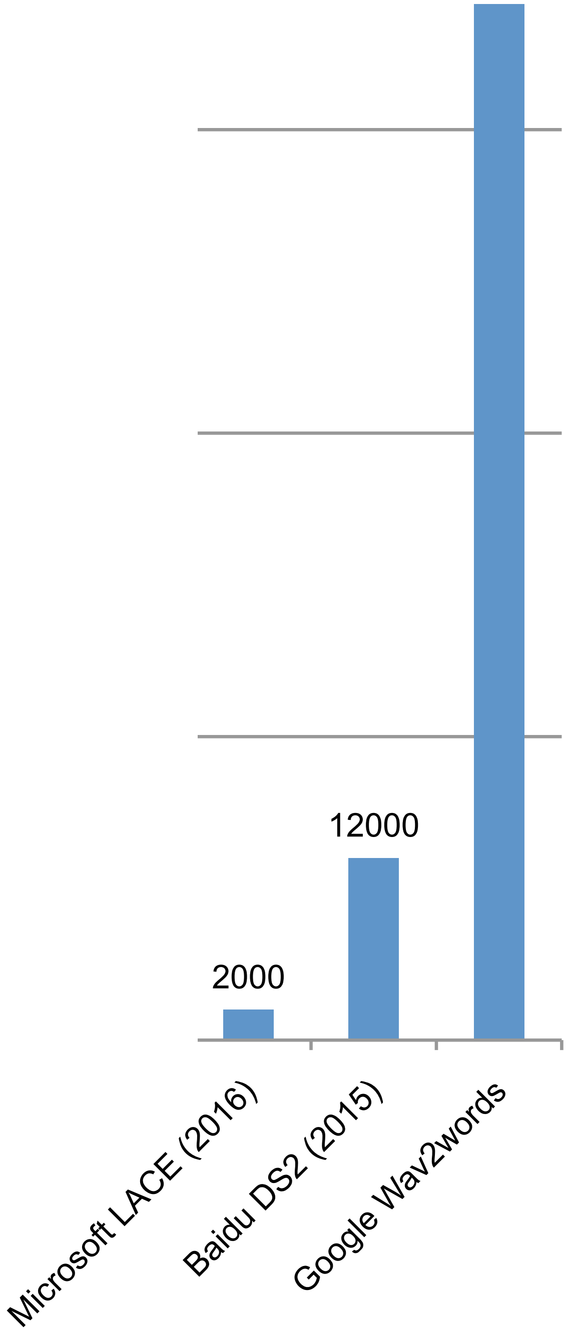
→ infants require less data, and no labels!



66h/y

800h/y

1200h/y

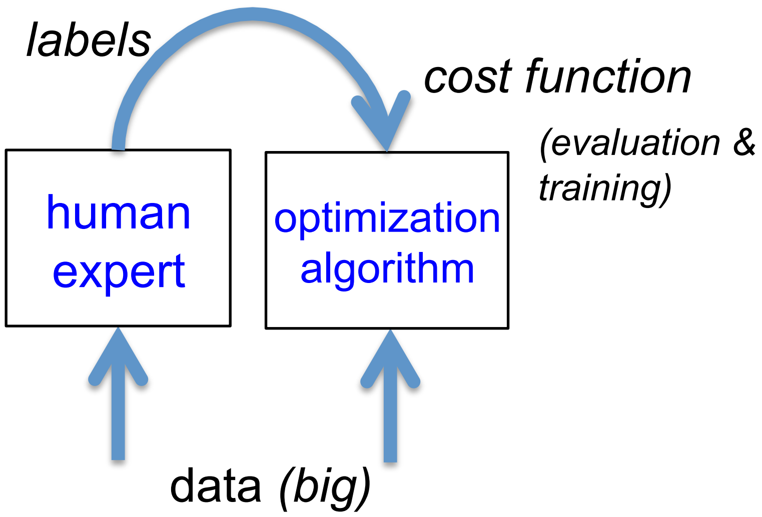


Cristia et al, 2017

‘Cognitive Machine Learning’

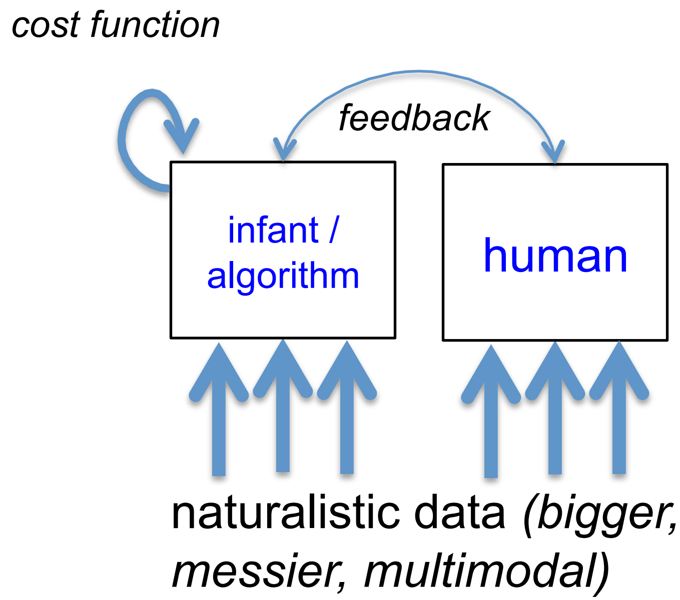
Standard Machine Learning

human supervision:
- *strong (unambiguous)*
- *dense (high bitrate)*
- *mono directional*



Human-like Machine Learning

human supervision:
- *weak (ambiguous)*
- *sparse (low bitrate)*
- *bi-directional*

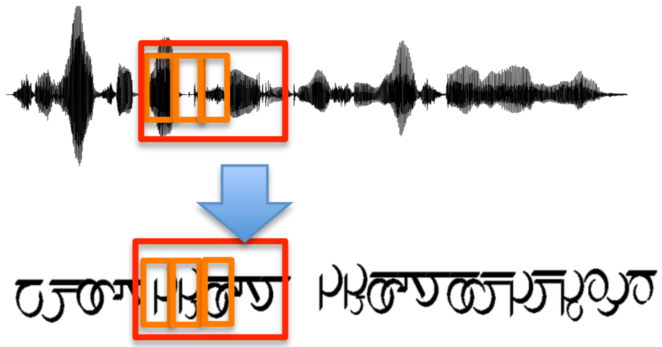


A new kind of challenge for AI

- The ‘ghost’ linguist conundrum:
  - you arrive in a foreign country
  - you want to construct a grammar for the language (list of phonemes, dictionary)
  - you cannot talk to the native, just listen and watch

→ *How would you do?*

The zero resource challenge(s)



- In an unknown language, from raw speech discover:
  - invariant subword units (Track 1)
  - words/terms (Track 2)
- ZR15 (Interspeech 2015)
  - English (casual, 12 speakers, 5 hours)
  - Xitsonga (read, 24 speakers, 2.5 hours)
- ZR17 (ASRU 2017)
  - 3 dev languages: English, French Mandarin (12-69 speakers, 2.5-45h)
  - 2 surprise languages: German, Wolof (24-30 spakers, 10-25h)
- JSALT 2017 Spoken Rosetta Stone Workshop, CMU

- *Aalto University, Finland*
- *KTH, Sweden*
- *University of Edinburgh, UK*
- *U. Tilburg, Netherlands*
- *Ecole Normale Sup, France*
- *Instituto Italiano di Tecnologia, Italy*
- *IIT Hyderabad, India*
- *Stellenbosch, U. South Africa*
- *National Taiwan U., Taiwan*
- *A*STAR, Singapore*
- *NAIST, Japan*
- *Carnegie Mellon, USA*
- *U. Chicago, USA*
- *Stanford Univ, USA*
- *Johns Hopkins, USA*
- *MIT, USA*

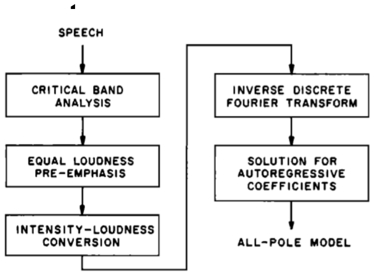
…

+ support from MSR, Google

www.zeropeech.com

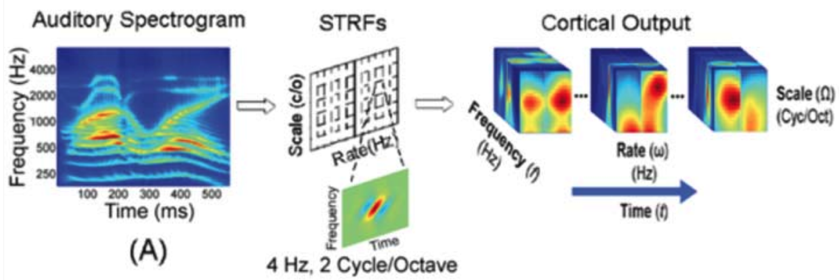
Learning acoustic representations from scratch

- Acoustic features
  PLP, RASTA



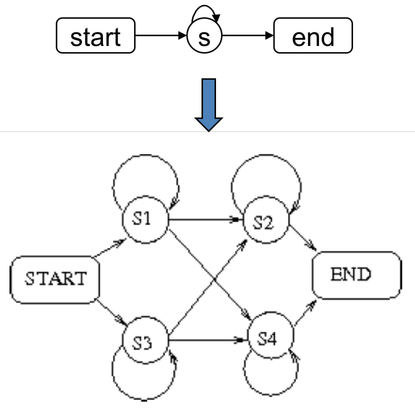
Hermanky (1990). *JASA*

- Auditory model



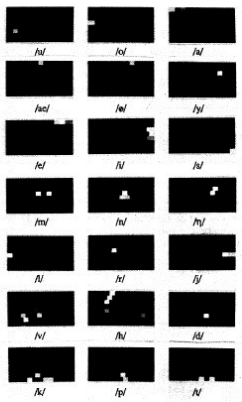
Chi, Ru, & Shamma (2005) *JASA*

- HMM state splitting



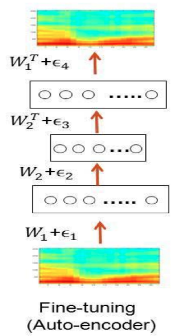
Varadarajan,
Khudanpur,Dupoux, (2008)

- Kohonen's maps



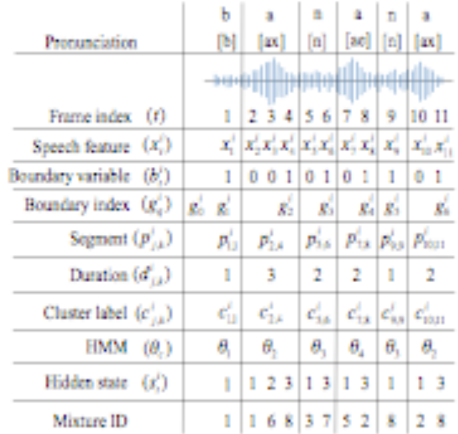
Kohonen (1988), *Computer*

- Deep autoencoders



Badino, Canevari, et al (2014),
ICASSP.

- Non Parametric Bayesian Clustering



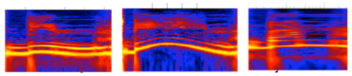
Lee & Glass, (2012). *Proc of ACL*

Learning acoustic representations: evaluation

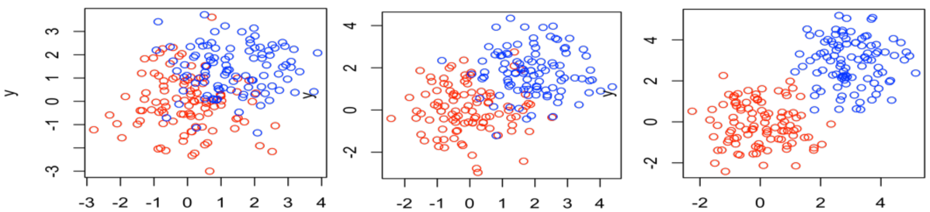
Minimal pairs ABX task

A B X

ba_{T1} ga_{T1} ga_{T2}



$$\theta(A,B) := \frac{1}{m(m-1)n} \sum_{a \in A} \sum_{b \in B} \sum_{x \in A \setminus \{a\}} \left(\mathbb{1}_{d(a,x)<d(b,x)} + \frac{1}{2} \mathbb{1}_{d(a,x)=d(b,x)} \right), \quad m = |A|, \; n = |B|$$



ABXerr=30% (1.6σ) | ABX=.20 (2σ) | ABX=.05 (2.4σ)

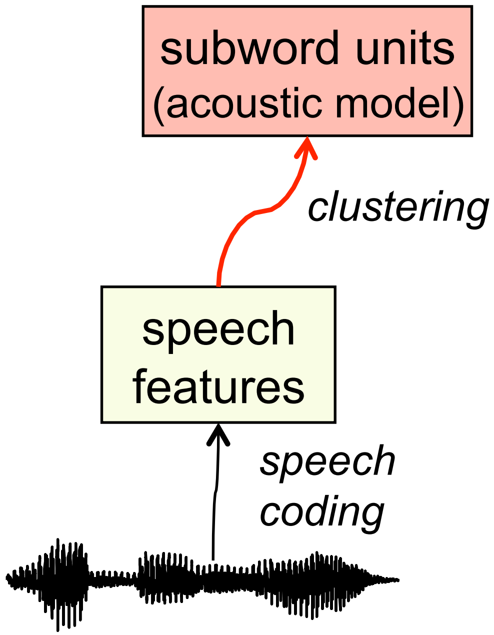
Comparison

- baseline: MFCC
- topline: supervised state-of–the-art system

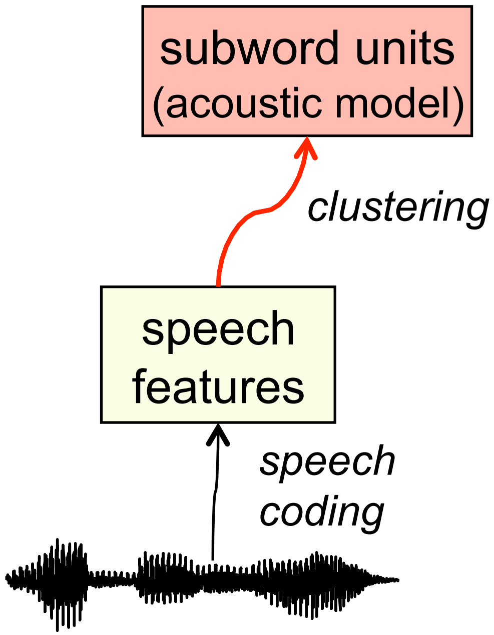
→ can we approach the topline?

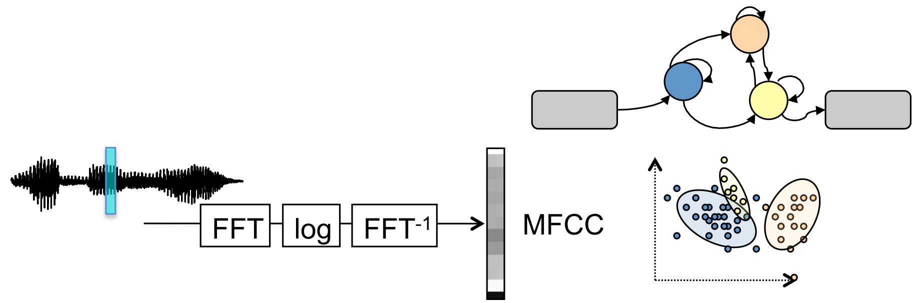
Schatz et al, 2013;2014

Idea #1: bottom up learning



Idea #1: bottom up learning



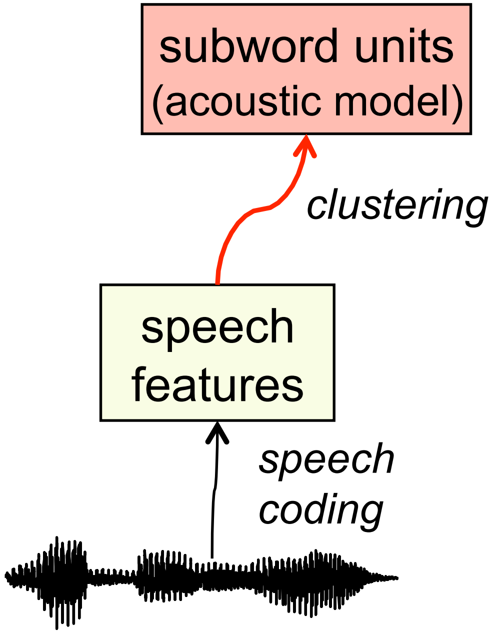


Successive State Splitting

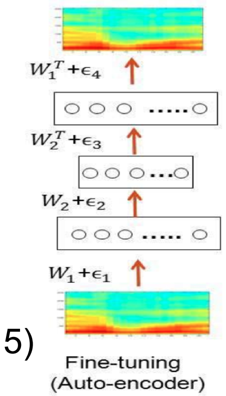
<table>
<tr><th>State seq</th><th>Allophones</th></tr>
<tr><td>11,28,32
15,17,2
3,17,2</td><td>[V]-t+[e|a|o]
[g|k]-[u|o]+[⋆]
[k|t|g|d]-a+[k|t|g|d]</td></tr>
<tr><td>31,5,13,5
17,2,31,11
3,30,22,34</td><td>[V]-[s|sj|sy]+[V]
[g|t|k|d]-[a|o]+[t|k]
[⋆]-a</td></tr>
<tr><td>6 24 8 15 22
22 35 11 28 32
4 17 24 2 31</td><td>[⋆]-o
[N|i|u|o]-[t|d]+[e|o|i]
[s|sy|z]-o+[t|d], [t|d]-o+[s|sy|z]</td></tr>
</table>

Varadarajan, Khudanpur,Dupoux, (2008)

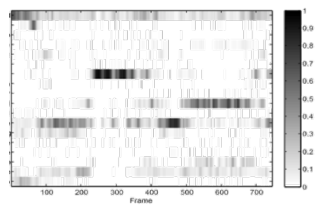
Idea #1: bottom up learning



- Low dimension continuous representations
  - Autoencoders (e.g. Badino et al. 2015)



- Probabilistic codes
  - posteriors of unsupervised GMMs (e.g. Heck et al 2015)



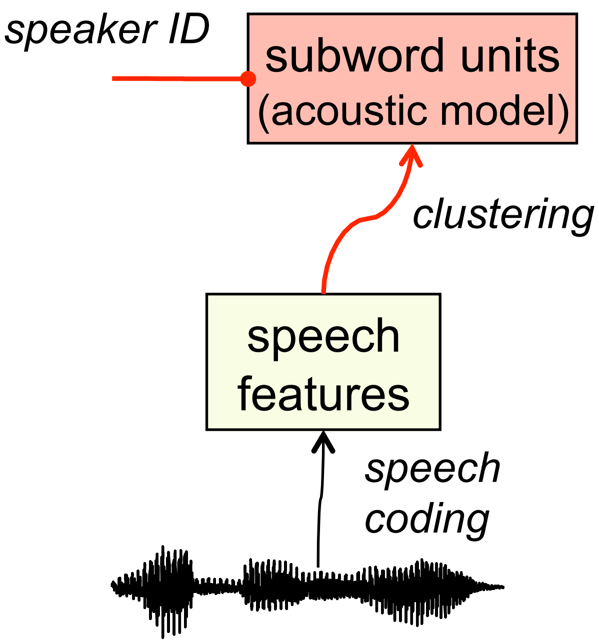
- Discrete codes
  - Unsupervised clustering, Hierarchical Bayesian (Lee & Glass, 2012; Ondel et al 2016), binarized DNNs (e.g. Myriam & Salvi 2017)

→ Simple idea, achieves interesting result, can be made more powerful with stronger priors, needs work on scalability

Main idea: information compression

- spectral information: 20800bit/sec,
- phoneme information: ~100bits/sec
- ***→ a 200x reduction !***

Idea #1b: invariant code

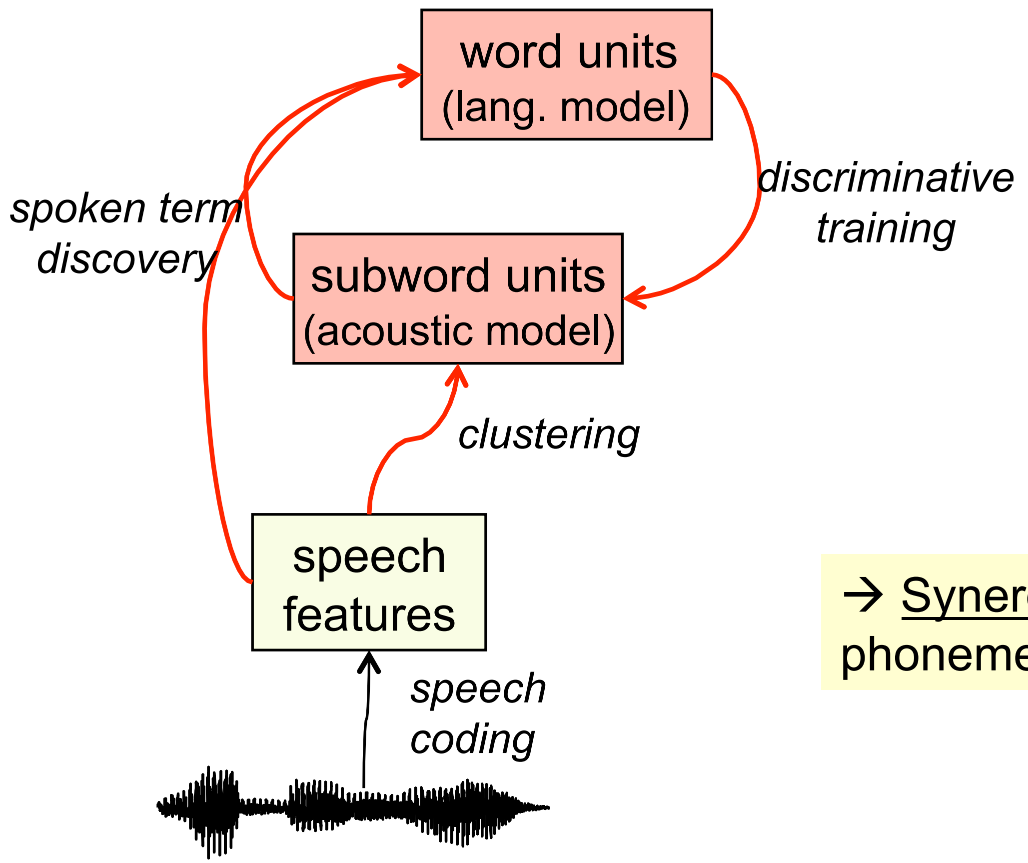


- speaker normalization
  - vocal tract normalization
  - fMMLR

(Heck et al. 2017)

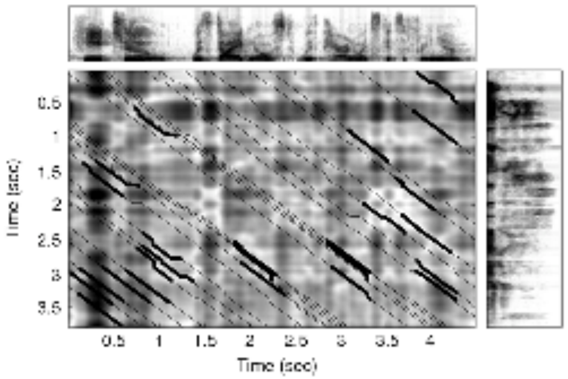
Main idea:
- assume infants know who is talking
- remove this information

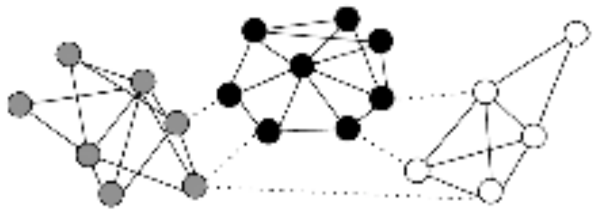
Idea #2: joint lexical-sublexical learning

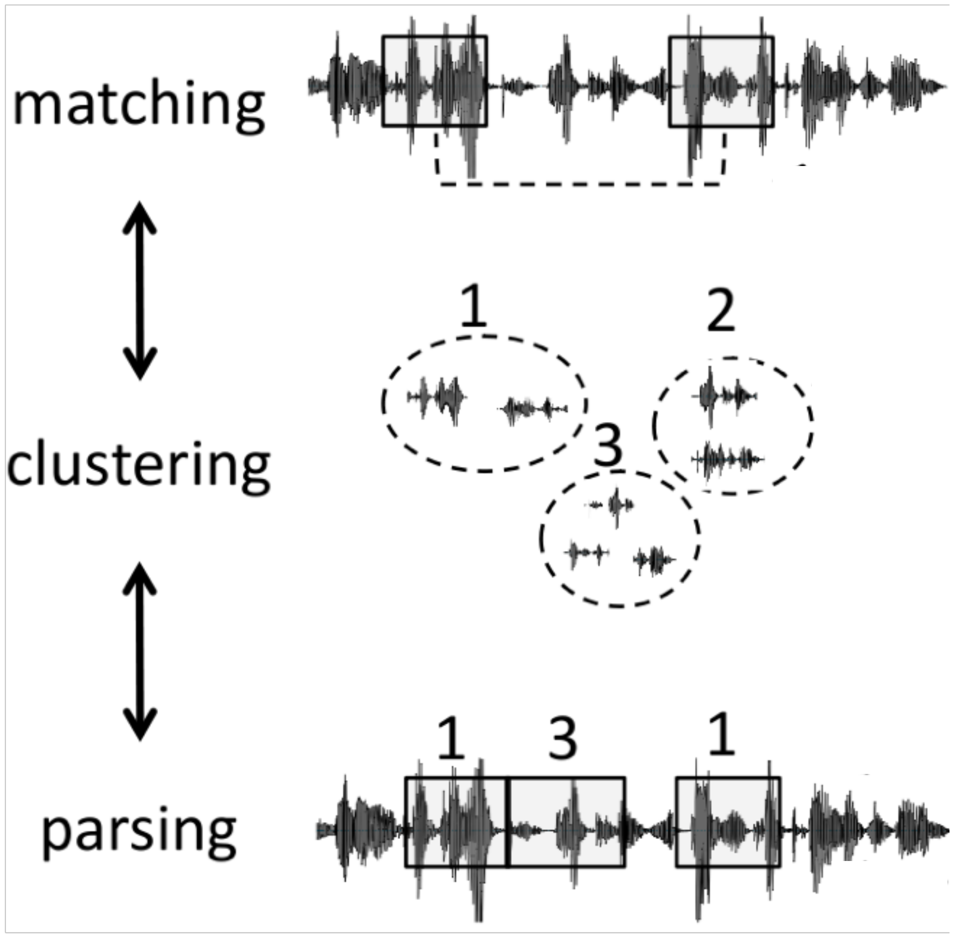


→ Synergies between word and phoneme learning

Spoken Term Discovery







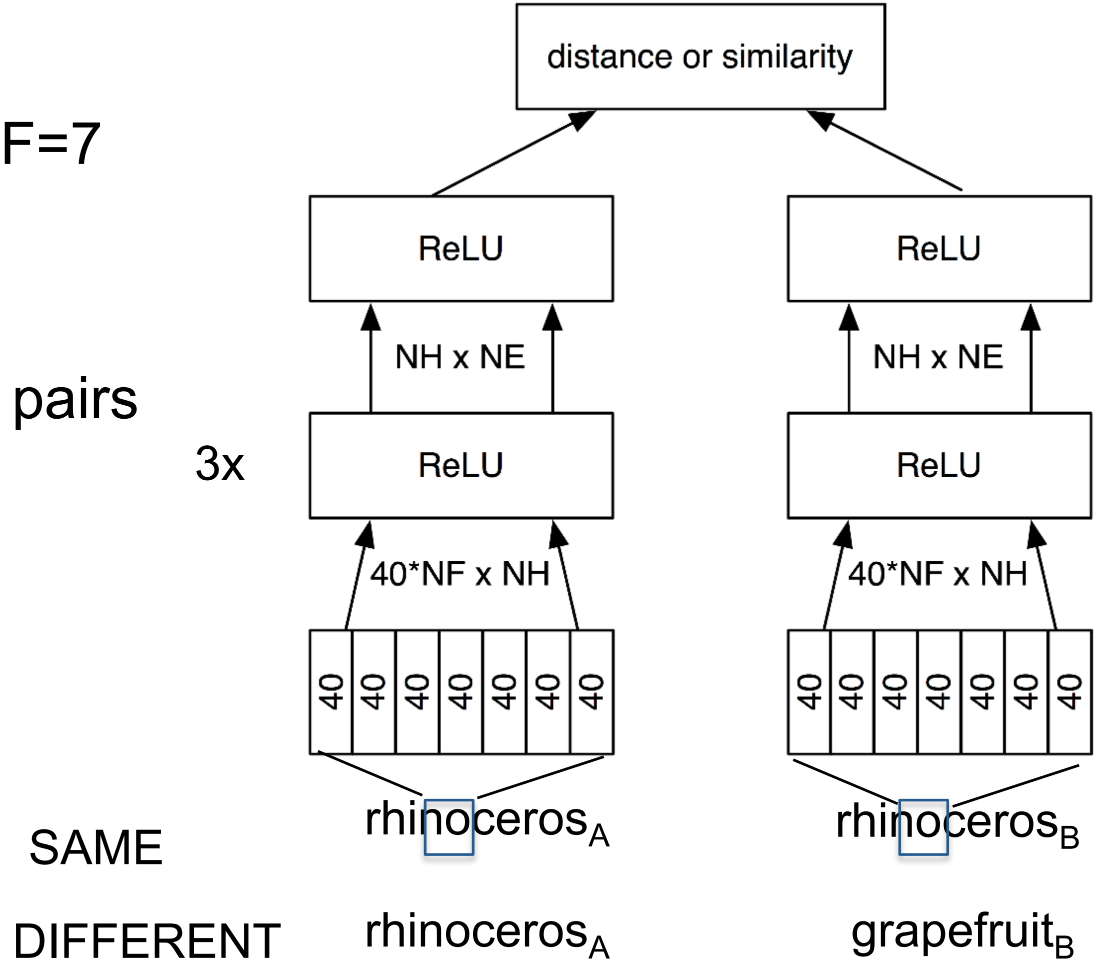
(Viterbi decoding)

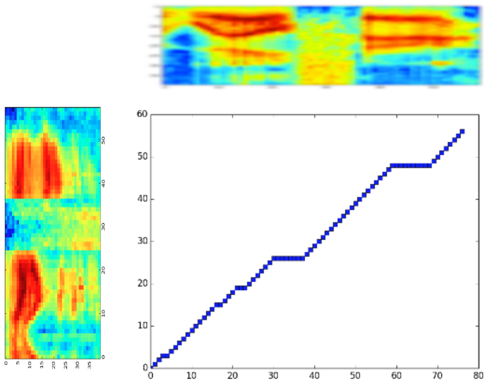
Algorithms: Park & Glass, (2008), Jansen et al. (2010), Muscariallo et al (2011)

Imagine you already have a lexicon of word forms

$$cost(X_A, X_B, \{\text{same}, \text{different}\})$$
$$\begin{cases} 1 - \cos(Y_A, Y_B) & \text{if same} \\ \cos^2(Y_A, Y_B) & \text{if different} \end{cases}$$

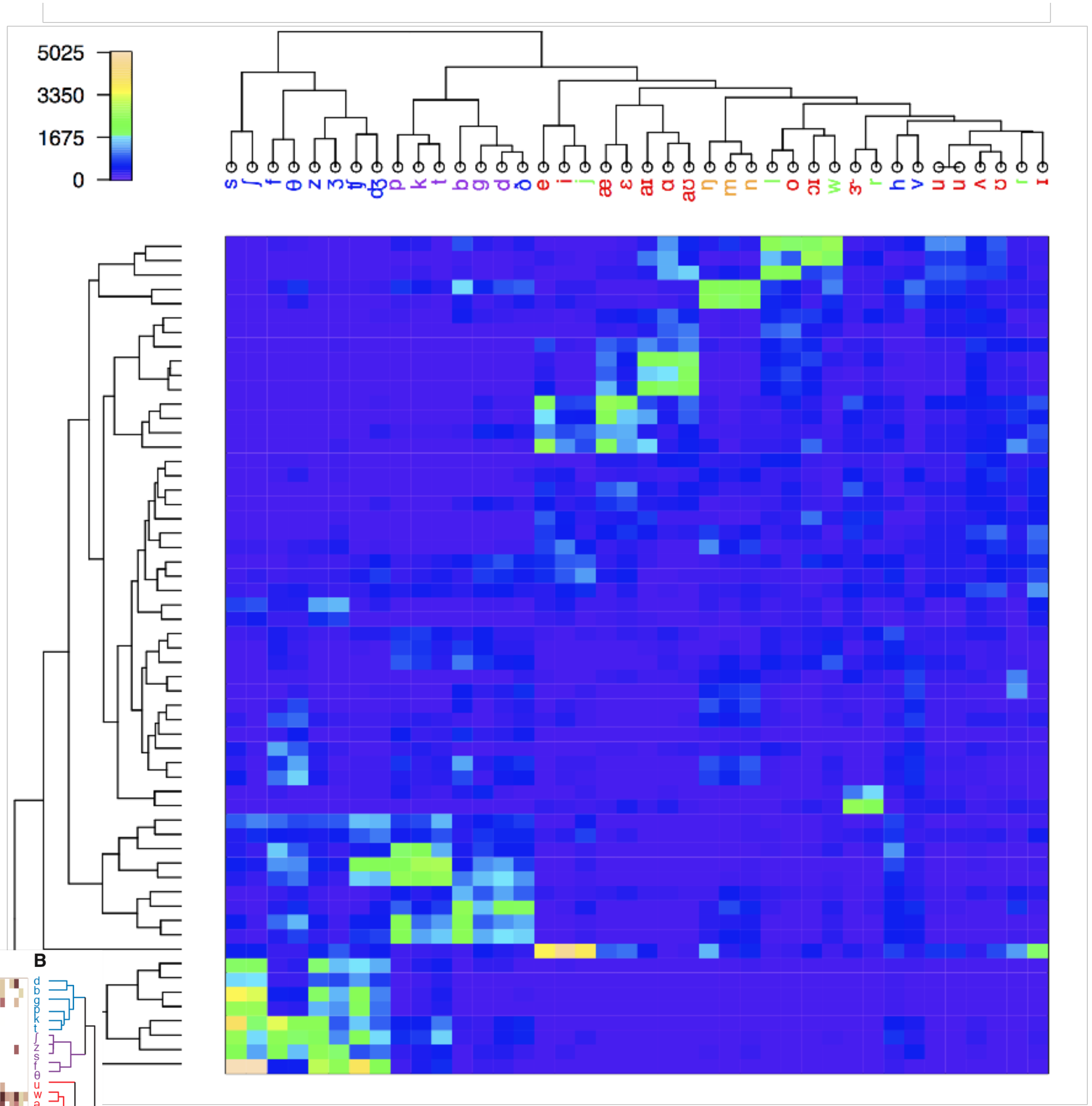
- NH = 1000, NE=100, NF=7
- 3 hidden layers
- TIMIT database
- 1737 word types → 62k pairs



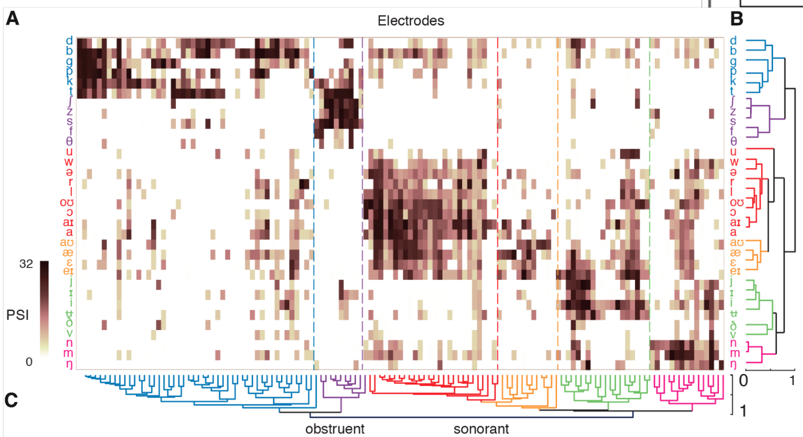


Dynamic Time Warping

- learns a sparse embedding



Synnaeve et al, 2014

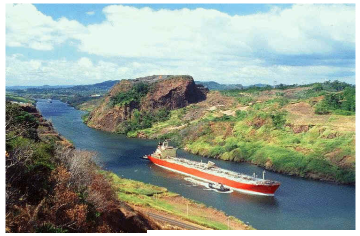


Mesgarani et al, 2014

A potential problem: allophones

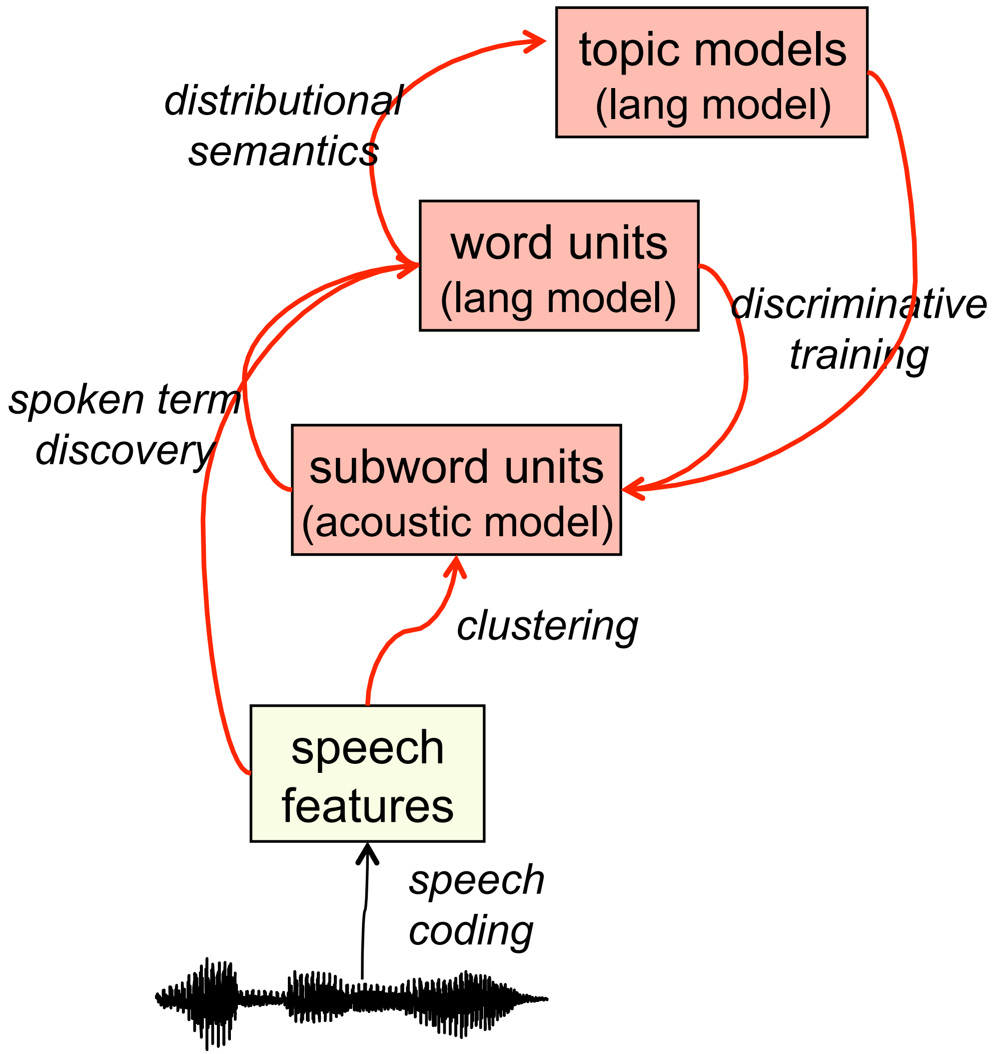
(/canaR/ vs /canaX/) ➔ (R,X) allophones

(/canaR/ vs /canaL/) ➔ (R,L) allophones



CANAL+

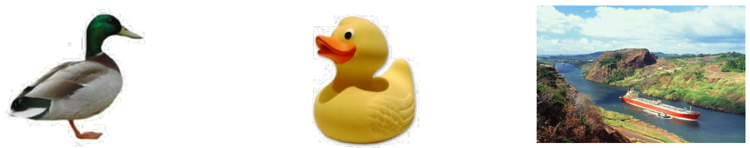
Idea #3: joint topic-lexical-sublexical learning



→ *scalable, can reach supervised systems*

<u>I. Learn topics on the basis of protolexicon</u>
→ *each protoword has now a vector representation*

<u>II. Use semantic distance to help subword clustering</u>
→ *'semantic' cosine distance combined with acoustic distance to cluster protophonemes*

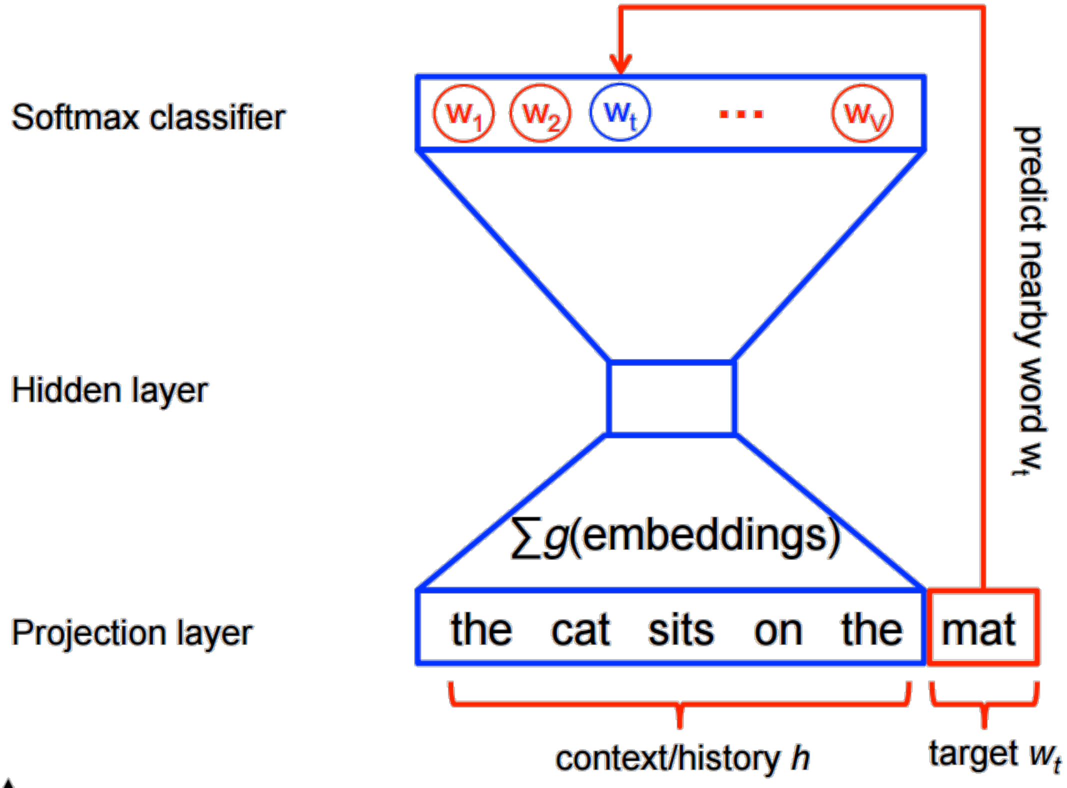


/kanaX/ vs. /kanaR/ vs. /kanal/

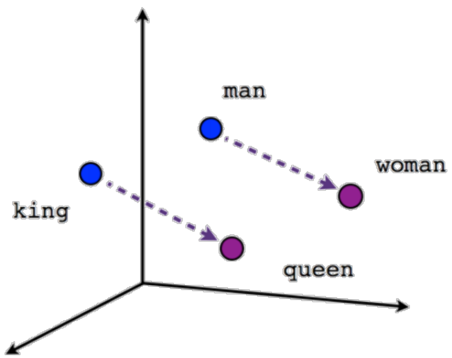
$\underbrace{\text{/kanaX/ vs. /kanaR/}}_{\textit{allophones}}$ $\underbrace{\text{/kanaR/ vs. /kanal/}}_{\textit{phonemes}}$

→ *Proof of principle with allophonic transcription; not done yet with raw speech*

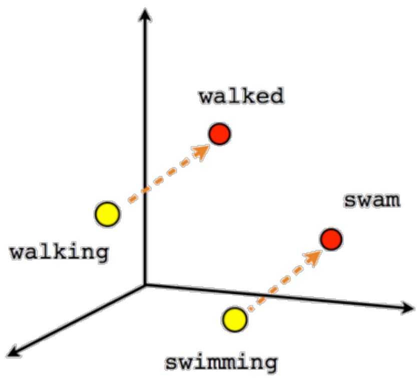
Fourtassi & Dupoux (2014)



https://www.tensorflow.org/tutorials/word2vec/



Male-Female



Verb tense

→can help to sort allophones from phonemes (Fourtassi & Dupoux, 2014)

Idea #3b: joint semantic lexical-sublexical learning

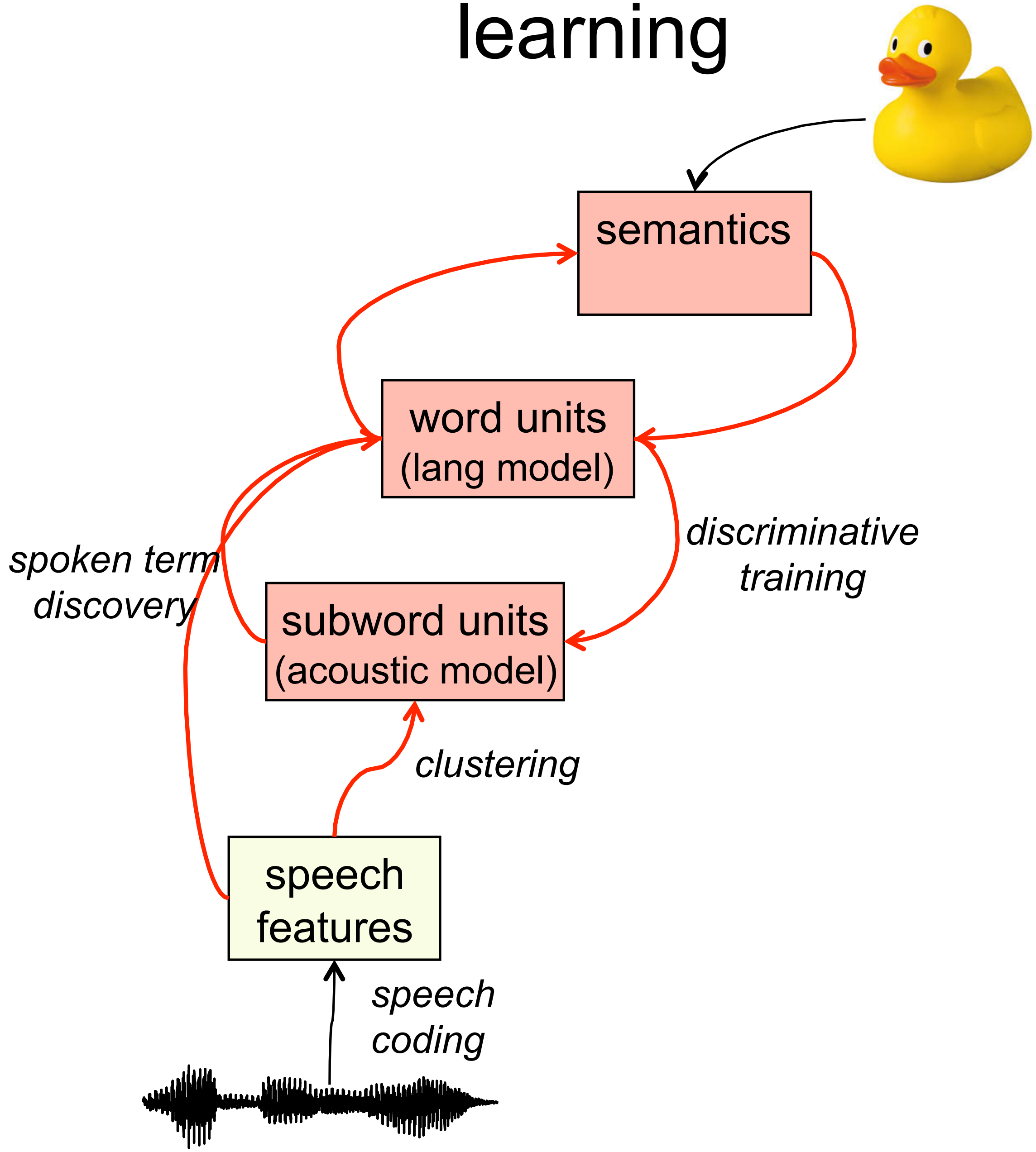
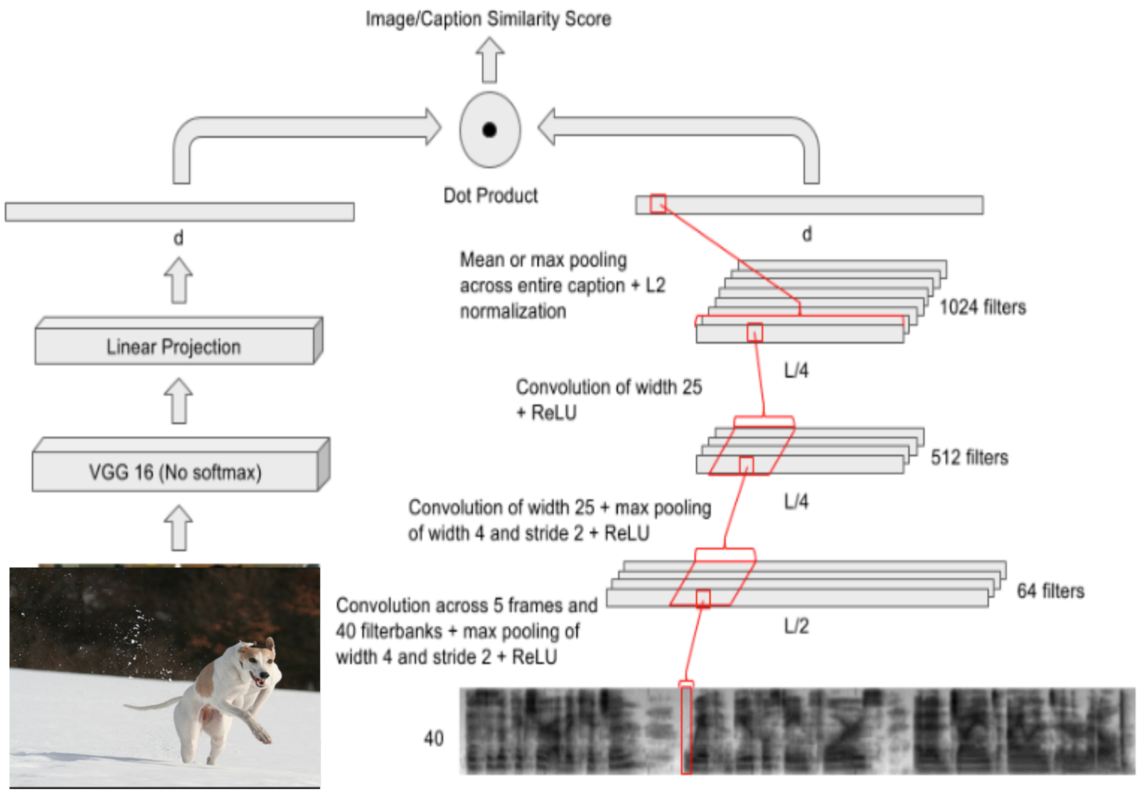


image retrieval



-A brown and white dog is running through the snow
-A dog is running in the snow
-A dog running through snow
-A white and brown dog is running through a snow covered field
-The white and brown dog is running over the surface of the snow

Harwarth & Glass 2015; 2016; 2017; Chrupala et al. 2017

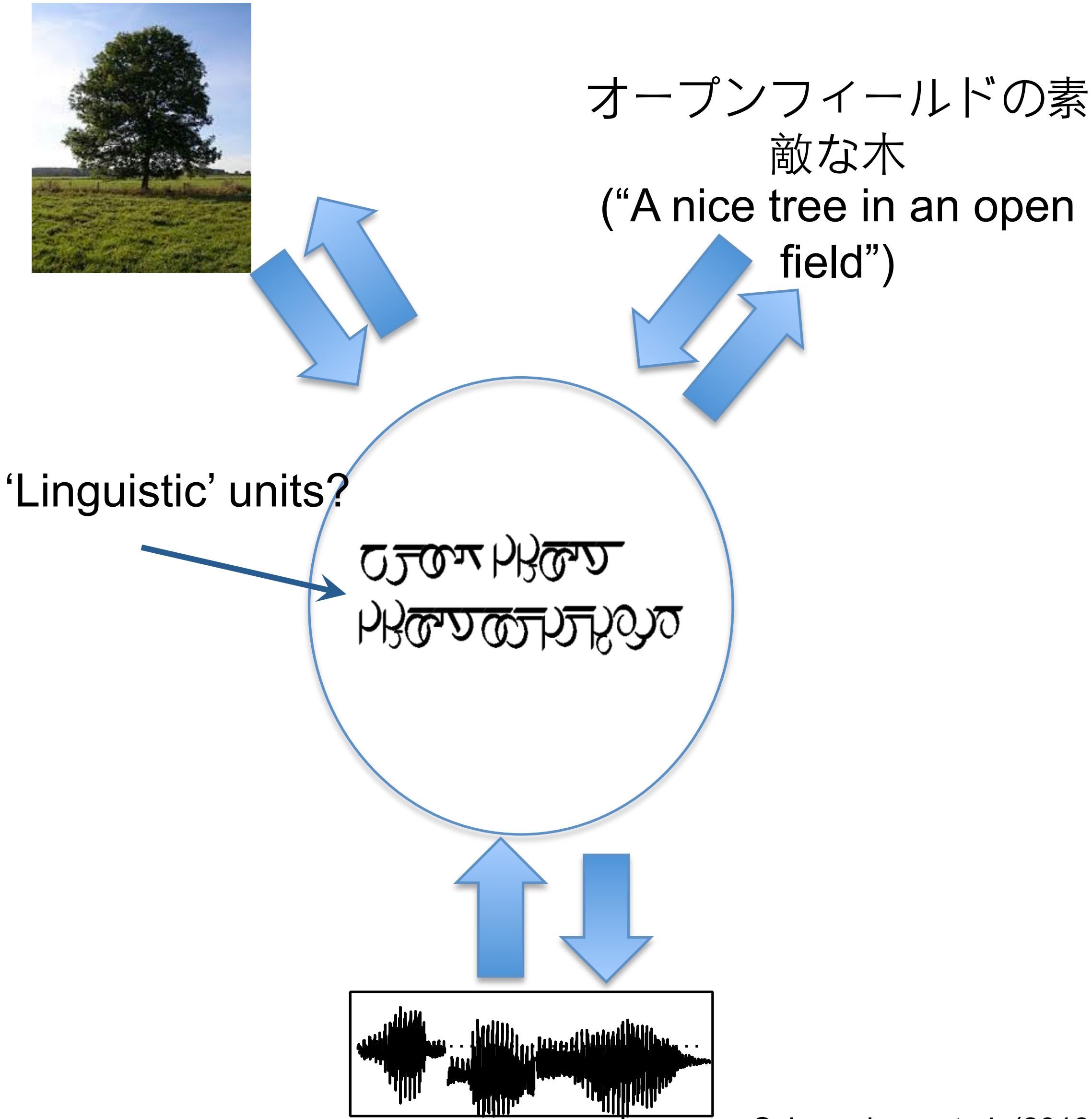
JSALT "Speaking Rosetta Project" CMU Summer 2017

Technological objective:

- Build speech technology for languages without orthographies

Scientific objective:

- Etudier l'émergence d'unités linguistiques symboliques
- dériver des prédictions



www.lti.cs.cmu.edu/2017-jelinek-workshop

Scharenborg et al, (2018). ICASSP

In brief

- Machine learning/AI could help understanding language acquisition
- But only if new, data efficient, unsupervised algorithms are constructed

III. What’s have we learned?

Testing old theories or deriving new predictions

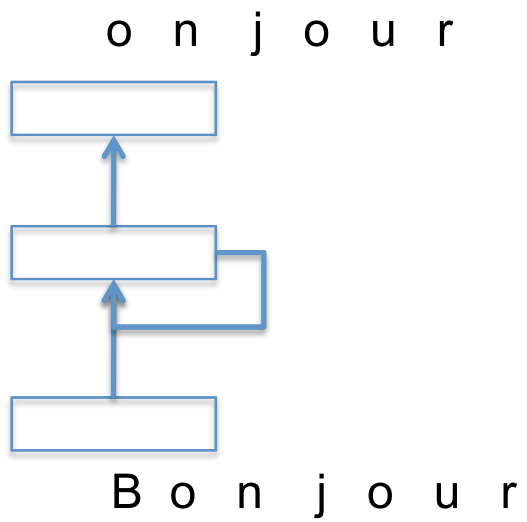
- Learnability in the limit: comparing with human adults
  - internal tests (comparison with gold standards: eg, segmentation F scores)
  - external tests (comparison with performance on behavioral tests: e.g. ABX discrimination tests)

→*already extremely constraining; most algorithms fail*

- Infant/Machine comparisons:
  - Testing old theories
  - Testing new predictions

Learning in the limit: AI Psycholinguistics

→ entrainer un réseau de neurone à prédire le caractère suivant



SCÈNE III.--ALCANTOR, BASQUE, MARIANE, DU CROISY,
BESTARIN, LE BARBOUILLÉ, MASCARILLE.

MASCARILLE.
Je ne puis davantage à propos.

MARIANE.
o ciel! de tout ce qu'il doit faire, et sa gloire à tous deux,
Qui sait se montrer des vœux de notre ressentiment:
Si bien de suivre le plus grand embarras?
Mais on puisse ravir à vous payer de vous faire l'ardeur.

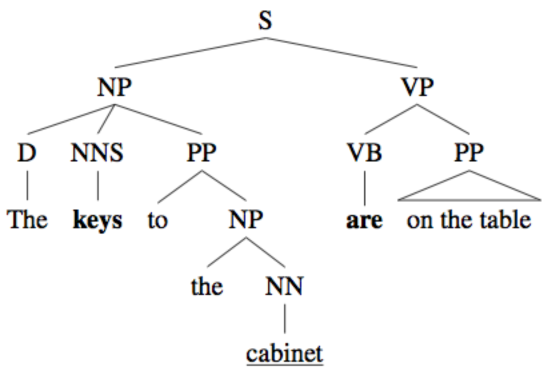
ÉRASTE.
Je ne sais.

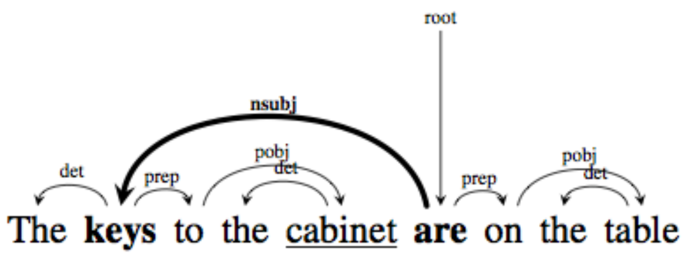
trained on Molière's works

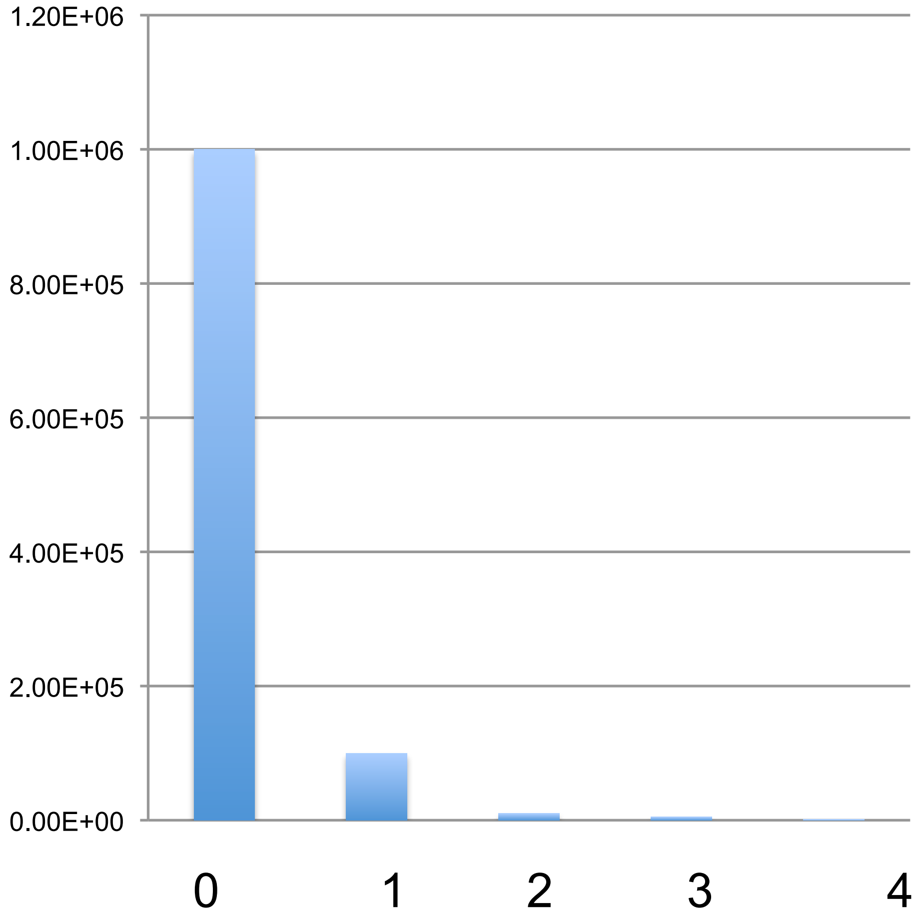
Karpathy (2015). The Unreasonable Effectiveness of Recurrent Neural Networks http://karpathy.github.io/2015/05/21/rnn-effectiveness/

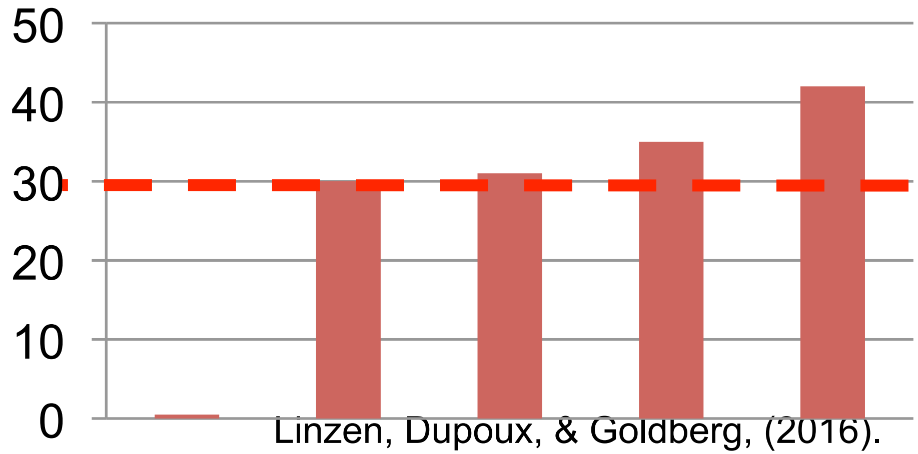
Yes, but…

a. The **key is** on the table.
b. *The **key are** on the table.
c. *The **keys is** on the table.
d. The **keys are** on the table.









Linzen, Dupoux, & Goldberg, (2016).

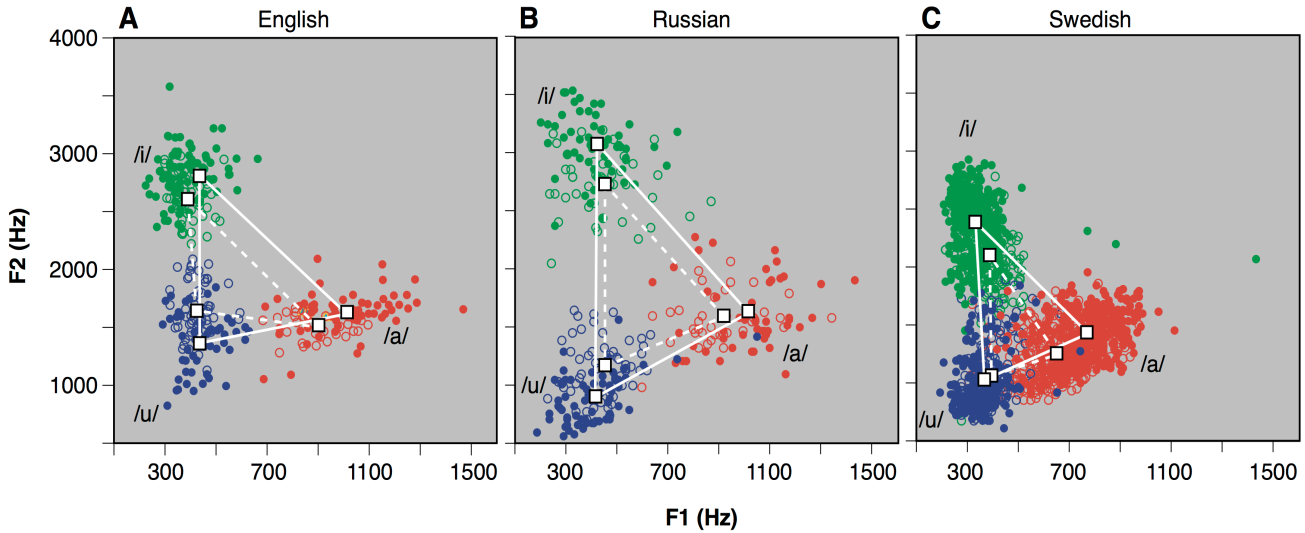
- Learnability in the limit: comparing with human adults
  - internal tests (comparison with gold standards: eg, segmentation F scores)
  - external tests (comparison with performance on behavioral tests: e.g. ABX discrimination tests)

- Infant/Machine comparisons:
  - Testing old theories
  - Testing new predictions

Testing old theories: Baby talk as hyperspeech

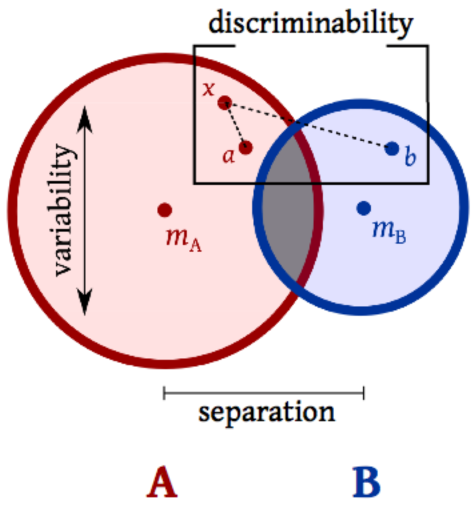
- The hyperspeech hypothesis: in IDS, parents facilitate the perception compared to ADS (Fernald, 2000).
- The hyperlearning hypothesis: in IDS parents facilitate phonetic learning (Kuhl et al 1997).

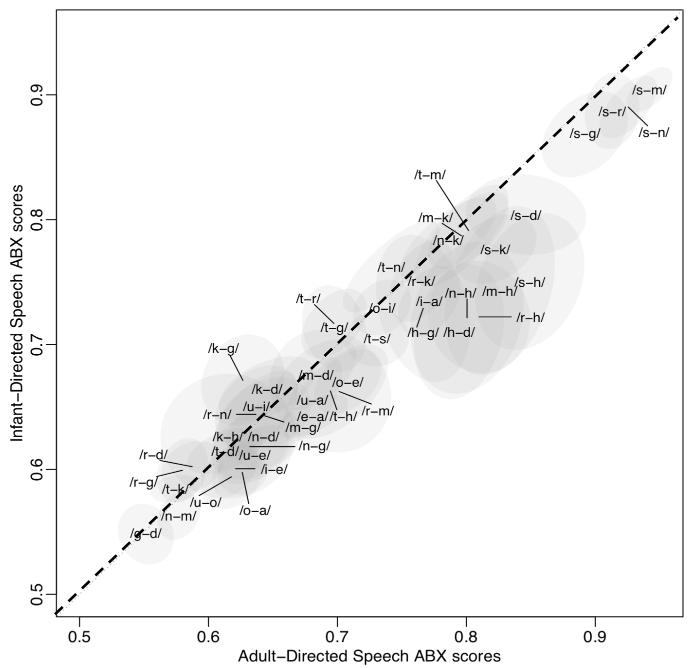
Moms of 6mo given 9 toys
/i a u/ stretched

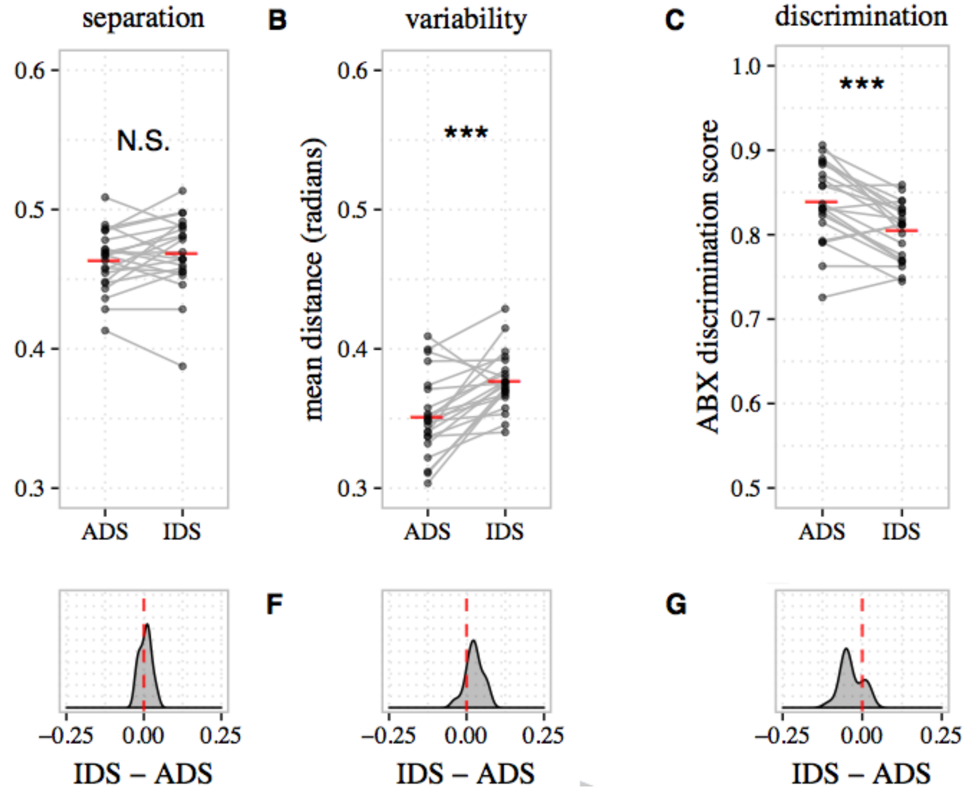


hyperarticulation hypothesis (Kuhl et al 1997)

- two counteracting forces
  - slightly more separation
  - much more phonetic variability
    - Guevarra-rukoz, et al (in prep)
    - Martin et al (2015)

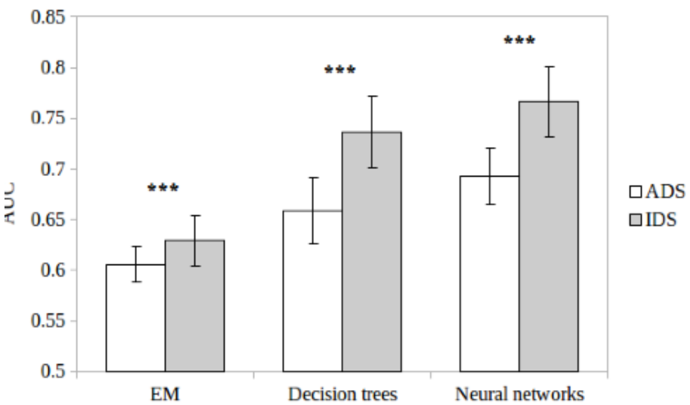


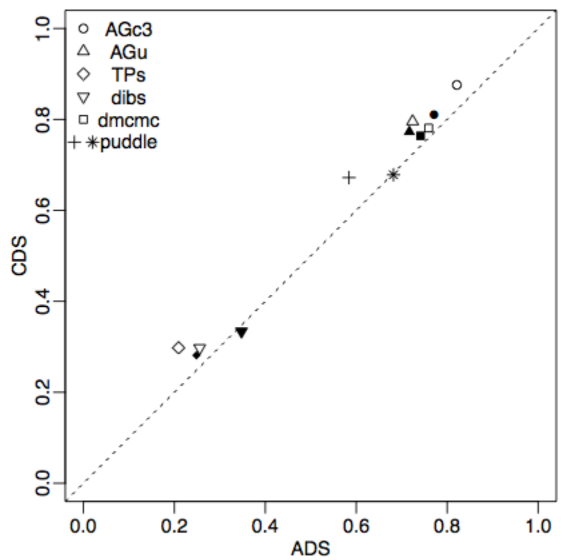




- other counteracting forces
  - slightly more distinct lexicon
    - onomatopeas
  - shorter sentences

  - better prosodic cues
    - Ludusan et al (2017)
    - pauses, F0 reset, duration



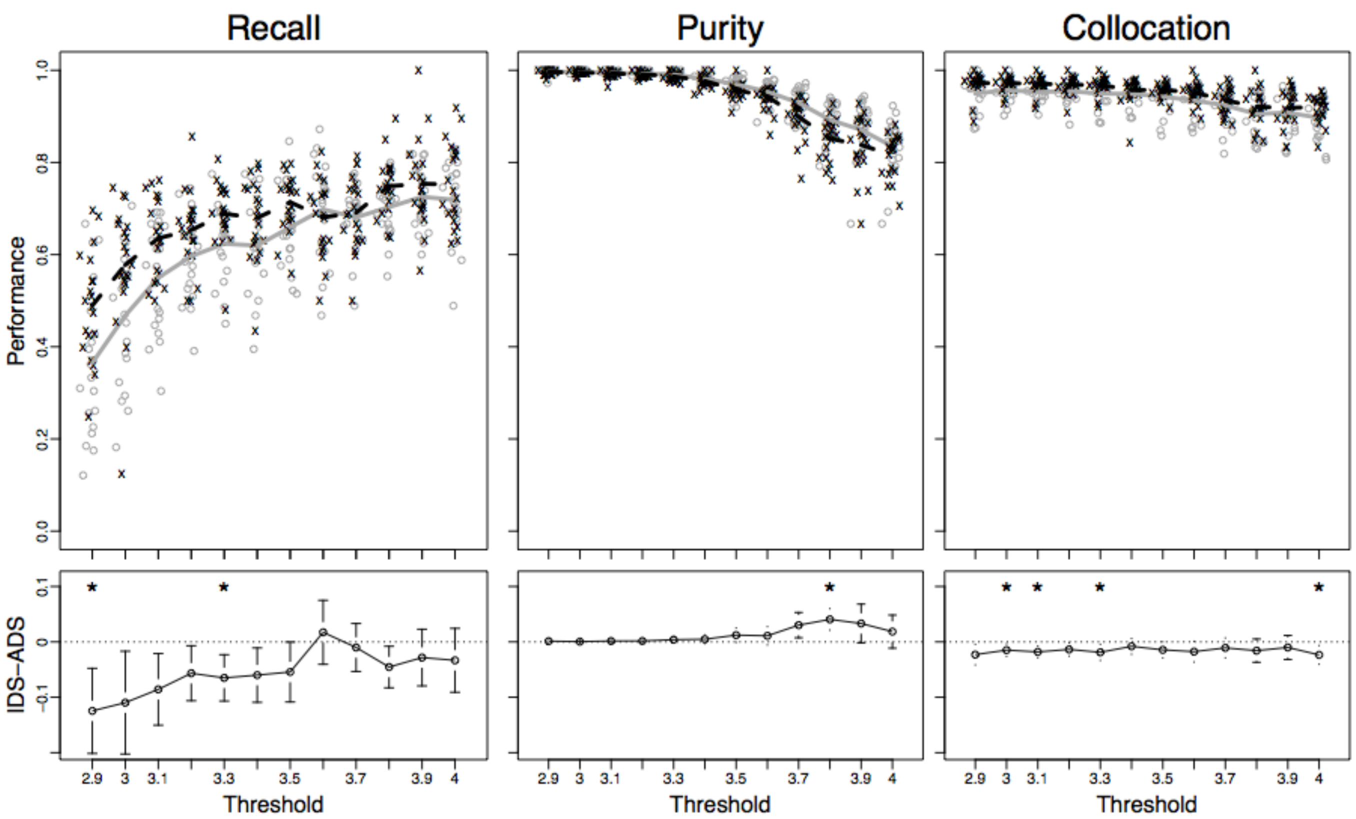


Cristia et al (in prep)

Overall effect

- spoken term discovery
  - 20 mothers
  - 20 object names
  - IDS vs ADS
  - MODIS system

→ not much difference

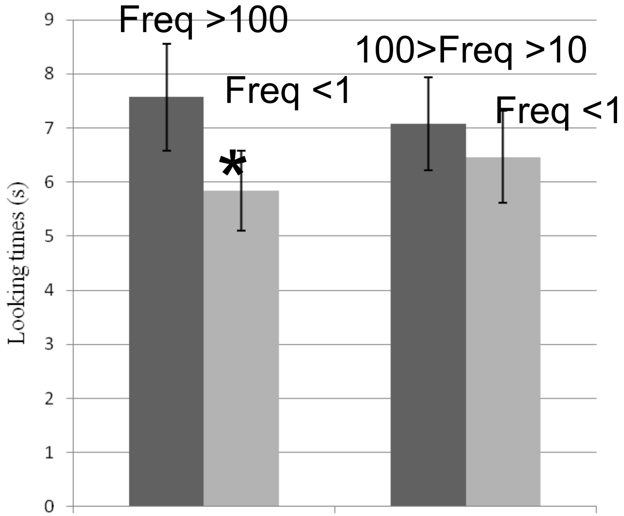


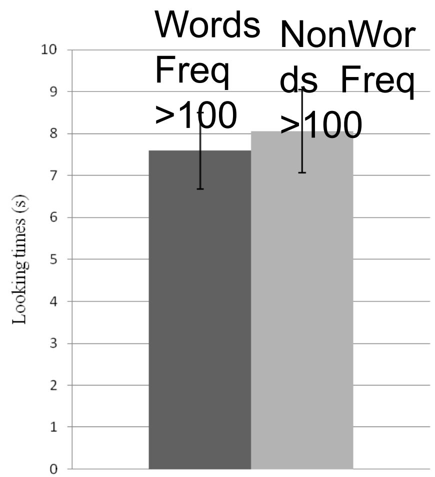
Ludusan, Seidl, Dupoux, Cristia (2015)

New predictions (I): missegmentations

- word discovery algorithms mis-segment words
  - do infants missegment too?
  - the infant protolexicon

[dɑ̃la]
[sepuʁ]
[kwasa]
[vafɛʁ]
[kɔʁɛ̃]
[mɛty]
[tule]
[akɛl]
[vɛpa]
[naply]
[pasyʁ]
[vødiʁ]





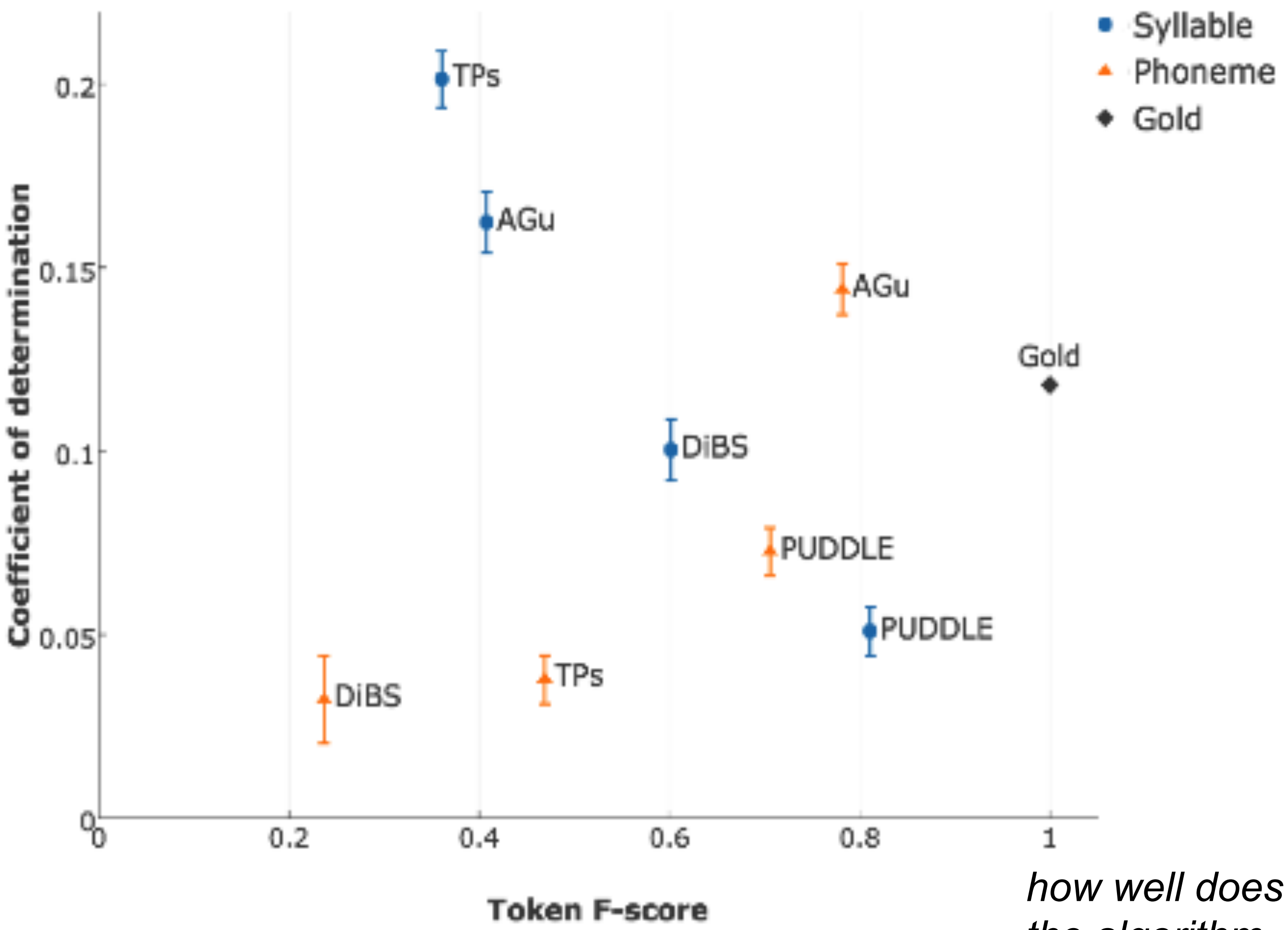
(N=16) (N=16) (N=16)

→*infants store many familiar bisyllable nonwords*

Ngon et al, (2012)

how well does the algorithm predict 13 mo infant's CDI vocabulary?

New predictions (II): predicted vocabulary

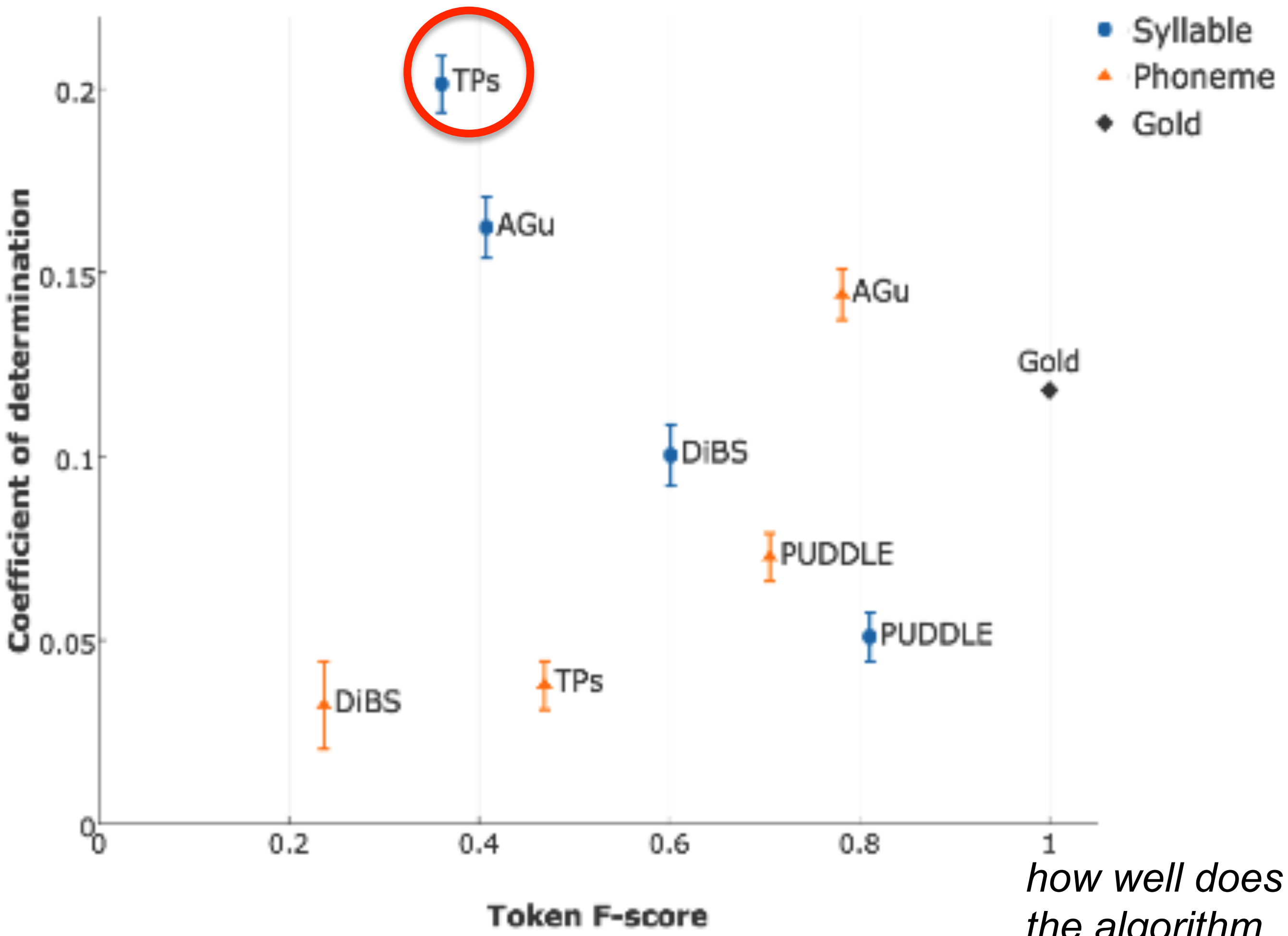


how well does the algorithm segments like adults do?

Larsen et al (2017), *Interspeech*

how well does the algorithm predict 13 mo infant's CDI vocabulary?

New predictions (II): predicted vocabulary



how well does the algorithm segments like adults do?

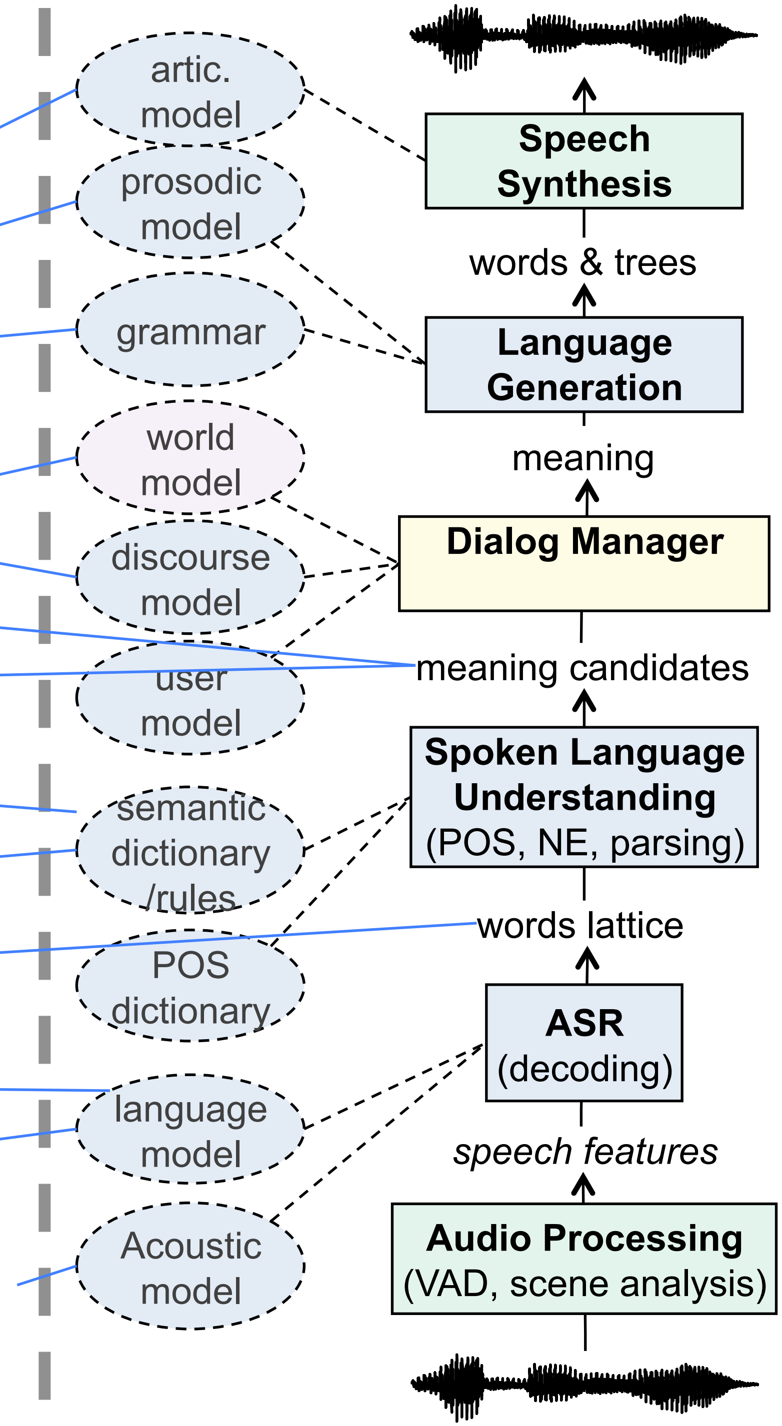
→ the algorithm that predicts the best infant vocabulary is under-optimal

Larsen et al (2017), *Interspeech*

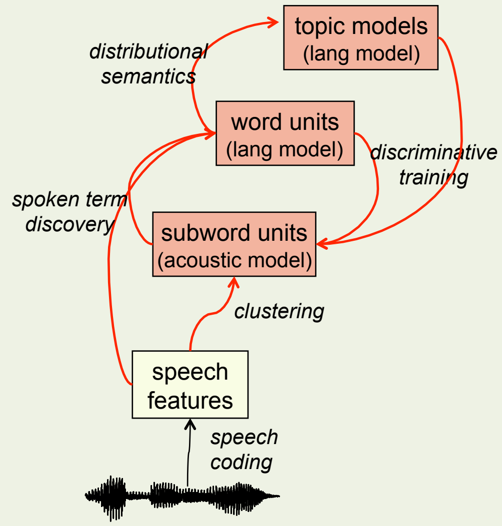
Towards language benchmarking

- simple psychophysical tasks (is X good? do X and Y match?)
- ground truth: humans/babies

18-24mo	Error analysis
5y	Felicity judgment
3y	Truth judgment
3y	Entailment judgment
2y	Grammaticality judgment
6mo	Word-Picture Matching
9mo	Word spotting
11mo	Lexical decision
9mo	Phonotactic judgment
newborn	Speaker invariant discrimination (ABX)



Summing up



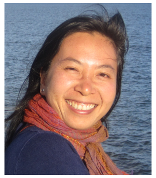
- Reverse engineering is feasible
  - realist data
  - scalable models
  - quantitative predictions
- It addresses the two deep puzzles
  - learnability problem:
    - bootstrap: provides a proof of principle that (some) learning is possible from raw sensory data, (provided a specific learning architecture, -- a computationally explicit LAD)
    - co-dependencies: not a problem, but an asset (*synergies*)
  - learning trajectories:
    - graduality & simulaneity:
      - can be explained through synergies
      - the possibility of sub-optimal algorithms
    - resilience:
      - still a lot to do here (data efficiency problem of machine learning)
      - we explored the functional role of infant directed speech

Project Bootphon

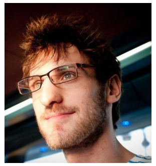
2012-2017

Thank you


Roberta Pinna


Xuan Nga Cao


Emmanuel Dupoux


Gabriel Synnaeve

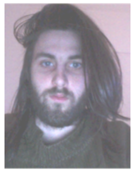

Maarten Versteegh


Ewan Dunbar


Bogdan Ludusan


Cristina Bergmann


Tal Linzen


Thomas Schatz

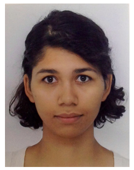

Mark Johnson


Aren Jansen


Abdellah Fourtassi


Hynek Hermansky


Julia Carbajal


Adriana Guevara


Sanjeev Khudanpur


Catherine Urban


Roland Thiolière


Mathieu Bernard

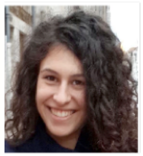

Juan Benjuema


Julien Karadayi


Rachid Riad


Rahma Chaabouni


Ronan Riochet


Elin Larsen


Neil Zeghidour

and many interns…


Sharon Peperkamp


Francis Bach


Alex Cristia

erc

ANR

îledeFrance

amazon web services

Facebook AI Research

Microsoft


Andy Martin


Vijay Peddinti


Reiko Mazuka