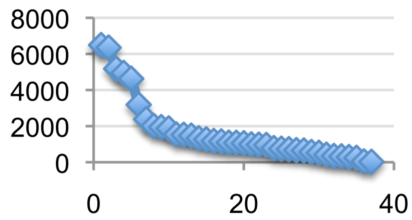
Algorithms for speech and language processing

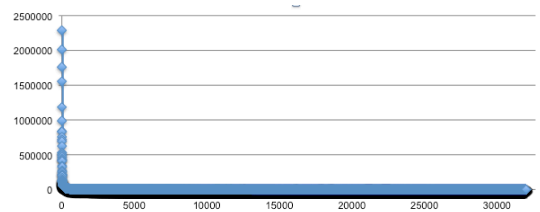
Challenges and perspectives

Power law, everywhere

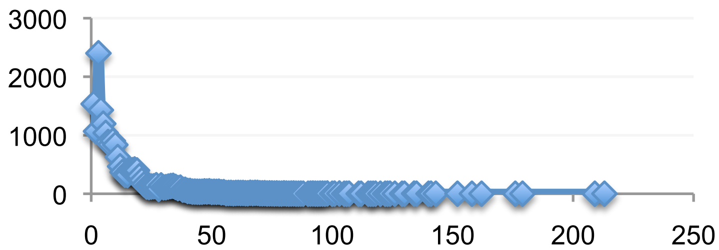
- sounds



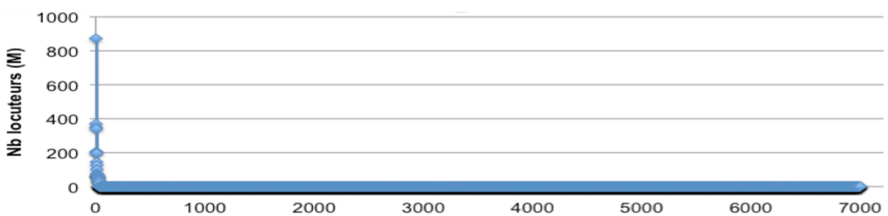
- words



- sentences



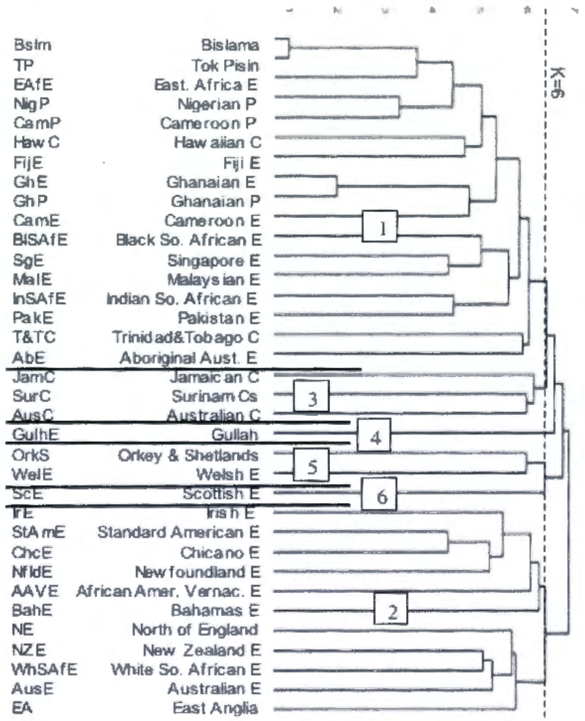
- languages



Sounds: dialects, registers. etc

- Google Speech API
  - 13 versions of English
    - English (Australia) en-AU
    - English (Canada) en-CA
    - English (Ghana) en-GH
    - English (Great Britain) en-GB
    - English (India) en-IN
    - English (Ireland) en-IE
    - English (Kenya) en-KE
    - English (New Zealand) en-NZ
    - English (Nigeria) en-NG
    - English (Philippines) en-PH
    - English (South Africa) en-ZA
    - English (Tanzania) en-TZ
    - English (United States) en-US
  - 20 versions of Spanish
  - etc

Trudgill 1999



Nagy et al, 2006

Sounds: dialects, registers, etc

- whispered speech

TABLE I. PERFORMANCES OF DIFFERENT NORMALIZATION TECHNIQUES IN SPEECH, WHISPER AND BIMODAL SPEECH RECOGNITION

Normalization techniques	WER (%)		
	Speech	Whisper	(Speech +Whisper)
CMN	1.5	53.9***	27.7***
MVN	2.2	66.0**	34.1**
QCN	0.9*	67.1**	34.0**
QCN-RASTALP	0.6*	68.2**	34.4**
CGN	1.5	70.3**	35.9**
without normalization	1.4	87.3	44.4
CVN	2.4	92.9	47.7

*(p < 0.05 *; p < 0.01 **; p < 0.005 ***; Confidence interval = 95%)*

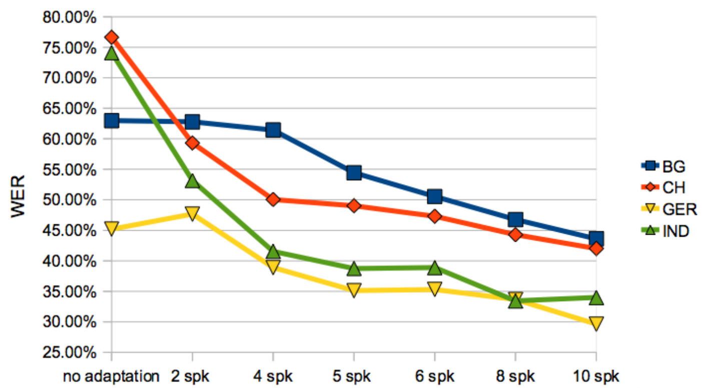
Grozdic, Jovicic et al. (2017)

- children's voice

- children's pitch higher than adult
→ MFCC features not good
- children voices and phonemes less stable
→ adding children and adult data improves children but degrades adult models

Dubagunta (ICASSP 2019)
Liao, Pundak et al (2015)

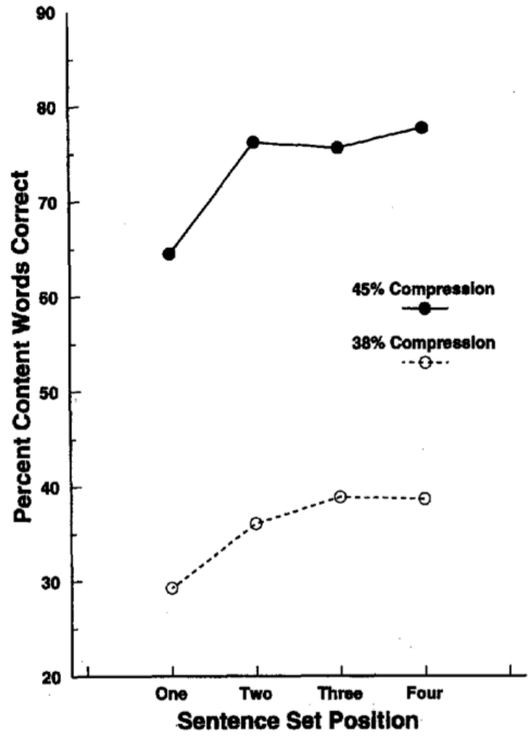
- foreign accented speech



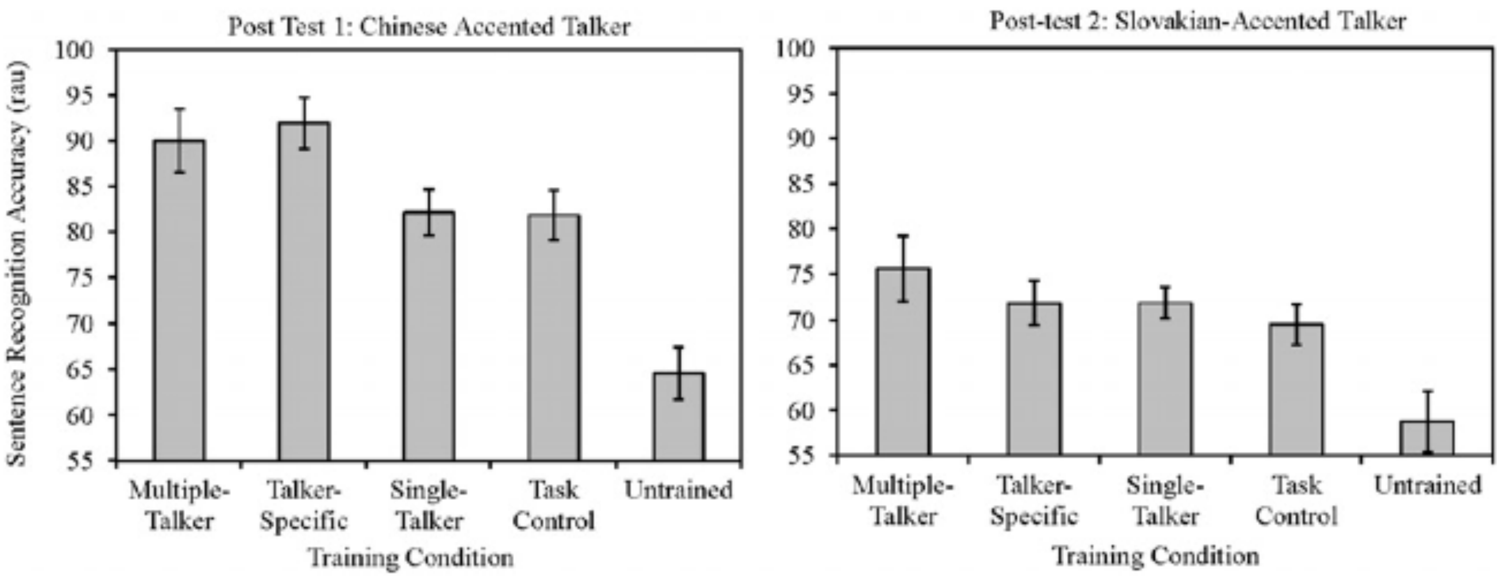
Mihaylova, (2011)

perceptual adaptation in humans

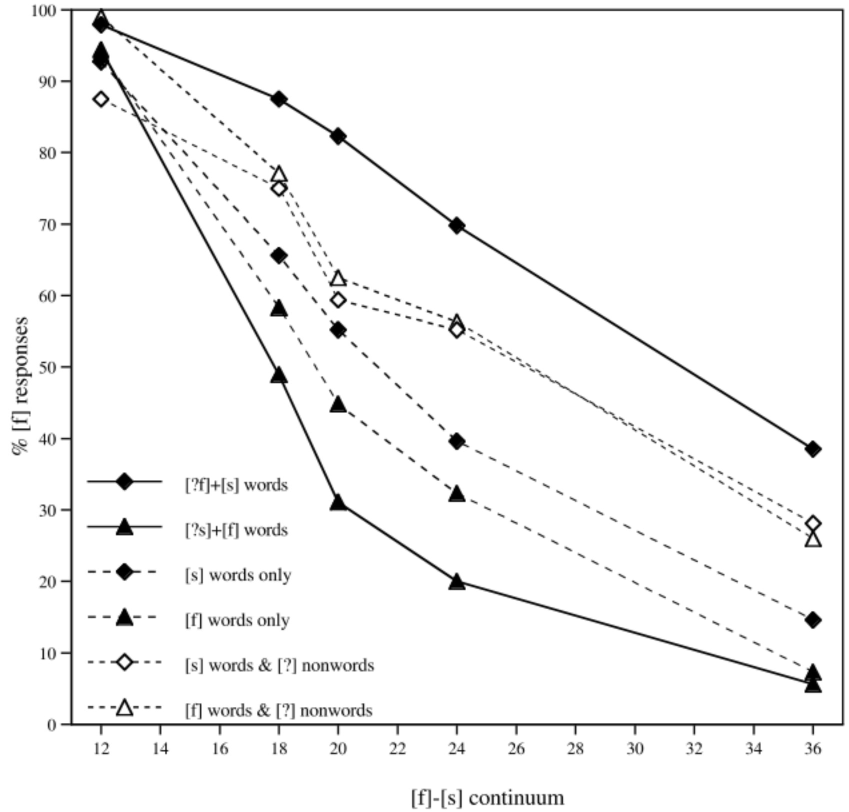
- fast dialect adaptation (Maye, Aslin, Tanenhaus, 2008)
  - the wicked witch of the west -> the weckup wetch of the wast
  - 20 minutes exposition
  - lexical decision: ‘wetch’ -> 69% word
- compressed speech (Dupoux & Green 1997)
  - 4 sets of 5 sentences



- accented speech (Bradlow & Bent 2008)
  - training: 5 repetitions of 16 sentences (in noise, no feedback)
  - test: 2 new sets of 16 sentences

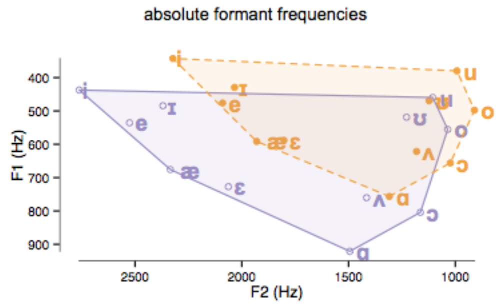


- perceptual learning
  (Norris, McQueen Cutler, 2003)
  - 20 words with ambiguous final s/f
  - (eg: astu[sf]e, or cara[sf]e)

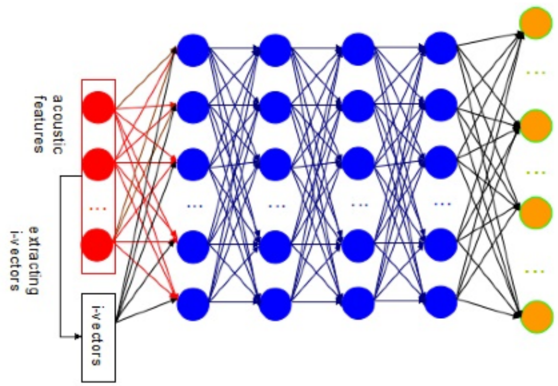


domain adaptation in machines

- fMLLR
  – x←Ax+b to maximize p(x|speaker)

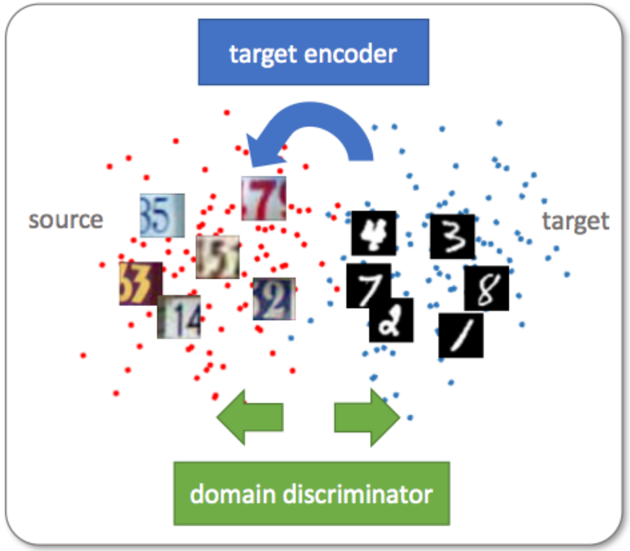


- i-vectors



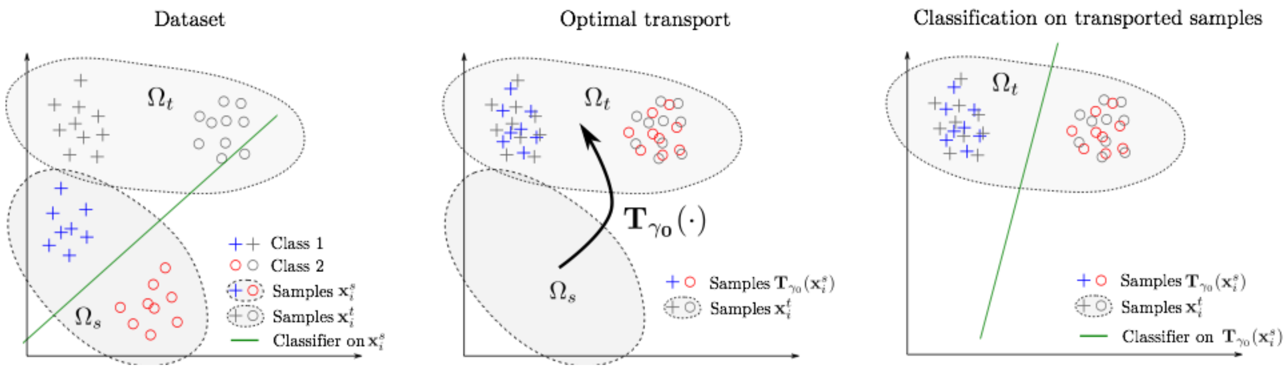
Chen, Liang et al (2015)

- adversarial domain adaptation



Tzeng, et al 2017

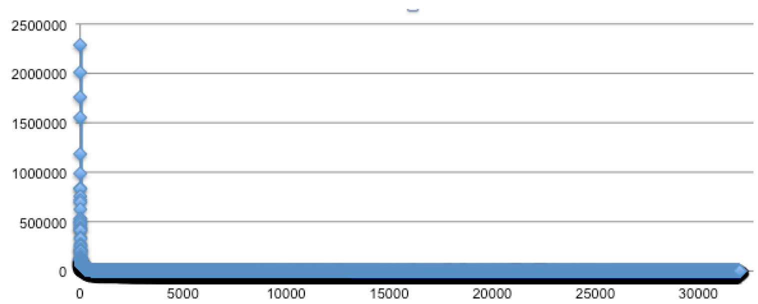
- optimal transport



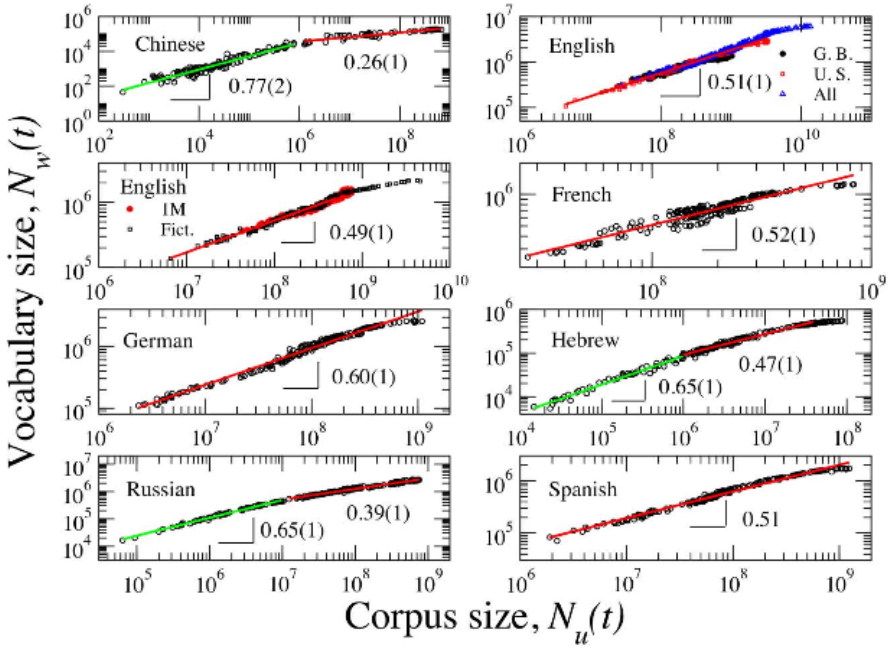
Courty et al (2016)

Voir survey (vision) Csurka (2017)

Words: the problem of oovs



pocket French 50Mw



- Heaps law

Petersen et al (2012)

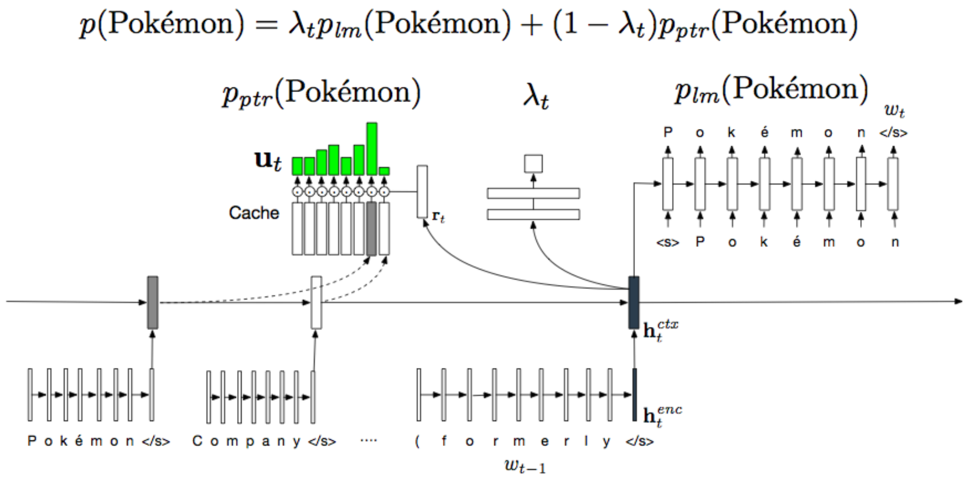
possible solution: subword codes

- eg:
  - character-based language models
  - word piece (eg. Byte Pair Encoding)

→ allows to generalize to new combinations of word parts, e.g.

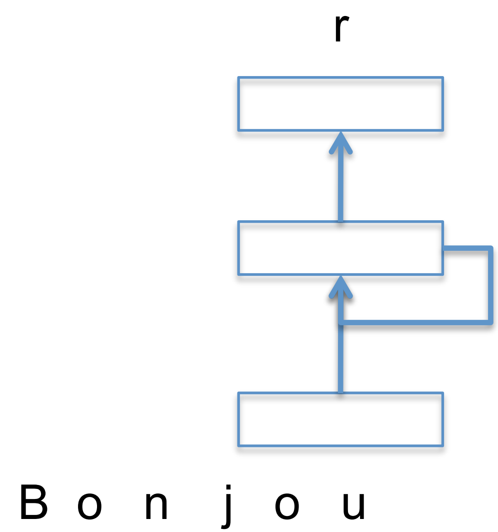
possible solution: episodic memory

- a cache based model of speech recognition Kuhn & de Mori (1990)
- Hybrid LSTM-Cache systems
  - Kawakami, Dyer, Blunsom (2017)
  - Grave Cisse Joulin (2017)



sentences: the problem of systematicity

→ RNN trained to predict the next character



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.

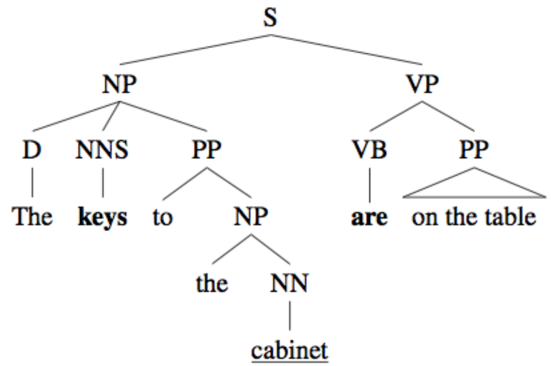
ÉRASTE.

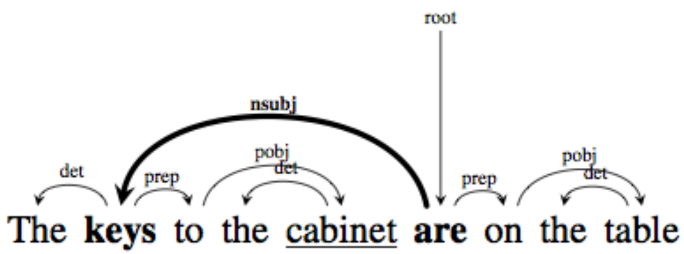
Je ne sais.

Trained on Molière's work

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

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.

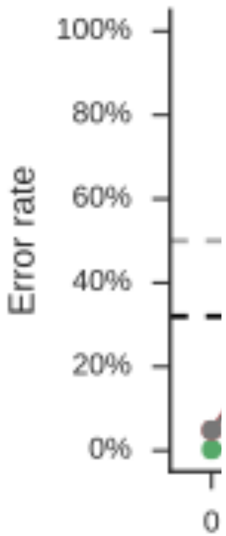




- infants acquire number agreement (in comprehension) at 18 months (Xi et al, in prep)
- Materials:
  – 1.3M sentences from 2013 Wikipedia dump
  – less than 50words
  – singular or plural noun subject
  – least one subject-verb dependency
- system: word level RNN/ LSTM

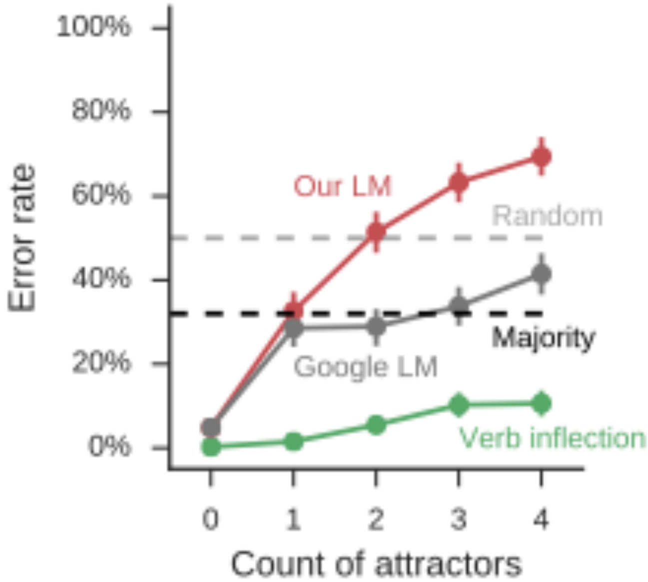
Linzen, Dupoux, & Goldberg, (2016).

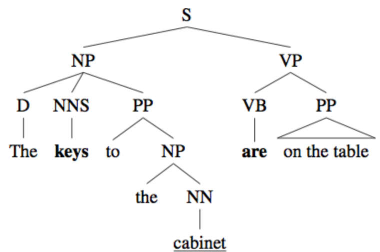
Training a language model
testing on number agreement
error rate: < 5%

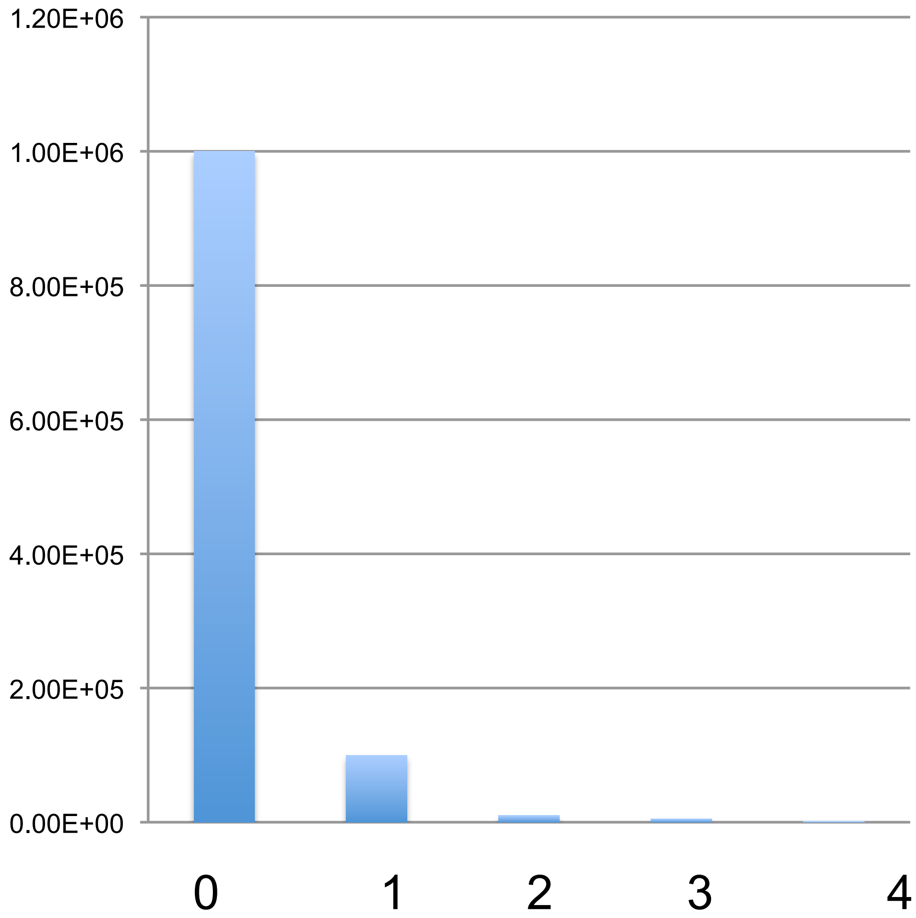


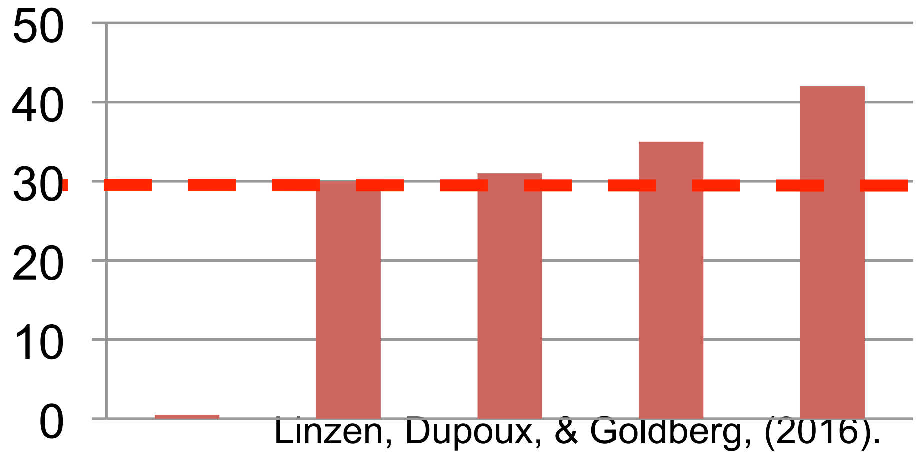
Linzen, Dupoux, & Goldberg, (2016).

Training a language model
testing on number agreement
error rate: < 5%









Linzen, Dupoux, & Goldberg, (2016).

compositionality

- He daxed

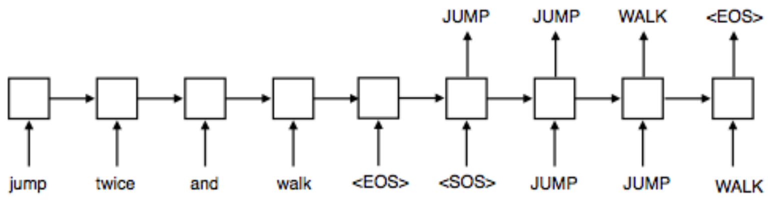
- He daxed twice

- He blicked

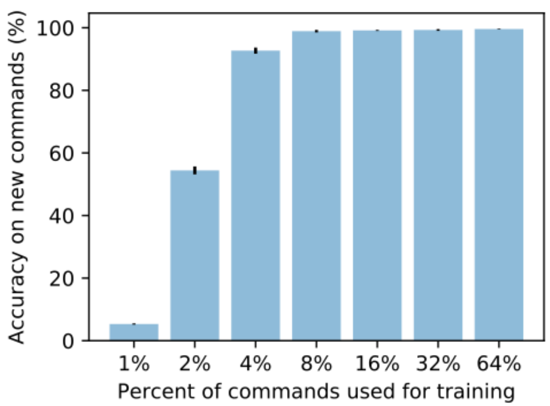
- He blicked twice

?

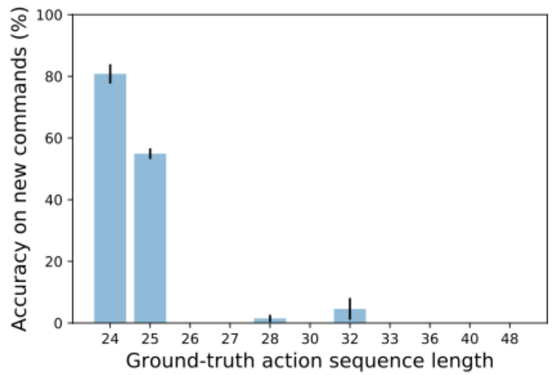
jump	⇒	JUMP
jump left	⇒	LTURN JUMP
jump around right	⇒	RTURN JUMP RTURN JUMP RTURN JUMP RTURN JUMP
turn left twice	⇒	LTURN LTURN
jump thrice	⇒	JUMP JUMP JUMP
jump opposite left and walk thrice	⇒	LTURN LTURN JUMP WALK WALK WALK
jump opposite left after walk around left	⇒	LTURN WALK LTURN WALK LTURN WALK LTURN WALK LTURN LTURN JUMP



Best architecture
2 layers LSTM, 200 hidden



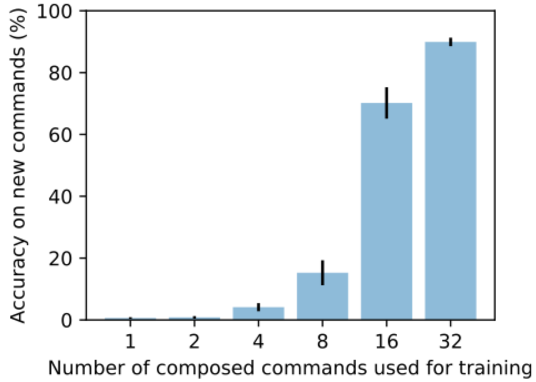
train: random subset



results as a function of action sequence length

Lake & Baroni (2017)

- train with
  - composed actions: (run, walk, look)+ (twice, thrice, left, and V, etc)
  - uncomposed novel action: jump
- test with
  - novel combinations of composed actions: 90% correct
  - combination of basic action: 0.08%-1.2% correct
- Same problem in translation



the machine needs a lot of evidence that the novel action can be composed to generalize

Lake & Baroni (2017)

The Fodor & Pylyshyn (1990) argument

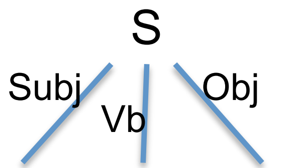
- mental representations have a constituent structure
  - they are not atomic or hollistic but have parts with specific roles
    - eg: the red cow; cheese or desert, Vx R(x), A->B
  - Some constituents can be recursive
    - eg: P. thinks that « M. is nice » -> J thinks that « P thinks that « M is nice » »
- mental processes are structure sensitive
  - eg: combinatorial semantics
    - semantics of « J. loves M. » derived from semantics of « J. », « loves » and « M. »
  - eg: logical inferences:
    - A->B, A   entails B ; this does not depend on the meaning of A and B but on the structure of the representations
- as a result, mental computations are
  - systematic
    - all humans are mortal -> John is moral, Mary is mortal, etc.
    - « Paul likes fruits » grammatical -> « Paul likes fruits » also grammatical, etc
  - productive (achieve discrete infinity)
    - the list of thoughts/sentences is not finite (I can construct new thoughts with old ones)
- connectionist representations have none of these properties

possible solutions

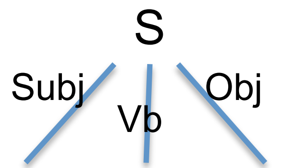
- symbols and rule systems
  – hand-written rules
  – non parametric Bayesian
- tensor products (Smolensky, 1990)

The Smolensky response: filler/roles & tensor products

The problem: representing tree structures



Paul loves Mary



Mary loves Paul

{Subj=Paul, Vb=loves, Obj=Mary}

{Subj=Mary, Vb=loves, Obj=Paul}

The recipe:
→1. define fillers and roles
→2. combine them into a representation with two basic operations:
binding (=), superposition (,)

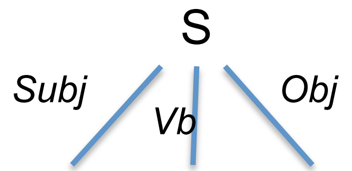
Note: the order of the components does not matter

{Subj=Paul, Vb=loves, Obj=Mary} = {Obj=Mary, Subj=Paul, Vb=loves}
!= {Obj=Paul, Subj=Mary, Vb=loves}

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structure in connectionnist systems. Artificial Intelligence, 46, 159-216.

The Smolensky response: filler/roles & tensor products

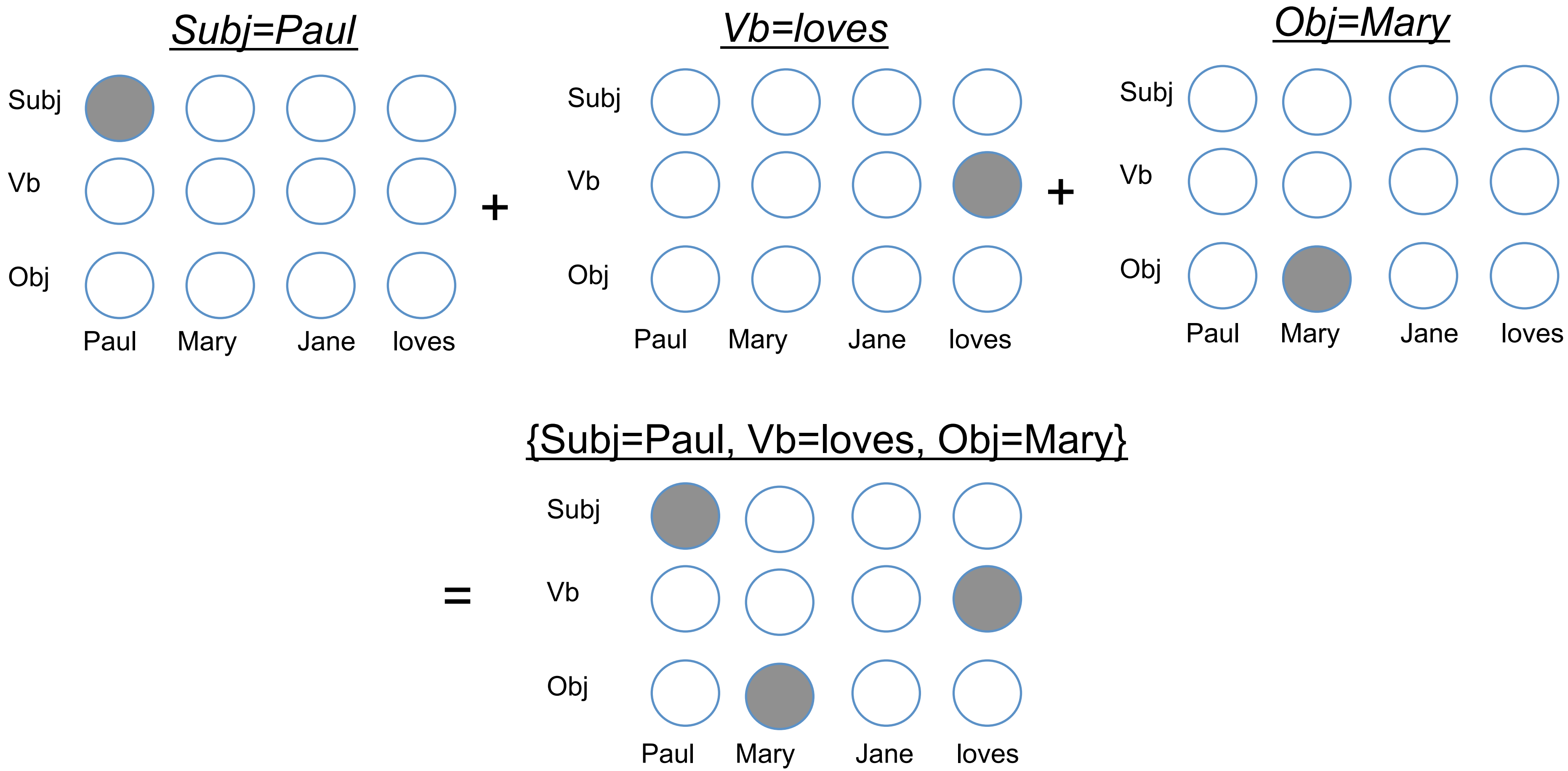
<u>The problem</u>: representing tree structures <u>with numbers</u>



Paul loves Mary

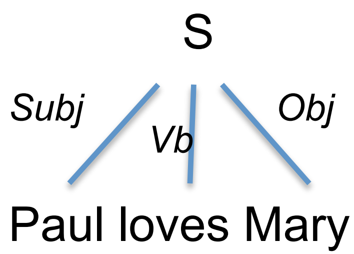
{Subj=Paul, Vb=loves, Obj=Mary}

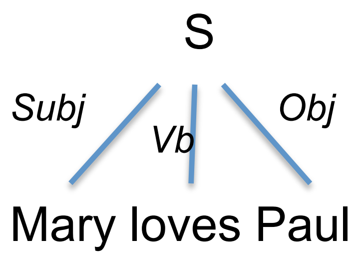
Binding: tensor product
Superposition: addition



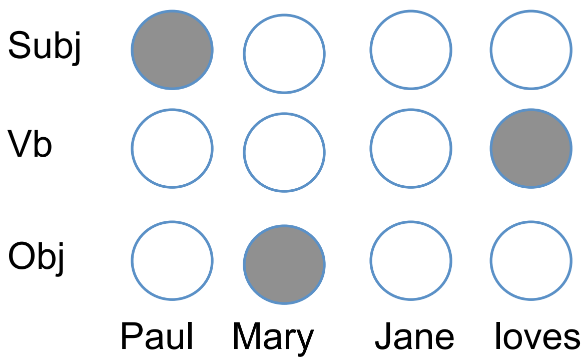
The Smolensky response: filler/roles & tensor products

<u>The problem</u>: representing tree structures <u>with numbers</u>

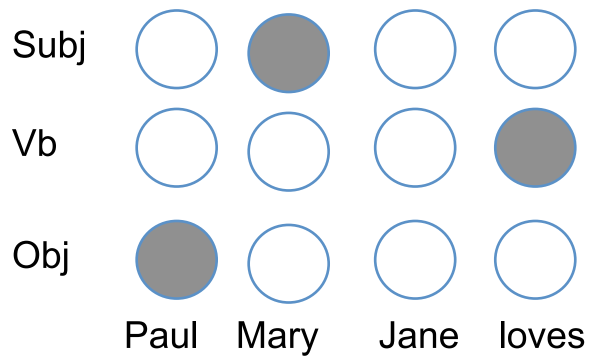




{Subj=Paul, Vb=loves, Obj=Mary}



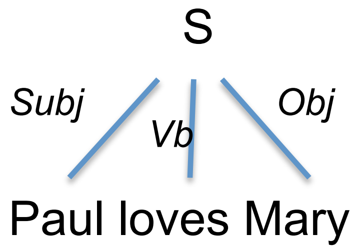
{Subj=Mary, Vb=loves, Obj=Paul}

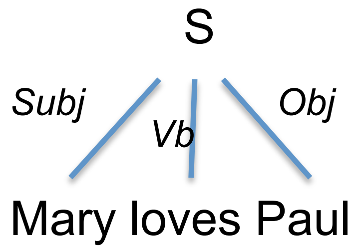


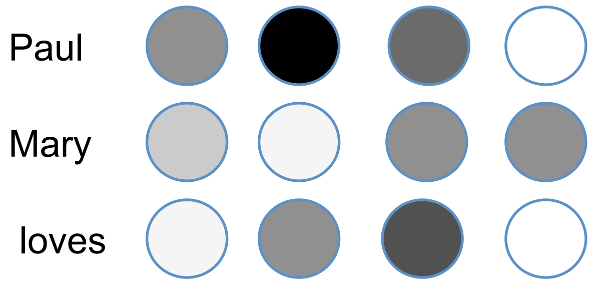
The Smolensky response: filler/roles & tensor products

<u>The problem</u>: representing tree structures <u>with numbers</u>

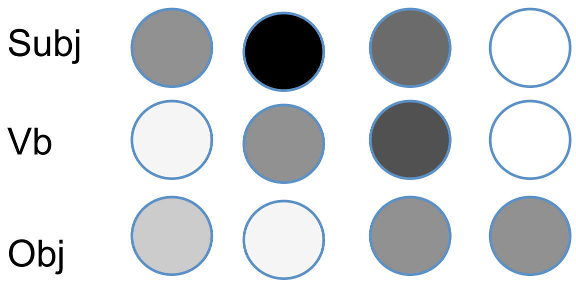
<u>Distributed representations for **fillers**</u>



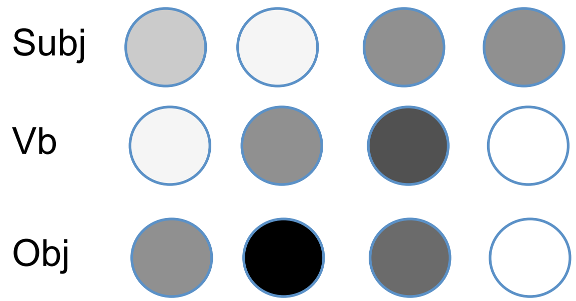




<u>{Subj=Paul, Vb=loves, Obj=Mary}</u>



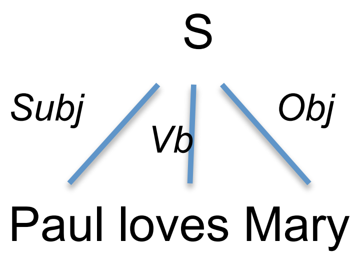
<u>{Subj=Mary, Vb=loves, Obj=Paul}</u>

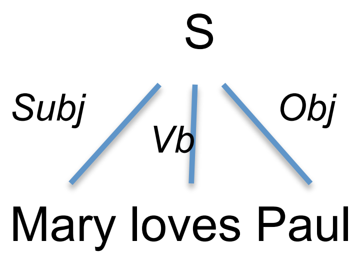


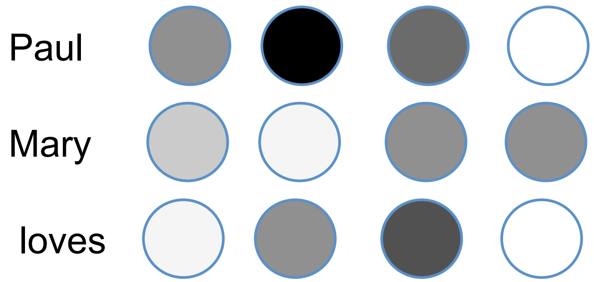
The Smolensky response: filler/roles & tensor products

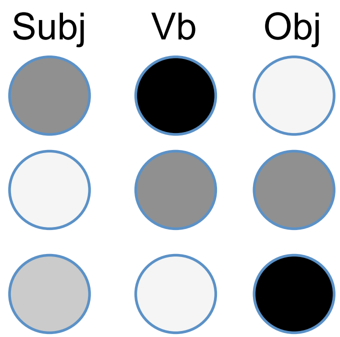
<u>The problem</u>: representing tree structures <u>with numbers</u>

<u>Distributed representations for **fillers & roles**</u>

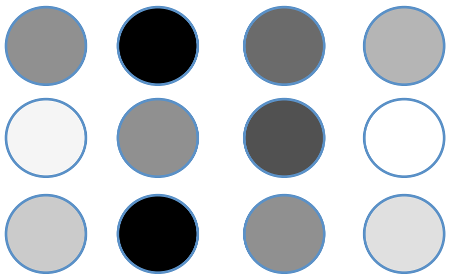




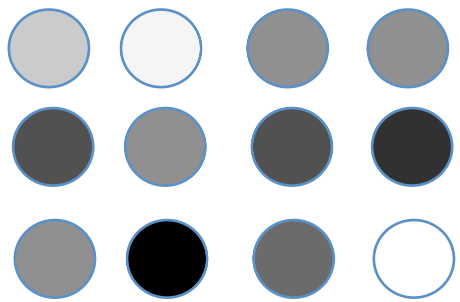




<u>{Subj=Paul, Vb=loves, Obj=Mary}</u>

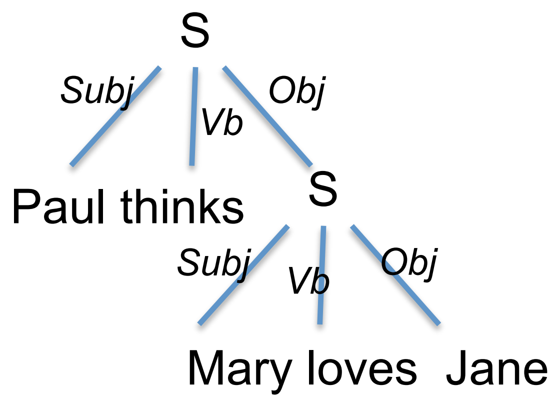


<u>{Subj=Mary, Vb=loves, Obj=Paul}</u>

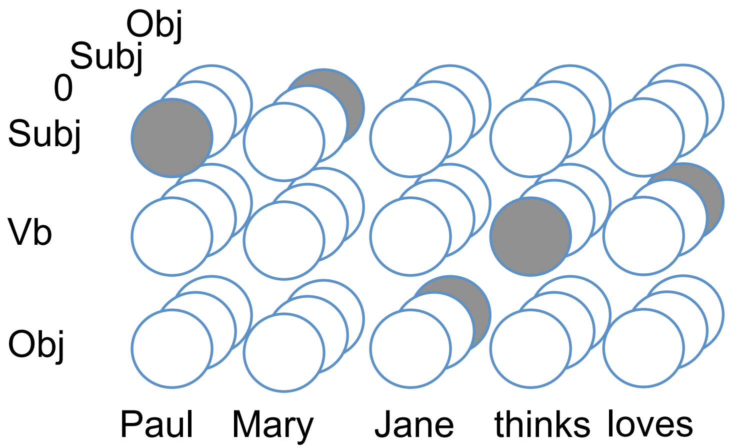


The Smolensky response: filler/roles & tensor products

The problem: representing recursive tree structures with numbers



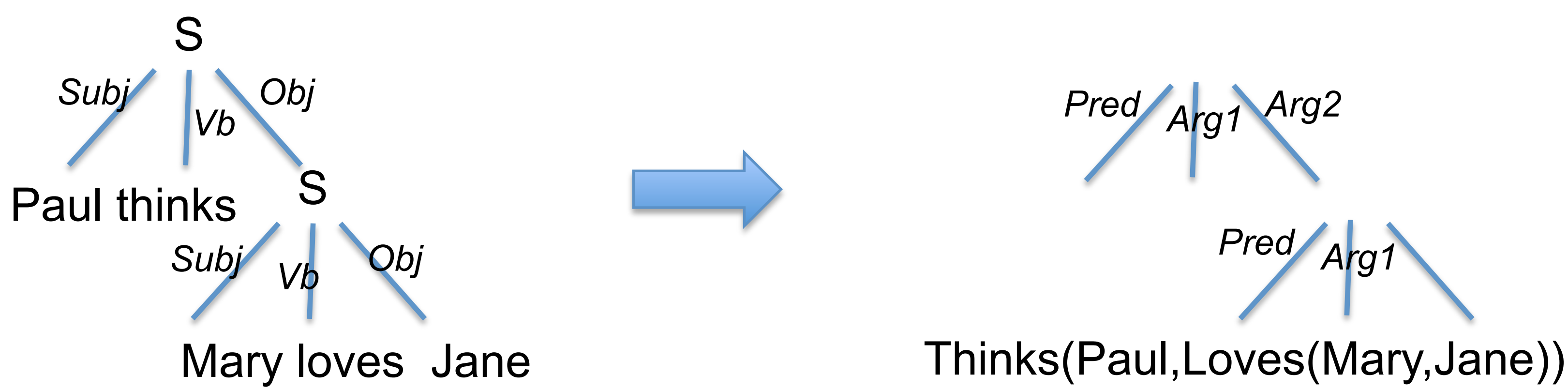
{0_Subj=Paul, 0_Vb=loves, Obj_Subj=Mary, Obj_Vb=loves, Obj_Obj=Jane}

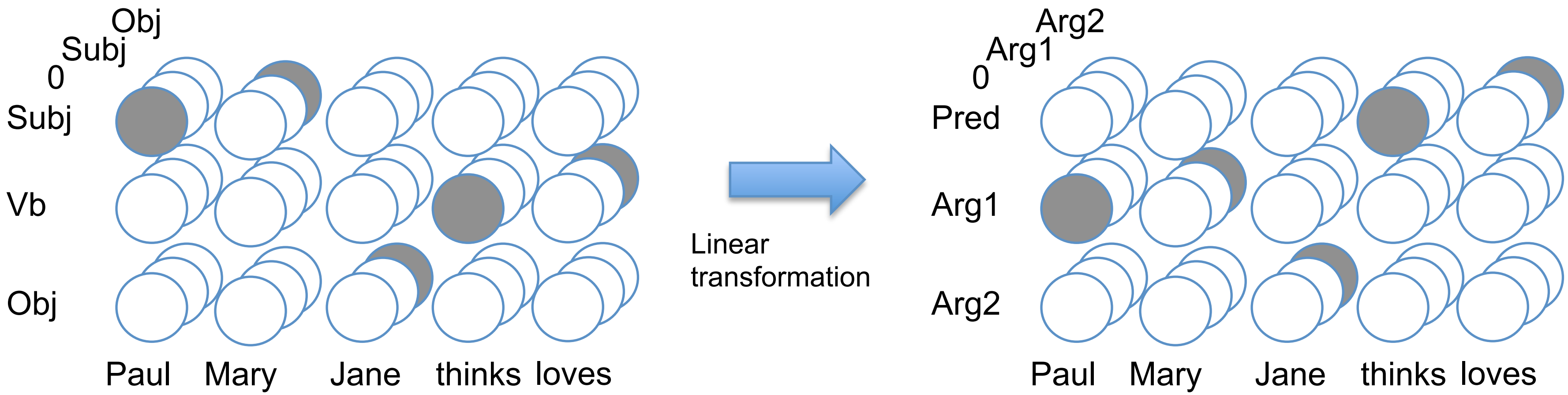


Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structure in connectionnist systems. Artificial Intelligence, 46, 159-216.

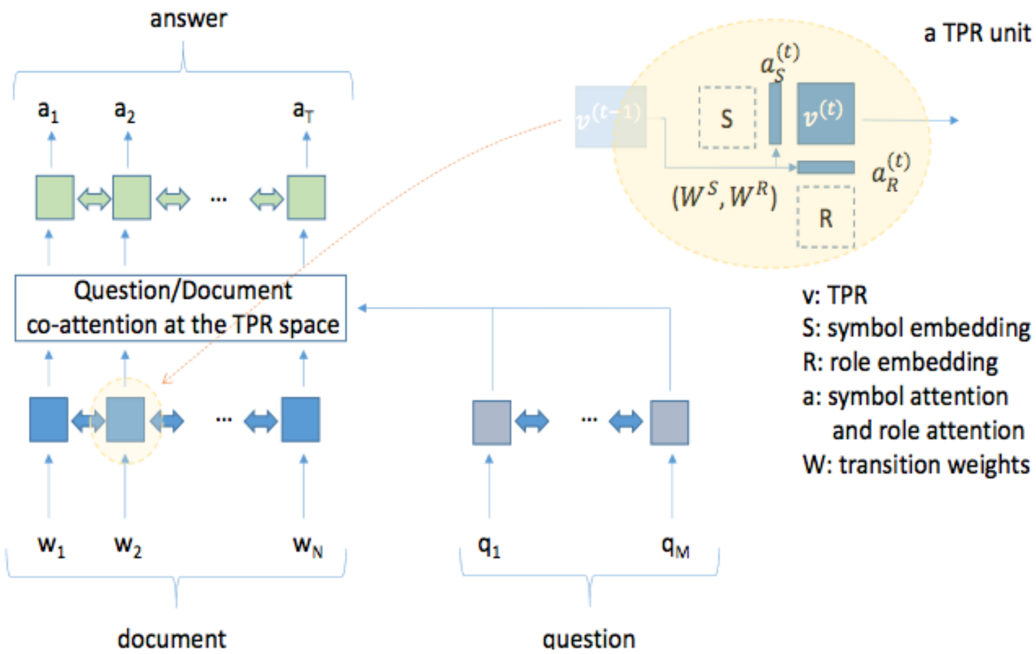
The Smolensky response: filler/roles & tensor products

The problem: manipulating recursive tree structures with numbers





Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structure in connectionnist systems. Artificial Intelligence, 46, 159-216.

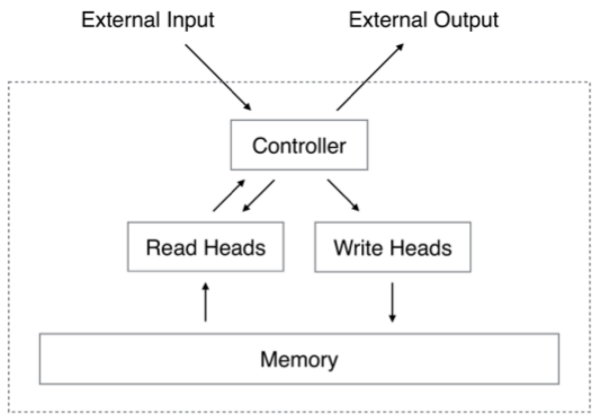


Tensor Product Recurrent Network

Palangi, Smolensky, He, Deng, 2017

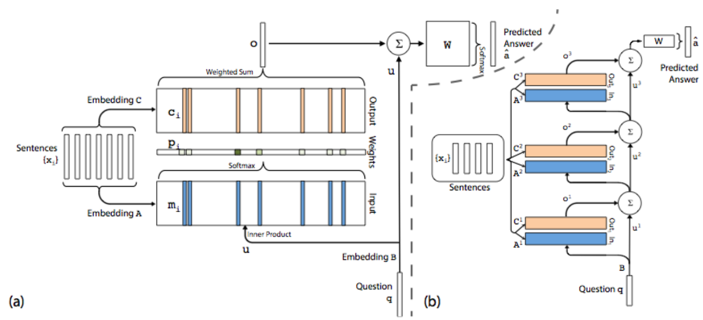
possible solutions

- symbols and rule systems
  - hand-written rules
  - non parametric Bayesian
- tensor products



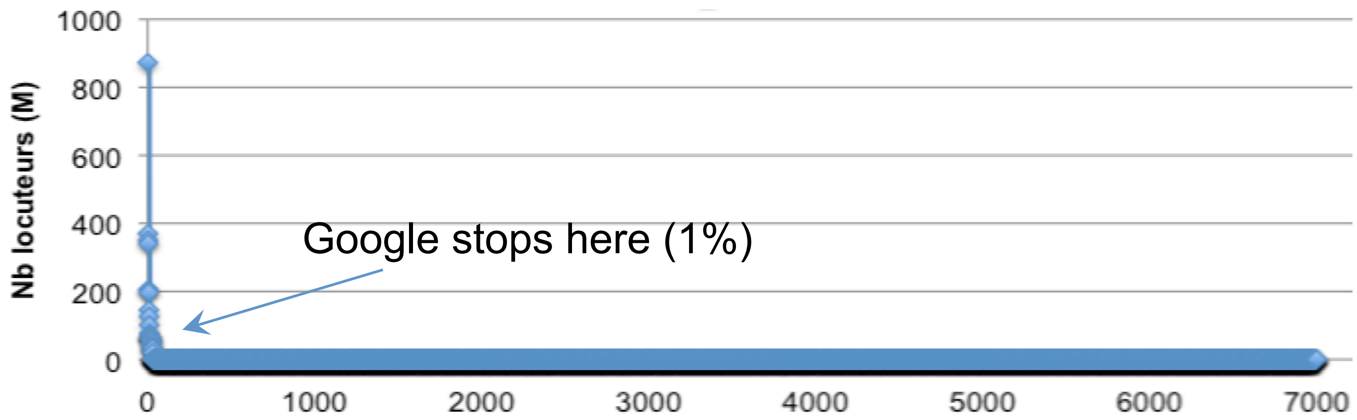
Graves, A., Wayne, G., & Danihelka (2014)

- memory networks?



Weston Chopra & Bordes (2015)

Languages: Paying the price of big data

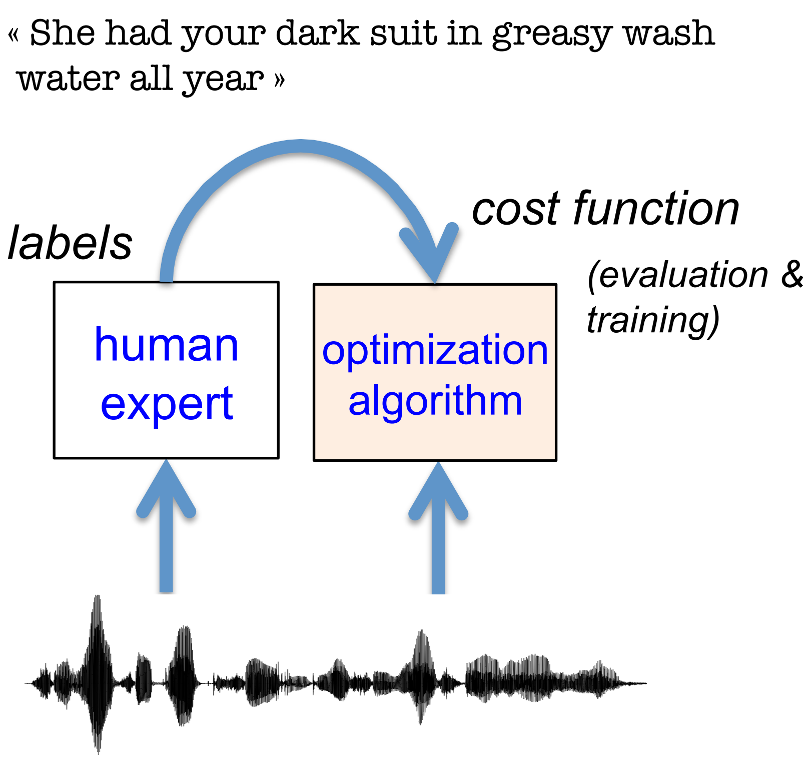


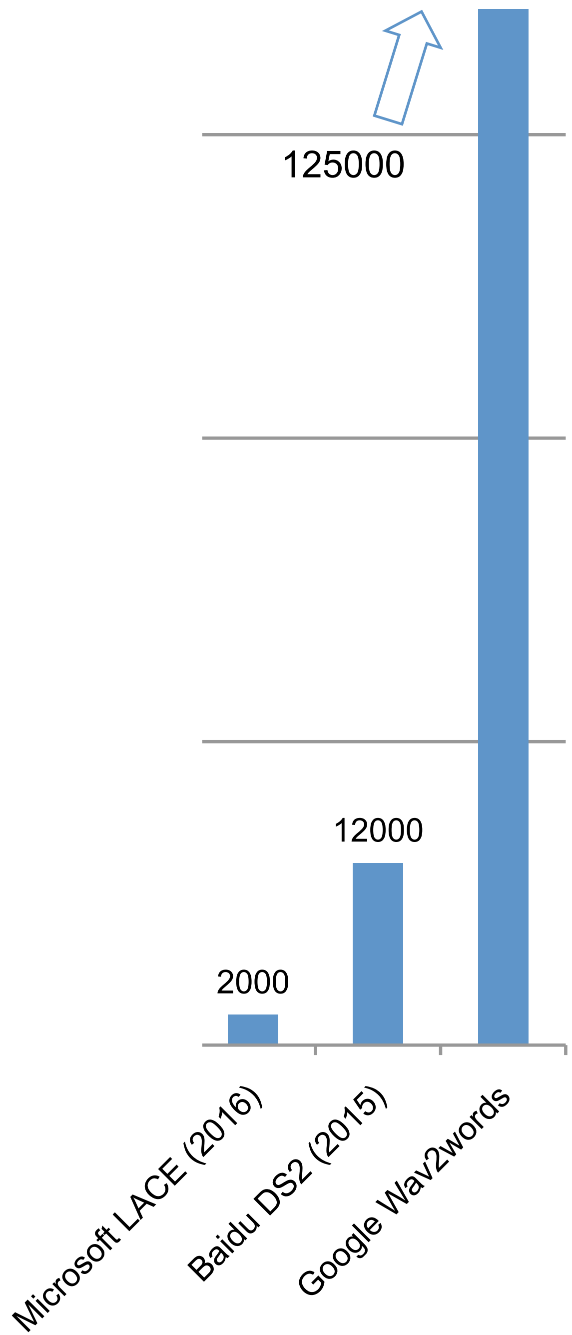
- the long tail
  - 50% of languages have no usable orthography
  - Moroccan Arabic (21M), Nigerian languages: Igbo (25M)
  - Even in English:
    - non standard channels (novel background noise)
    - non standard articulation (children, seniors, etc)
    - non standard pronounciation (sociolects, idiolects, foreing accent)
    - non standard language models (casual versus formal speech)

→ *but infants can learn languages without any orthographic input!*

The data efficiency problem

End-to-end ASR

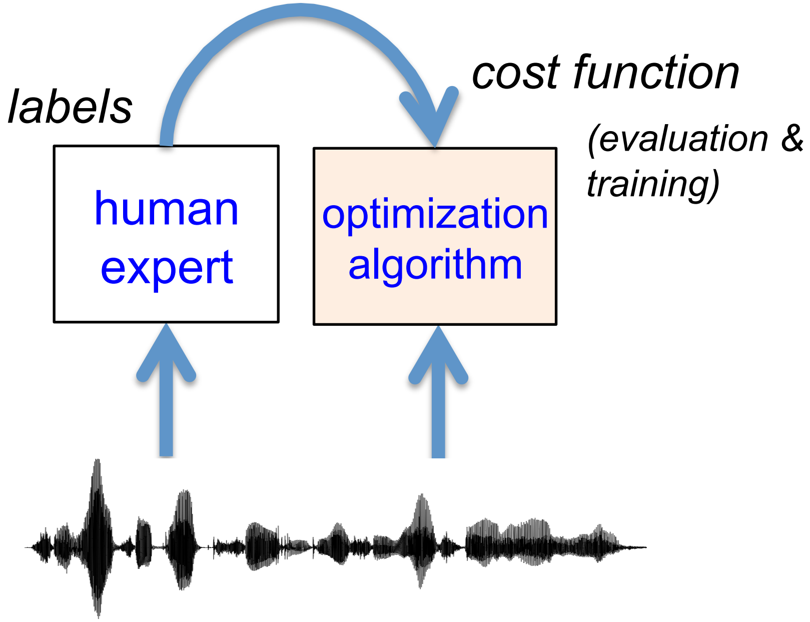




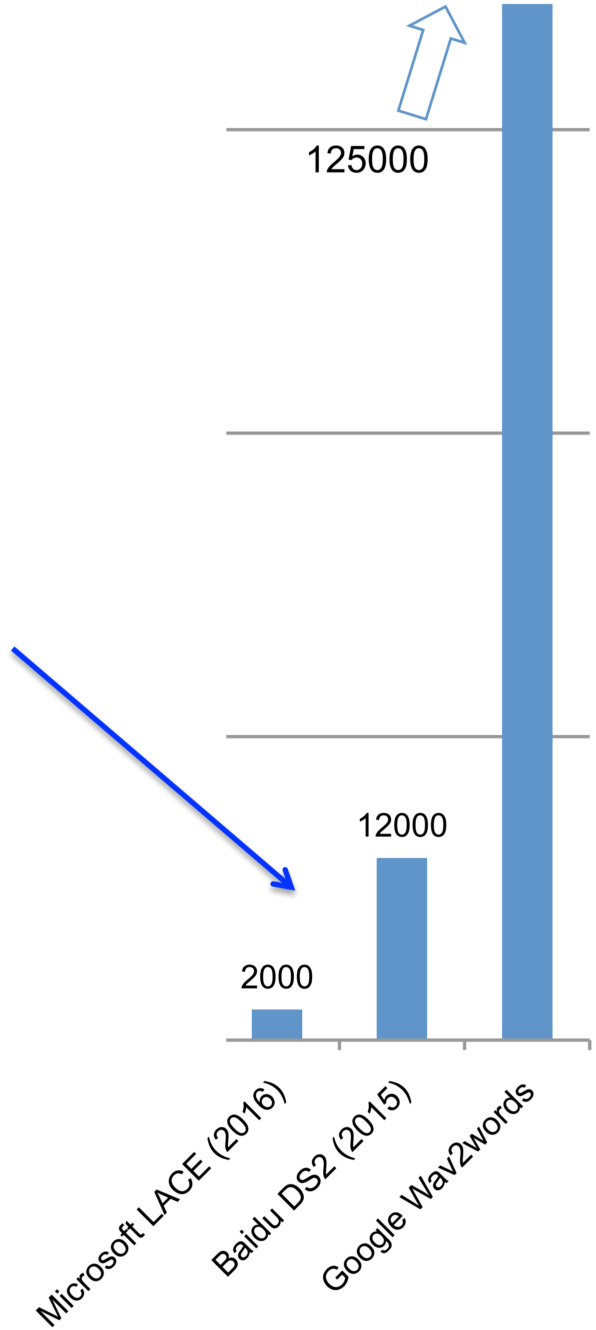
The data efficiency problem

End-to-end ASR

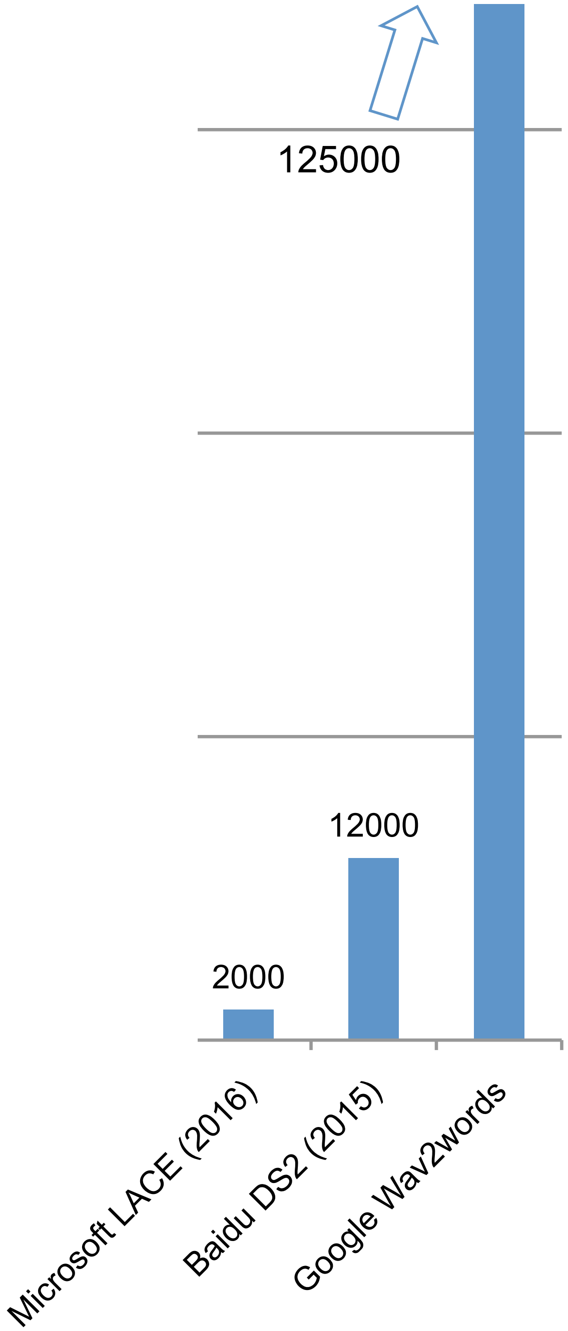
« She had your dark suit in greasy wash water all year »



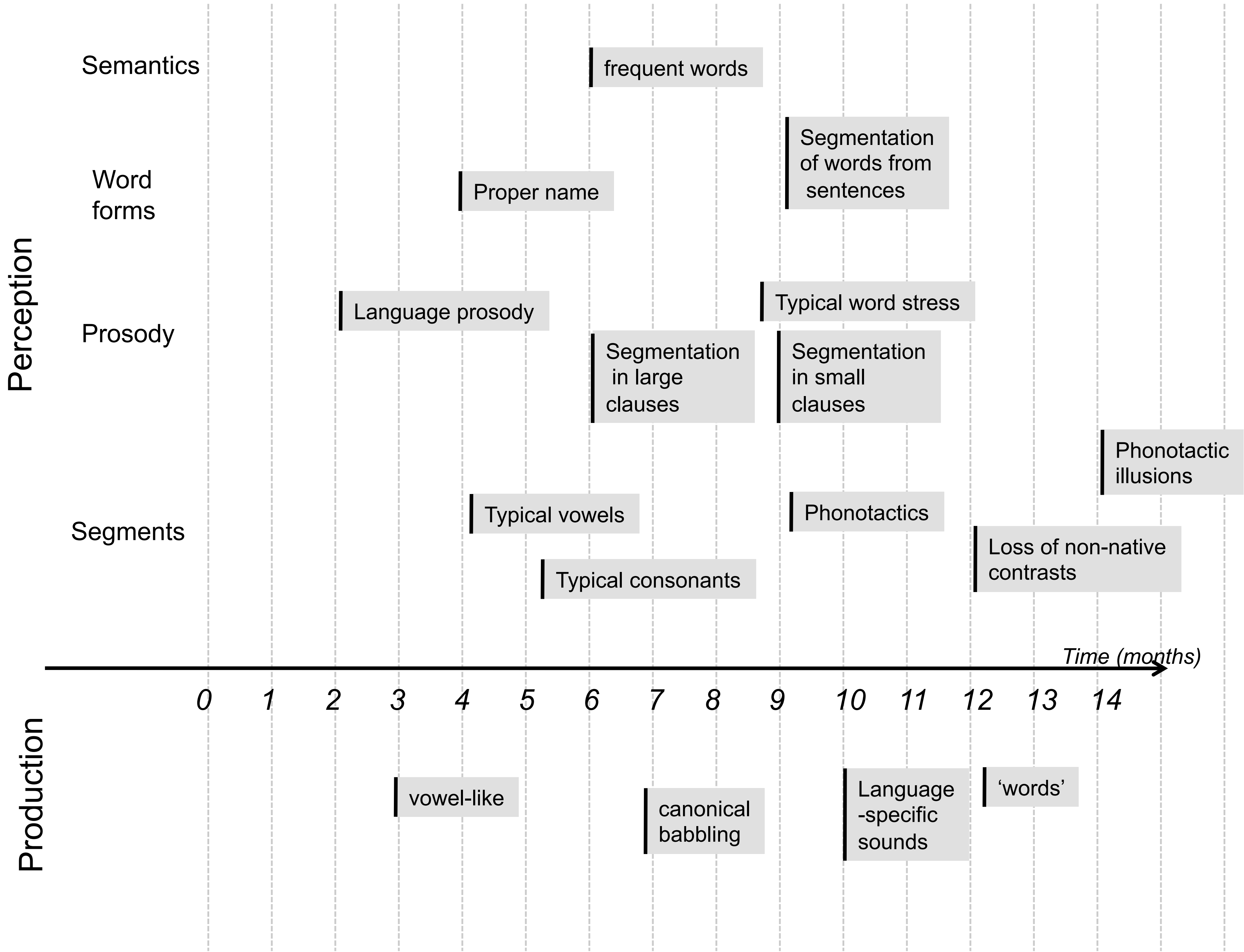
+ 1000000000 words of text for language modeling!
(10000 books)



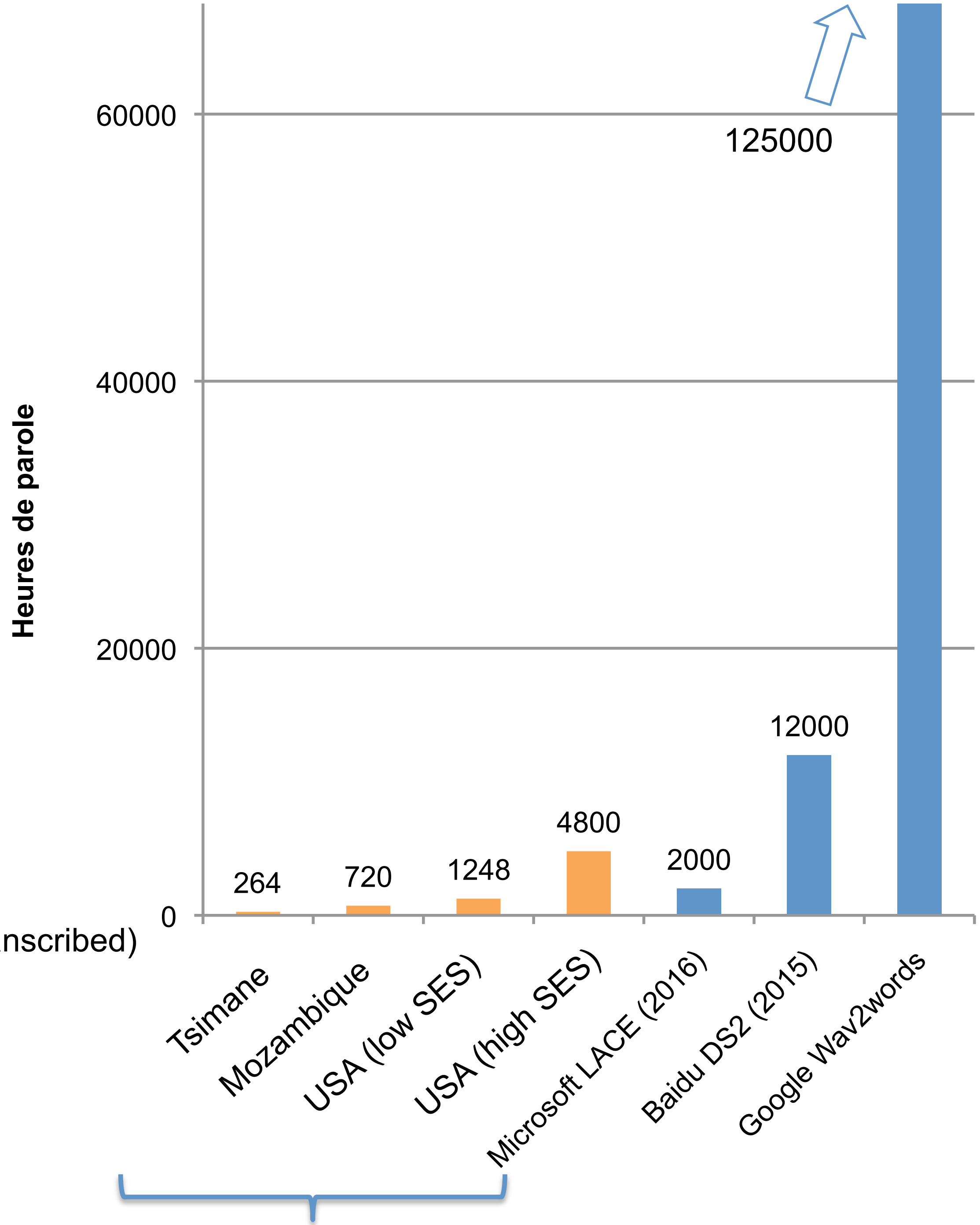
How much data?



Language acquisition trajectories



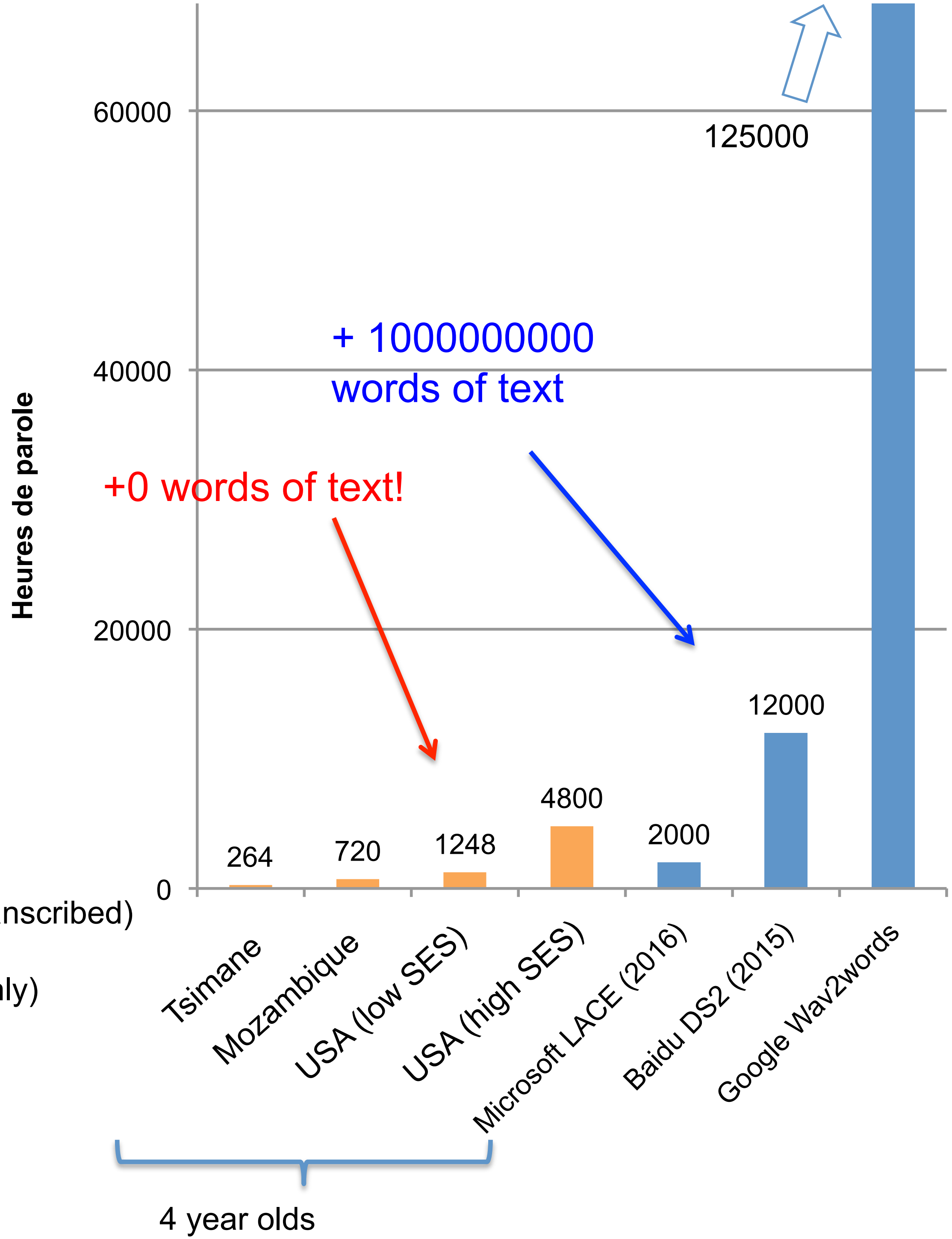
How much data?



Supervised (orthographically transcribed)

Cristia et al, in press

How much data?

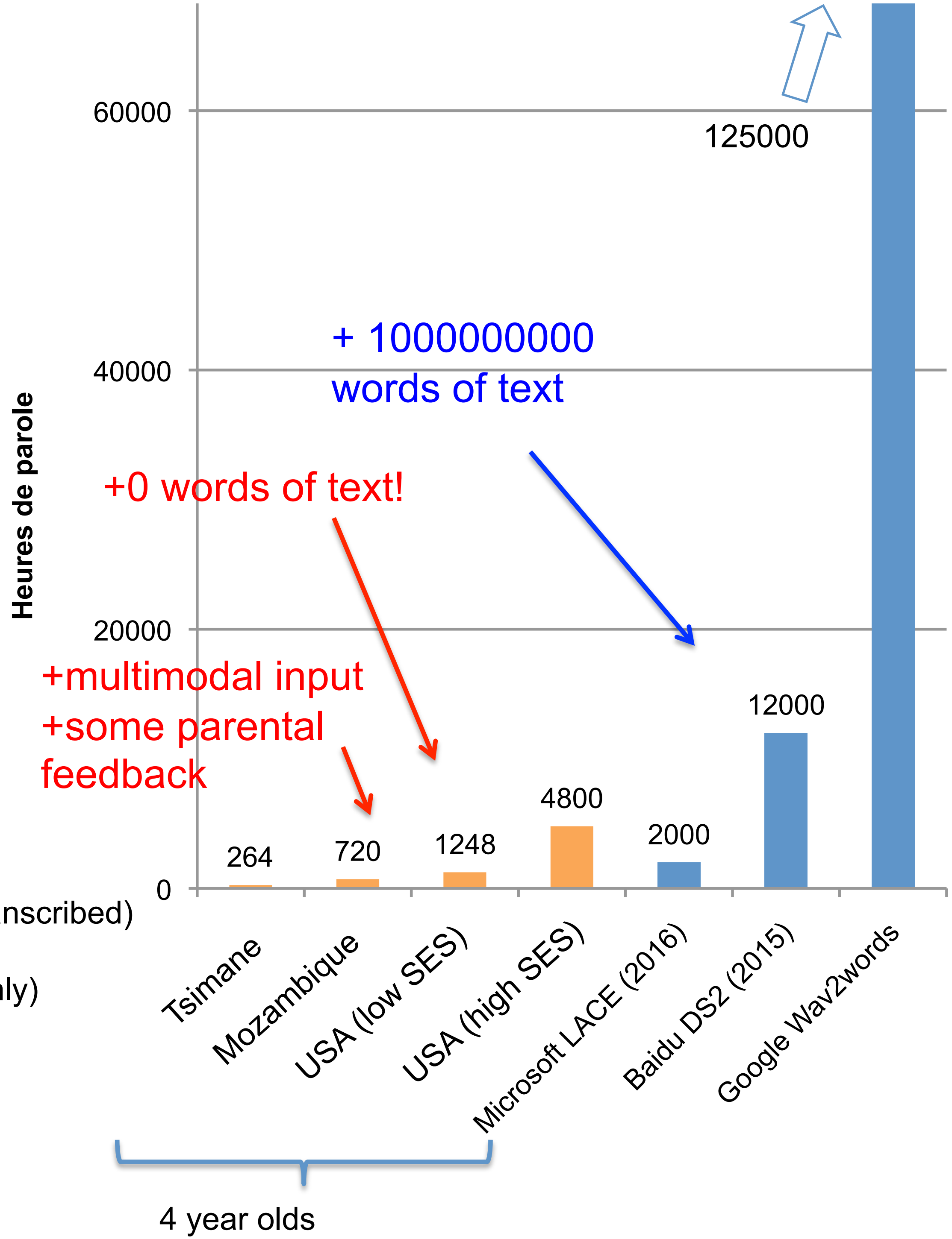


Supervised (orthographically transcribed)

Unsupervised (sensory data only)

Cristia et al, in press

How much data?



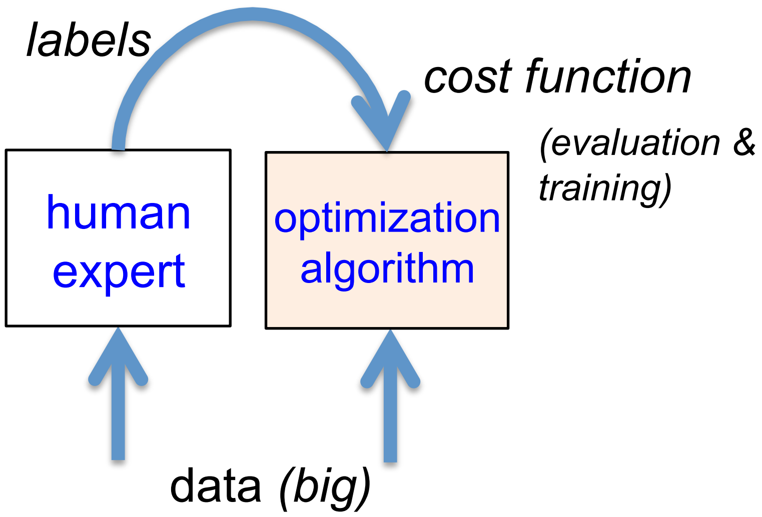
Supervised (orthographically transcribed)

Unsupervised (sensory data only)

Cristia et al, in press

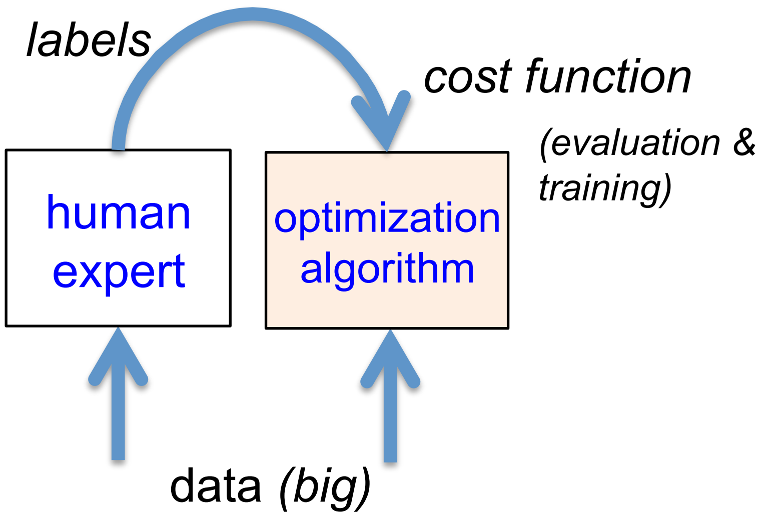
Standard Machine Learning

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



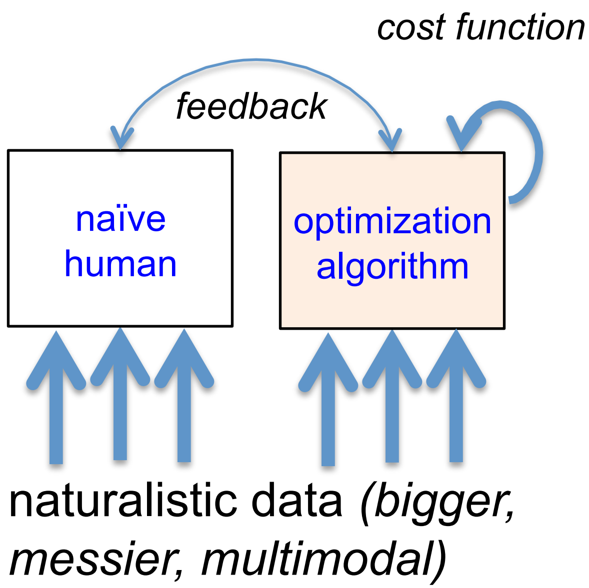
Standard Machine Learning

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



Human-like Machine Learning

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



Developmental Artificial Intelligence

Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an <u>appropriate course of education</u> one would obtain the adult brain.

Turing (1950)

Not standard machine learning!

(my emphasis)

Developmental Artificial Intelligence

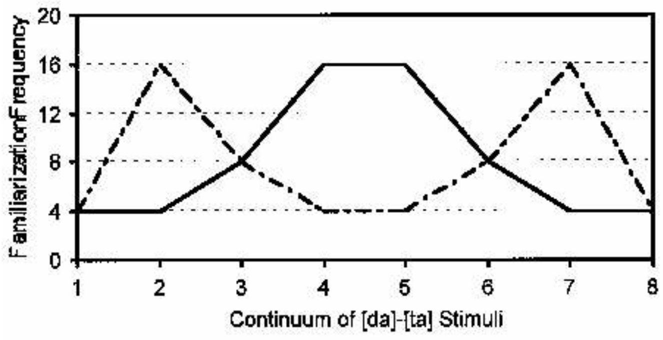
- language learning=learning a compact representation for the input (Kelley, 1967, de Marcken, 1996)

- language learning=learning to translate between surface input to underlying concepts (Siklossy, 1968; Siskind, 1996)

- language learning=learning to communicate (Bruner 1975)

In infants

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



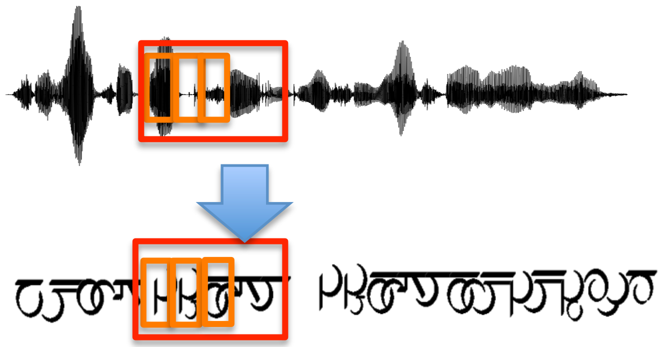
  – word segmentation (Saffran et al. 1996)

gakotibunitagakotiralépy

  – 'algebraic' rule learning (Marcus, et al., 1999)

ABB vs ABC

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)

- *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

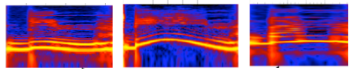
www.zeropeech.com

Track 1. ideas

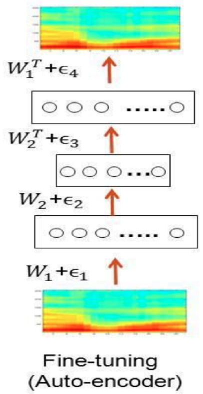
Evaluation: psychophysical ABX task

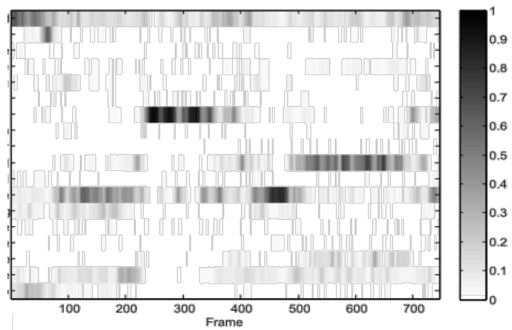
A B X

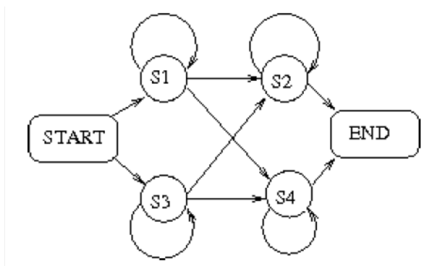
ba_{T1} ga_{T1} ga_{T2}



- 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, binarized DNNs (e.g. Myriam & Salvi 2017)



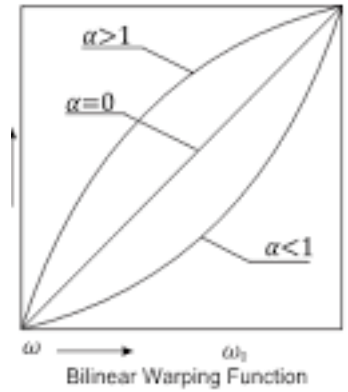




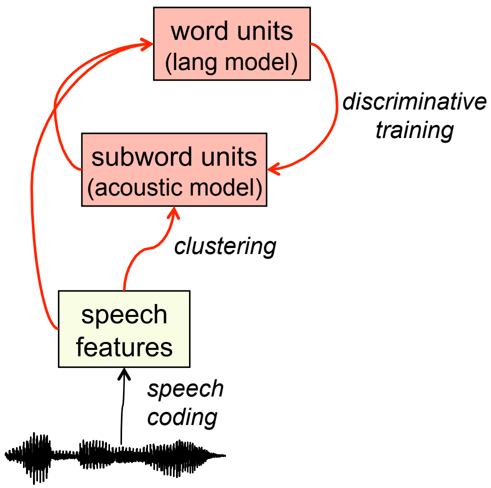
Main idea: information compression

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

Track 1: more ideas



- Talker Normalization
  - remove linguistically irrelevant information (talker identity, channel, emotional states, etc).
  - method:
    - features level normalization: VTLN (Chen et al. 2015)
    - model level normalization: fMMLR, etc (Heck et al 2016; 2017)
- Using Track 2 to help Track 1
  - principle: minimal pairs are rare
    - same words → 100% phonemes same
    - different words → 90% phonemes are different
  - method: top-down signal
    - Correspondance AE (Kamper et al 2015)
    - Siamese network (Synnaeve et al., 2014)



Track 1: Leaderboard

[Eng, Xit]

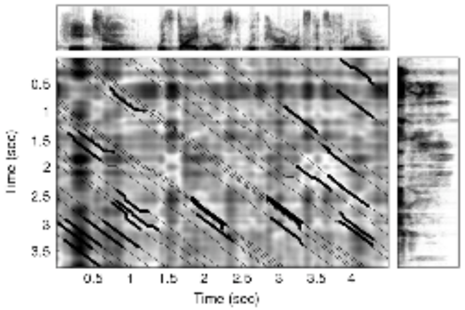
	Paper	Compression	Normalization	Contrast	Articulation	Result (%supervized)
ZR15	Chen et al. 2015	DPGMM	VTLN			[95%, 56%]
ZR15	**Heck et al. 2016**	**DPGMM**	**VTLN+LDA**			**[100%, 72%]**
ZR15	Renshaw et al. 2015			Corr AE		[58%, 49%]
ZR15	Thiollière et al. 2015			Siamese		[84%-58%]
ZR17	**Heck et al 2017**	**DPGMM**	**VTLN+LDA+MLLT+ +fMMLR**			**[109%, 74%]**

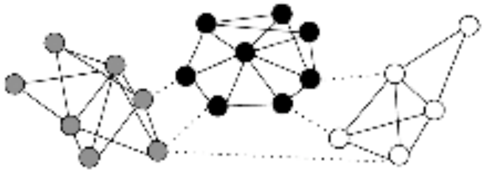
[Ger, Wol]

Attention: these are NOT phonemes!

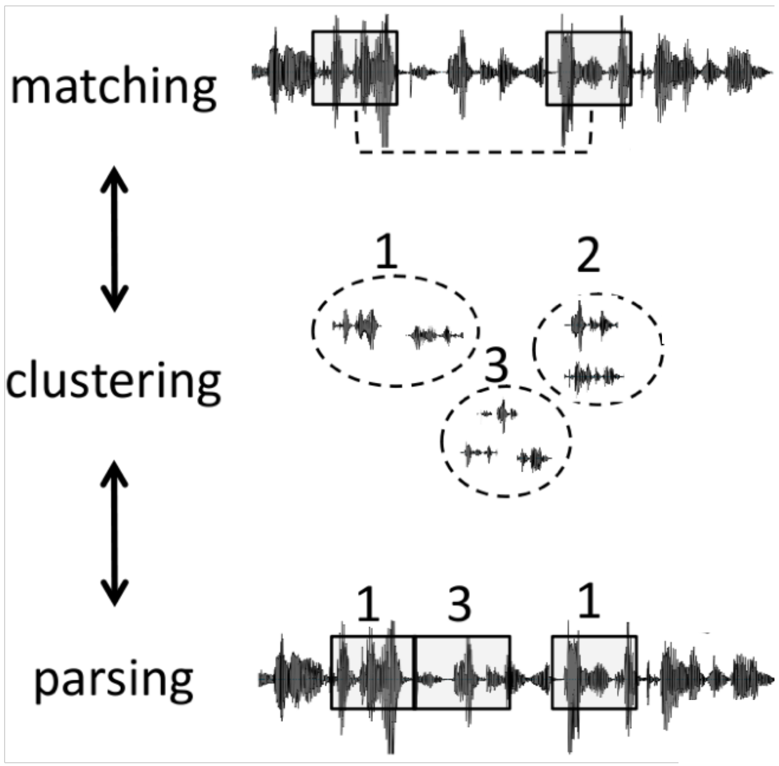
Track 2

task





(Viterbi decoding)



Finding repeated patterns

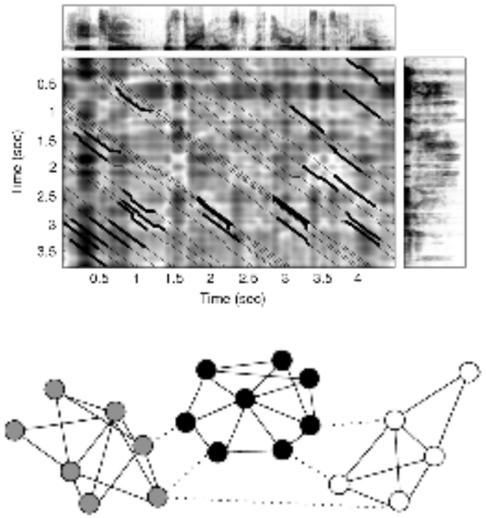
Clustering these patterns

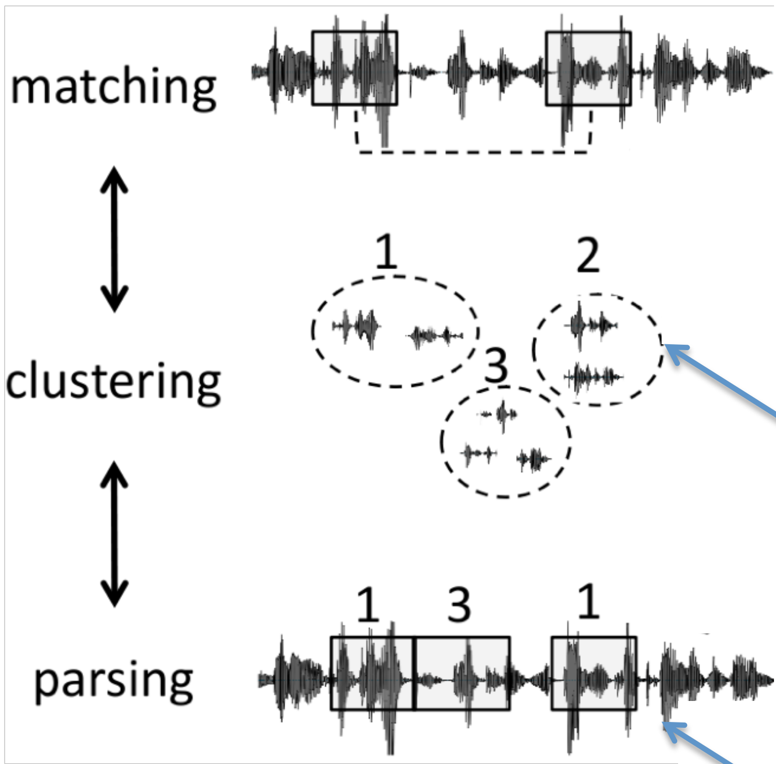
Using known words to find new ones

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

Track 2: evaluations

<u>task</u>



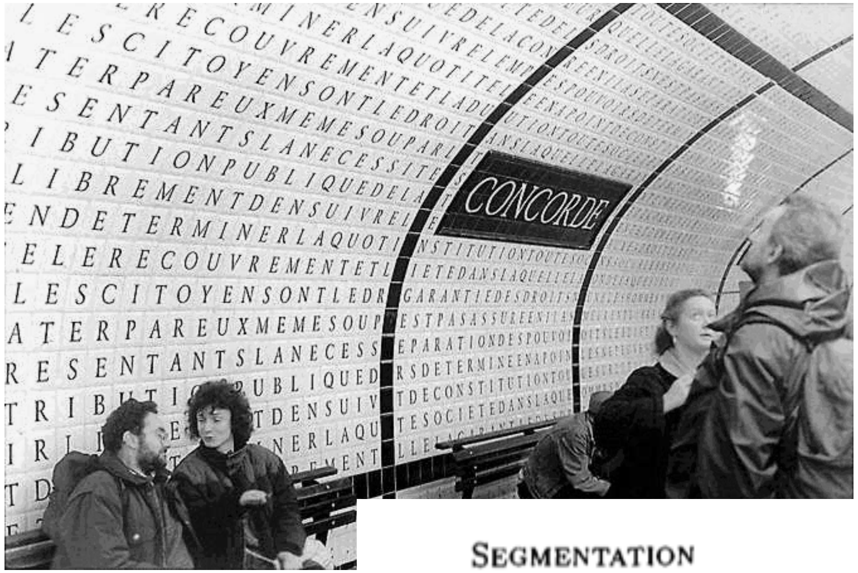


compare with 'gold' words

(Viterbi decoding)

compare with 'gold' boundaries

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

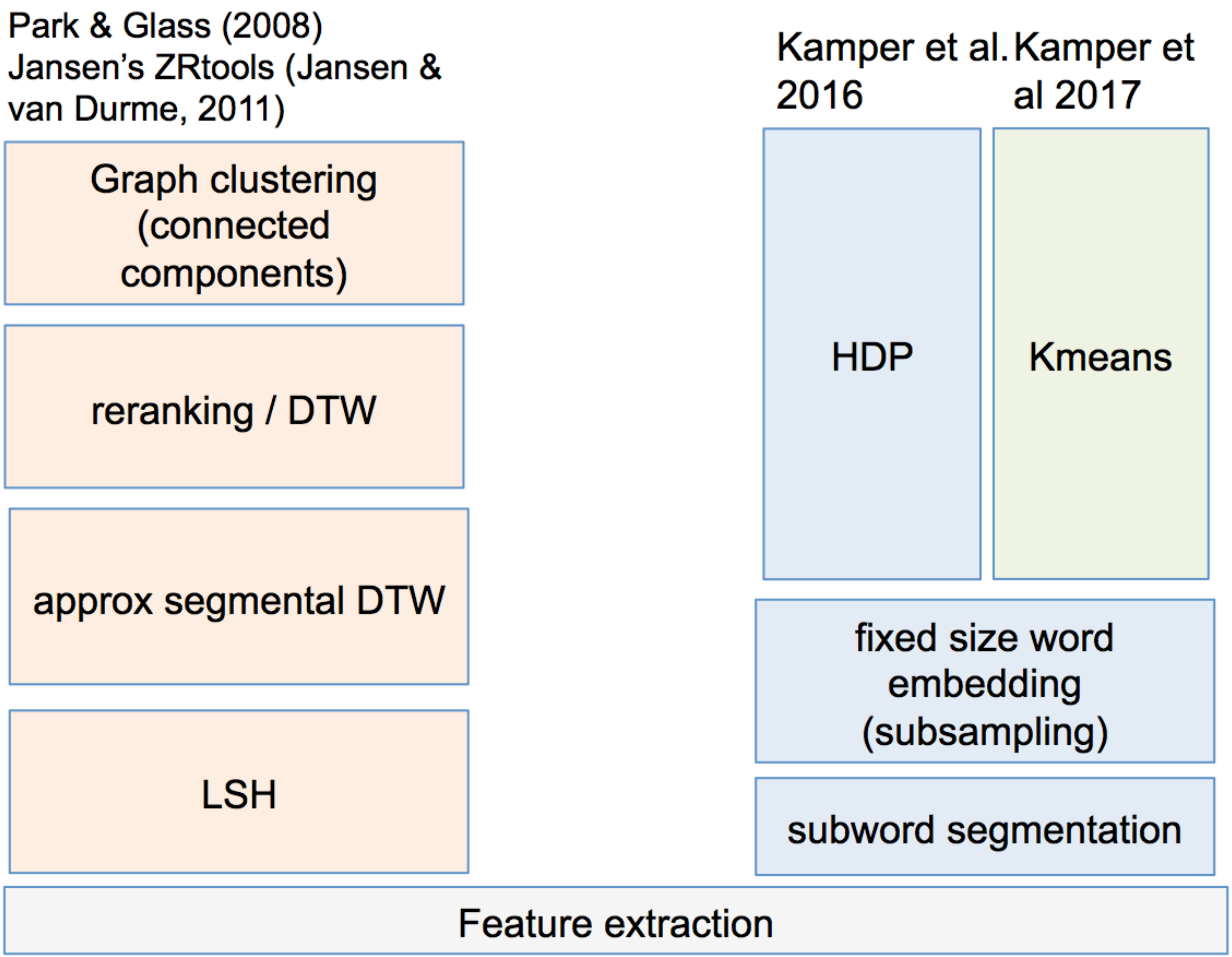


the case of text

- <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 $\boxed{\text{thekitty}}$	1 do 2 $\boxed{\text{thekitty}}$ 3 you	1 3 5 $\boxed{2}$	25+10=35
see $\boxed{\text{thekitty}}$	4 like 5 see	5 $\boxed{2}$	
do you like $\boxed{\text{thekitty}}$		1 3 4 $\boxed{2}$	
do you see the kitty	1 do 2 the 3 you	1 3 5 2 6	26+13=39
see the kitty	4 like 5 see 6 kitty	5 2 6	
do you like the kitty		1 3 4 2 6	
do $\boxed{\text{yousee}}$ the kitty	1 do 2 the 3 you	1 $\boxed{7}$ 2 6	33+12=45
see the kitty	4 like 5 see 6 kitty	5 2 6	
do you like the kitty	$\boxed{7}$ $\boxed{\text{yousee}}$	1 3 4 2 6	

- <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)


Park & Glass (2008)
Jansen's ZRtools (Jansen & van Durme, 2011)
Graph clustering (connected components)
reranking / DTW
approx segmental DTW
LSH
Kamper et al. 2016
Kamper et al 2017
HDP
Kmeans
fixed size word embedding (subsampling)
subword segmentation
Feature extraction

Track 2: leaderboard

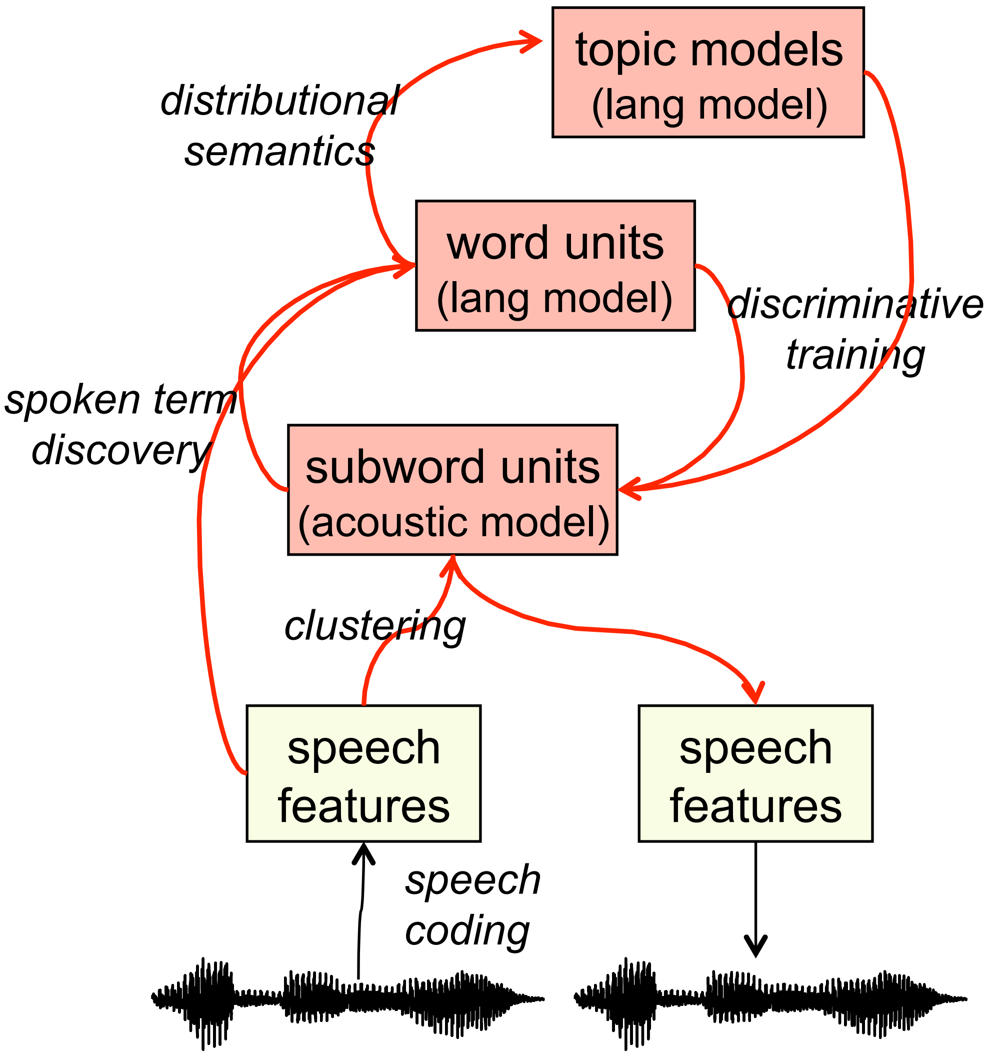
[Eng, Xit]

	Paper	Method	Result (Boundaries) (%supervized)	Results (Type) (%supervized)
ZR15	Rasanen et al. 2015	SyllOsc	[48%, 36%]	[21.1%, 9%]
ZR15	Lizinski et al. 2015	FDPLS-DTW+ConnComp	[36%, 40%]	[3.4%, **40%**]
ZR15	Kamper et al. 2016	Word embeddings+ DPGMM	**[68%, 50%]**	**[30%**, 24%]
ZR17	Garcia-Granada et all. 2017	Out of domain ASR + DTW	[30%,32%]	[13%,**15%**]
ZR17	Kamper et al. 2017	Word embedding + kmeans	**[58%,49%]**	**[24%,**14%**]**

[Ger, Wol]

→ *Still long way to go for word discovery*

Next step



ZR 2021? Language modeling from speech

see: audio word-to-vec (Chung & Glass, 2018),

ZR 2019. Unsupervised speech synthesis

see: Muthukumar & Black, 2014; Scharenborg et al. 2018, JSALT 2017; Chorowski, RJ Weiss, S Bengio, A Oord (2019)

Developmental Artificial Intelligence

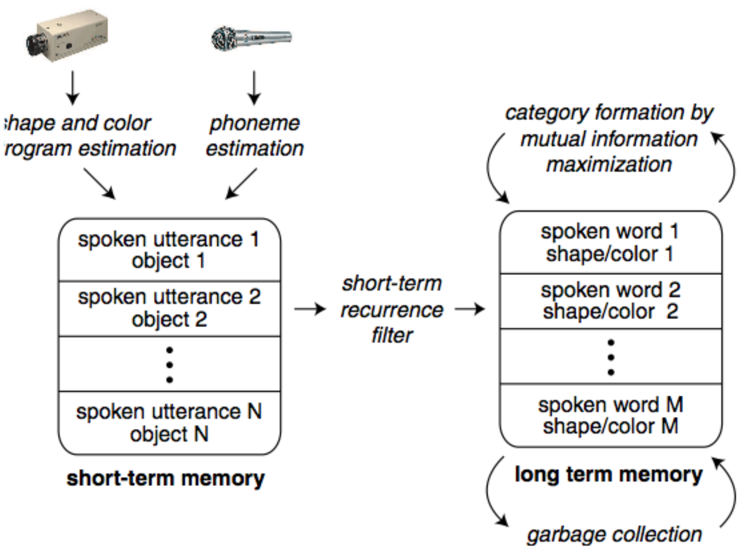
- language learning=learning a compact representation for the input (Kelley, 1967, de Marcken, 1996)

- language learning=learning to translate between surface input to underlying concepts (Siklossy, 1968; Siskind, 1996)

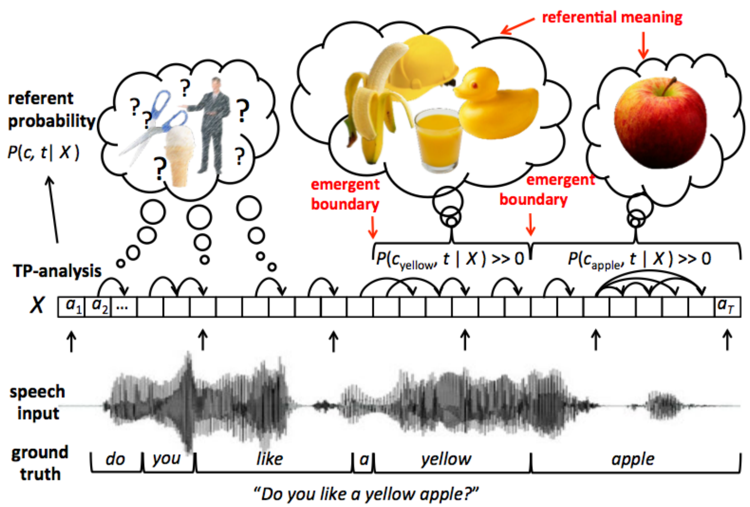
- language learning=learning to communicate (Bruner 1975)

word learning

Cross situational learning
learning the correspondance between words and meaning across many examples



Roy & Pentland, 2002



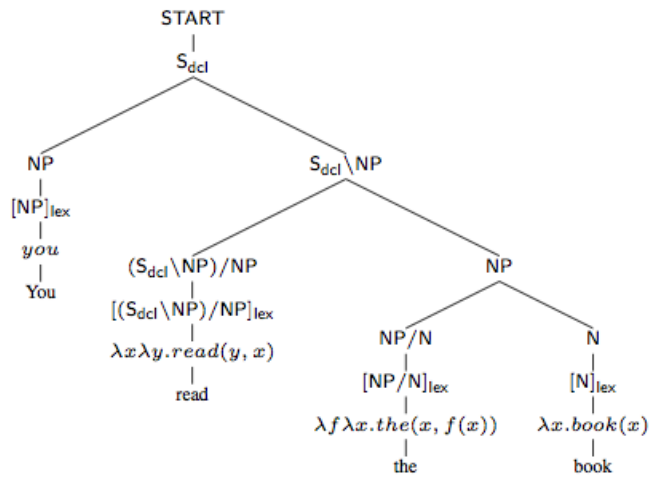
Rasanen & Rasilo, 2005

learning sentence parsing

Utterance : you have another cookie

Candidate Meanings $\begin{cases} have(you, another(x, cookie(x))) \\ eat(you, your(x, cake(x))) \\ want(i, another(x, cookie(x))) \end{cases}$

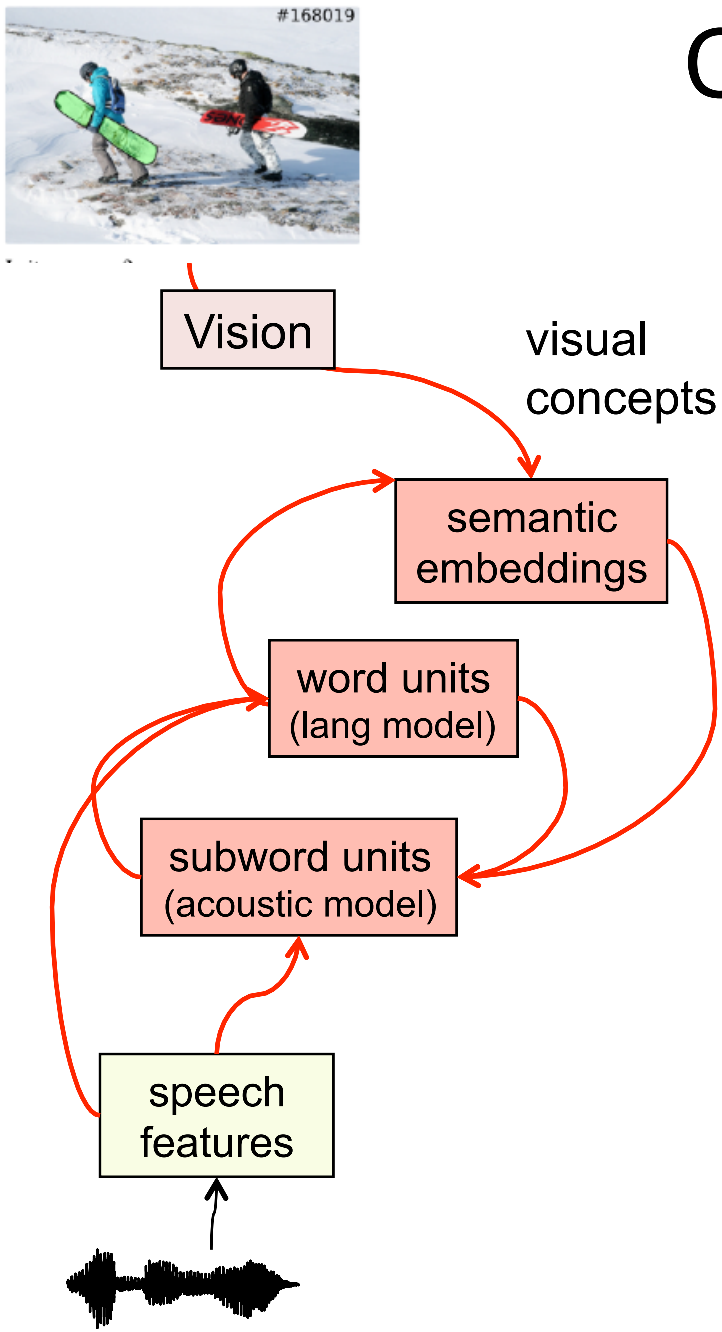
$$\begin{aligned} \text{You} &\vdash \mathsf{NP} : you \\ \text{read} &\vdash \mathsf{S\backslash NP/NP} : \lambda x \lambda y.read(y, x) \\ \text{the} &\vdash \mathsf{NP/N} : \lambda f.the(x, f(x)) \\ \text{book} &\vdash \mathsf{N} : \lambda x.book(x) \end{aligned}$$



see also Siskind 1996

Kwiatkowski et al 2012
Abend,et al, 2017

Cross modal grounding



Cross-modal context

→ *pragmatic use of language induce correlations between context and content*

a. Descriptive/Referential
- people talk about the here and now
- infants: cross-situational learning (Yu & Smith, 2007)
- computers: (Roy & Pentland, 2002, Boves, Ten Bosch, & Moore, 2007, Siskind, 1996, Harwath & Glass, 2015,2016,2017; Chrupala et al. 2017; Alishahi et al, 2017, etc)

b. Imperatives
- people use language to program each other
- infants: (Schaffer, H. R., & Crook, 1980; Rheingold, et al, 1987, etc)
- adults (Tsividis et al 2017)
- robots (*Saponaro et al 2017*); videogames (Kaplan et al 2017)

c. Questions/Answers
- people use language to exchange information
- infants: Jusczyk, et al (2003), etc
- computers: (Liang et al. 2011; Sontoro et al 2017, …)

Developmental Artificial Intelligence

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

language as a social ability

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

Learning language with impoverished input: ‘creoles’

- ‘pidgins’: emerge in adult communities from different language backgrounds
  - inconsistent word order, no prefix/suffix, no tense marking; only one proposition at a time,
  - no consistent way of indicating who did what to whom
- ‘creoles’: emerge among infants who are immersed in pidgins (eg;hawaian creole)
  - emergence of grammatical words (auxiliaries, case marking, relative pronouns, etc)

Language creation with zero input: Nicaraguan Sign Language

In 1978, Nicaraguan governement opens the first public school for deaf children. The children did not have a sing language, but had develop some element of ‘home signs’ in their homes

Example English home signer:

http://goldin-meadow-lab.uchicago.edu/Images/shovel.mov

At school, during the breaks, they develop a new sign language
(very different from signed Spanish)

http://www.pbs.org/wgbh/evolution/library/07/2/l_072_04.html

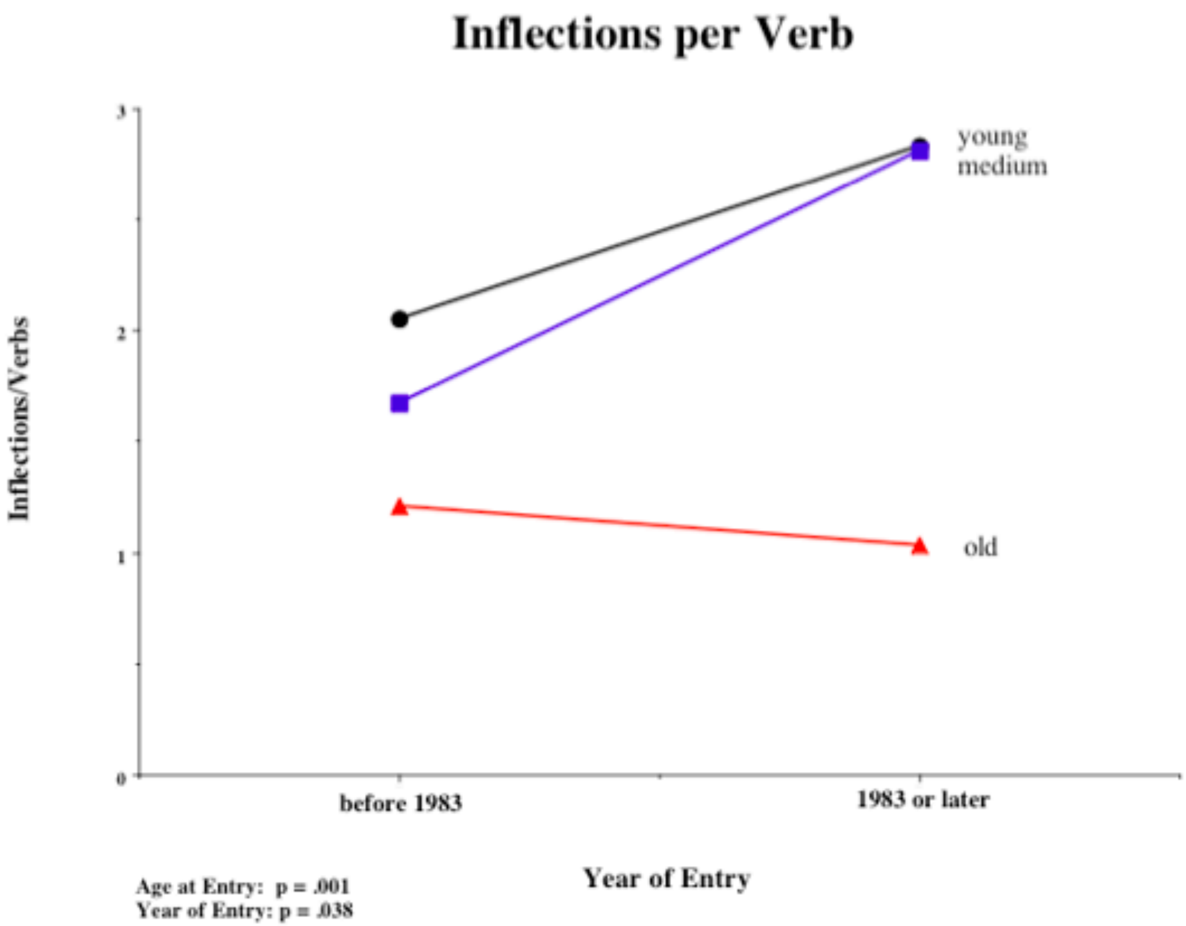


Figure 2. The number of inflections per verb is greater overall for signers who entered the community in 1983 or later, and for signers who were exposed to the language at a *young* or *medium* age. The *young* and *medium* Age at Entry signers are particularly affected by a later Year of Entry.

Inflection:

He likes me.

(as opposed to “he like me”)

→effet d’age et de génération

Senghas & Coppola 2001

language emergence in machines

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

Presumably the child-brain is something like a notebook as one buys it from the stationer's. <u>Rather little mechanism,</u> and <u>lots of blank sheets.</u> (...) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed.

Turing (1950)

→ Wishful thinking!!

(my emphasis)

Summing up

- power law everywhere
- the head of the distribution is easy (lots of data, interpolation)
- the tail of the distribution is hard (scarce data, extrapolation)
- humans deal with the problem though a host of mechanisms
  - domain adaptation (new conditions, domains)
  - episodic memory (new words)
  - rule systems (new combinations)
  - unsupervised/weakly supervised learning (small, noisy data)
- mimicking these with machine learning an active & exciting area of research