Statistics and the Scientific Study of Language

What do they have to do with each other?

Mark Johnson

Brown University

ESSLLI 2005

Outline

Why Statistics?

Learning probabilistic context-free grammars

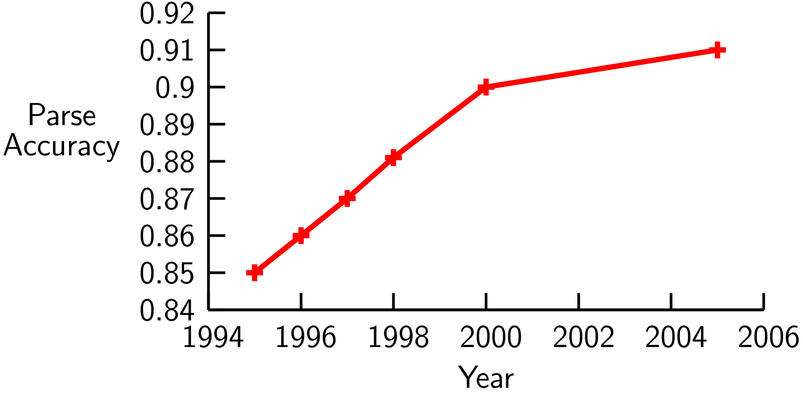
Factoring learning into simpler components

The Janus-faced nature of computational linguistics

Conclusion

Statistical revolution in computational linguistics

- Speech recognition
- Syntactic parsing
- Machine translation



Statistical models in computational linguistics

- *Supervised learning:* structure to be learned is *visible*
  - speech transcripts, treebank, proposition bank, translation pairs
  - more information than available to a child
  - annotation requires (linguistic) knowledge
    - a more practical method of making information available to a computer than writing a grammar by hand
- *Unsupervised learning:* structure to be learned is *hidden*
  - alien radio, alien TV

Chomsky's "Three Questions"

- *What constitutes knowledge of language?*
  - grammar (universal, language specific)
- *How is knowledge of language acquired?*
  - language acquisition
- *How is knowledge of language put to use?*
  - psycholinguistics

(last two questions are about inference)

The centrality of inference

- “poverty of the stimulus”
  - ⇒ innate knowledge of language (universal grammar)
  - ⇒ intricate grammar with rich deductive structure

The centrality of inference

- “poverty of the stimulus”
  - ⇒ innate knowledge of language (universal grammar)
  - ⇒ intricate grammar with rich deductive structure
- Statistics is the theory of *optimal inference* in the presence of *uncertainty*
  - We can define probability distributions over structured objects
  - ⇒ no inherent contradiction between statistical inference and linguistic structure
  - probabilistic models are *declarative*
  - probabilistic models can be systematically *combined*

$$\mathrm{P}(X, Y) = \mathrm{P}(X)\mathrm{P}(Y|X)$$

Questions that statistical models might answer

- What information is required to learn language?
- How useful are different kinds of information to language learners?
  - Bayesian inference can utilize *prior knowledge*
  - Prior can encode “soft” markedness preferences and “hard” universal constraints
- Are there *synergies* between different information sources?
  - Does knowledge of phonology or morphology make word segmentation easier?
  - May provide hints about human language acquisition

Outline

Why Statistics?

Learning probabilistic context-free grammars

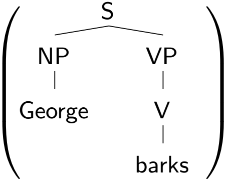
Factoring learning into simpler components

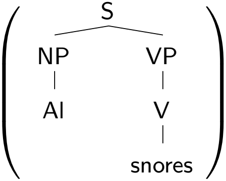
The Janus-faced nature of computational linguistics

Conclusion

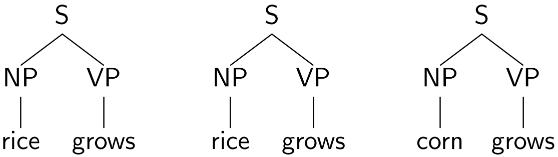
Probabilistic Context-Free Grammars

1.0	S → NP VP	1.0	VP → V
0.75	NP → George	0.25	NP → Al
0.6	V → barks	0.4	V → snores

P  = 0.45

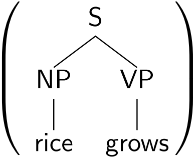
P  = 0.1

Estimating PCFGs from *visible data*

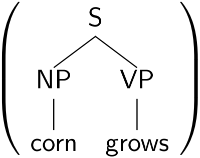


Rule	**Count**	**Rel Freq**
S → NP VP	3	1
NP → rice	2	2/3
NP → corn	1	1/3
VP → grows	3	1

Rel freq is *maximum likelihood estimator* (selects rule probabilities that maximize probability of trees)

P


$= 2/3$

P


$= 1/3$

Estimating PCFGs from *hidden data*

- Training data consists of strings w alone
- Maximum likelihood selects rule probabilities that maximize the *marginal probability* of the strings w
- *Expectation maximization* is a way of building hidden data estimators out of visible data estimators
  - parse trees of iteration i are training data for rule probabilities at iteration $i+1$
- Each iteration is *guaranteed* not to decrease $\mathrm{P}(w)$ (but can get trapped in local minima)
- This can be done without enumerating the parses

Example: The EM algorithm with a toy PCFG

Initial rule probs

rule	**prob**
...	...
VP → V	0.2
VP → V NP	0.2
VP → NP V	0.2
VP → V NP NP	0.2
VP → NP NP V	0.2
...	...
Det → the	0.1
N → the	0.1
V → the	0.1

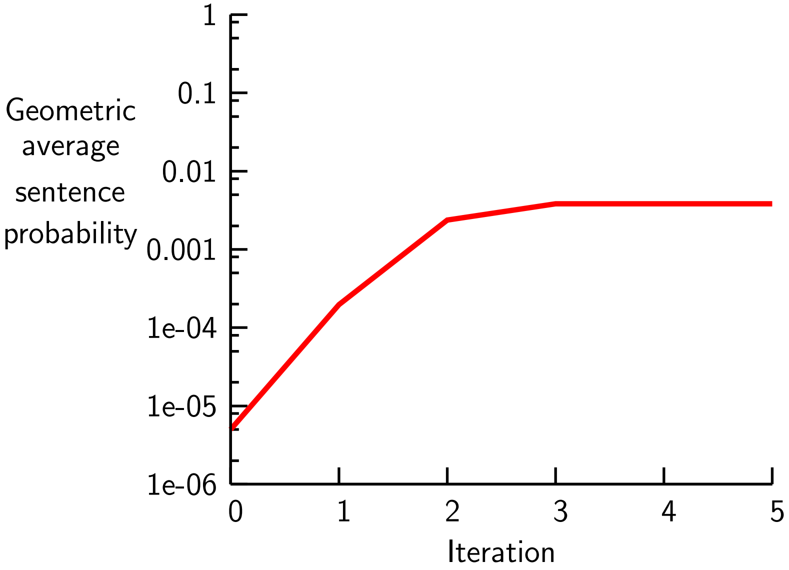
"English" input

the dog bites
the dog bites a man
a man gives the dog a bone
...

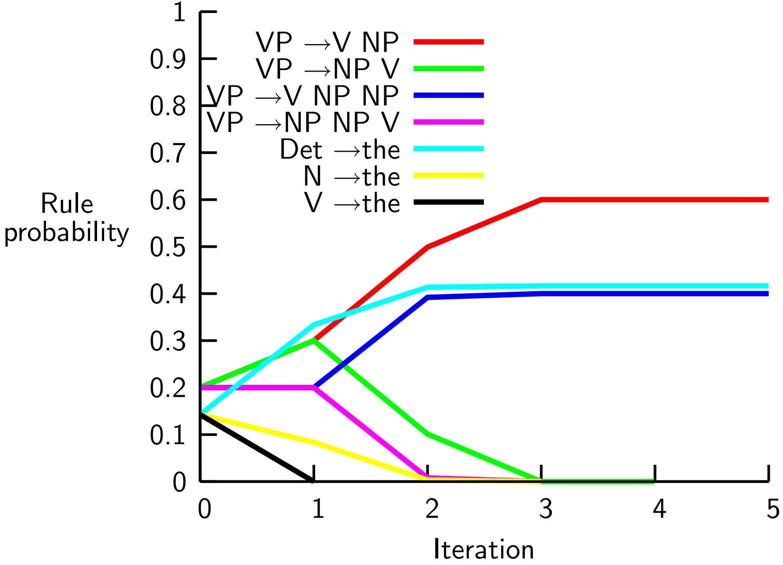
"pseudo-Japanese" input

the dog bites
the dog a man bites
a man the dog a bone gives
...

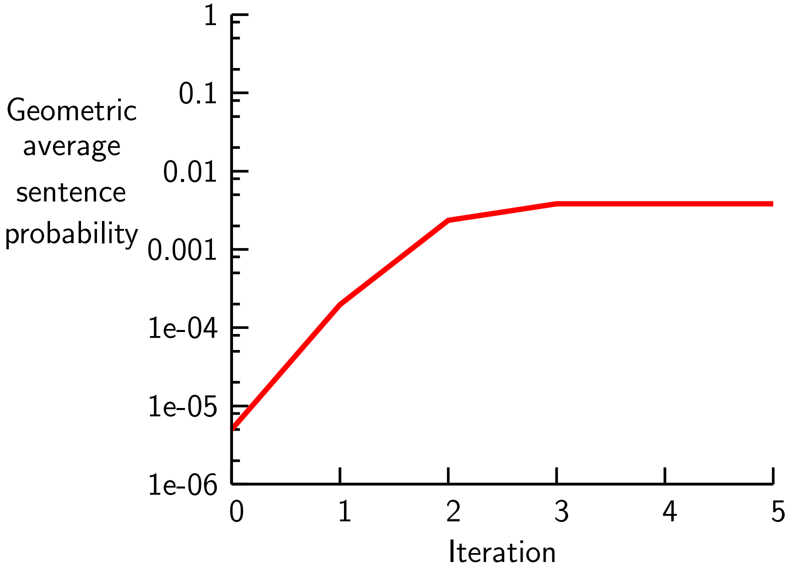
Probability of “English”



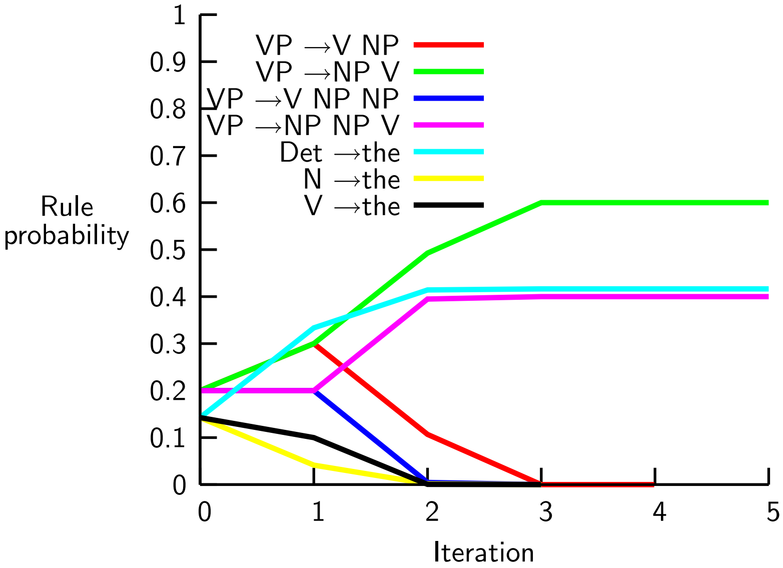
Rule probabilities from “English”



Probability of “Japanese”



Rule probabilities from “Japanese”

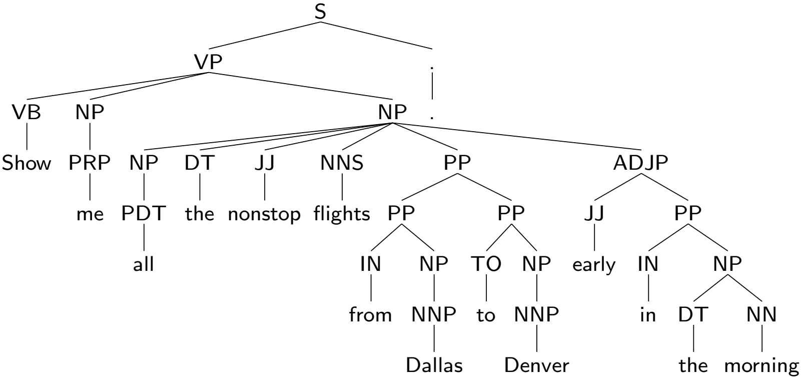


Learning in statistical paradigm

- The likelihood is a differentiable function of rule probabilities
  $\Rightarrow$ learning can involve small, incremental updates
- Learning structure (rules) is hard, but . . .
- Parameter estimation can approximate rule learning
  - start with “superset” grammar
  - estimate rule probabilities
  - discard low probability rules
- Parameters can be associated with other things besides rules (e.g., HeadInitial, HeadFinal)

Applying EM to real data

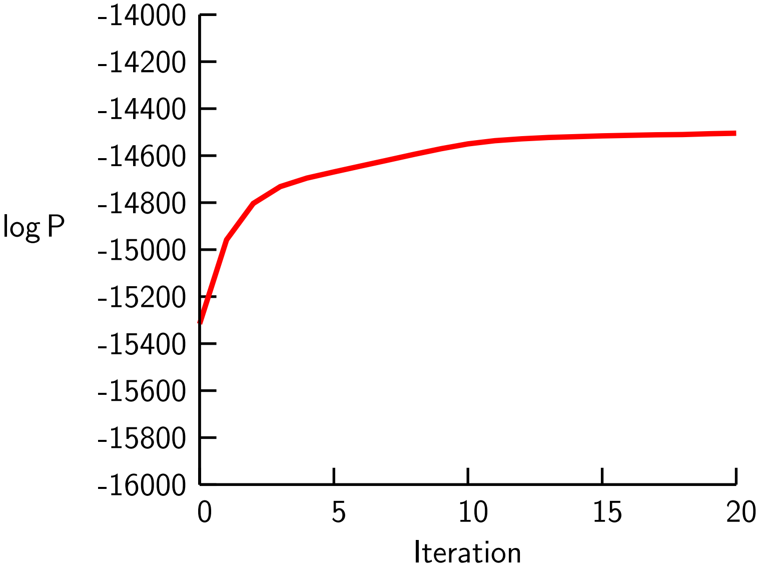
- ATIS treebank consists of 1,300 hand-constructed parse trees
- ignore the words (in this experiment)
- about 1,000 PCFG rules are needed to build these trees



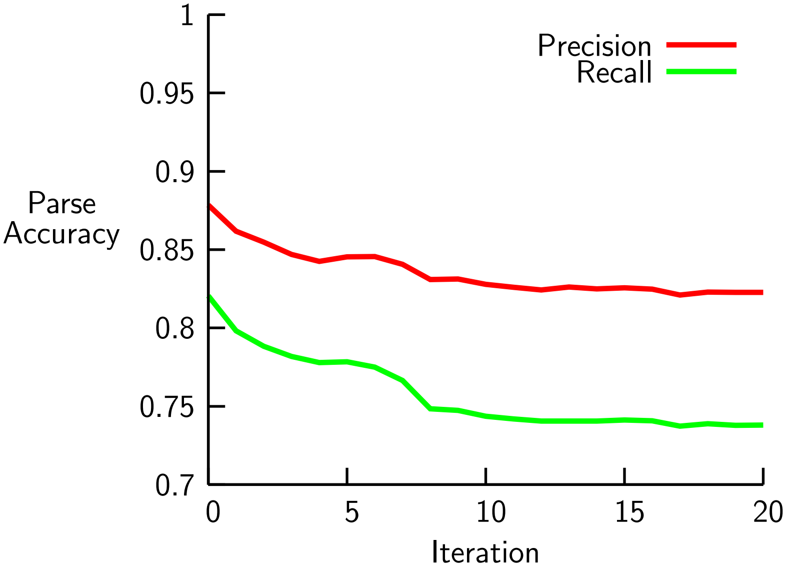
Experiments with EM

1. Extract productions from trees and estimate probabilities probabilities from trees to produce PCFG.
2. Initialize EM with the treebank grammar and MLE probabilities
3. Apply EM (to strings alone) to re-estimate production probabilities.
4. At each iteration:
    - Measure the likelihood of the training data and the quality of the parses produced by each grammar.
    - Test on training data (so poor performance is not due to overlearning).

Log likelihood of training strings



Quality of ML parses



Why does it work so poorly?

- *Wrong data:* grammar is a transduction between form and meaning ⇒ learn from form/meaning pairs
  - exactly what contextual information is available to a language learner?
- *Wrong model:* PCFGs are poor models of syntax
- *Wrong objective function:* Maximum likelihood makes the sentences as likely as possible, but syntax isn't intended to predict sentences (Klein and Manning)
- How can information about the *marginal distribution of strings* $\mathrm{P}(w)$ provide information about the *conditional distribution of parses* t *given strings* $\mathrm{P}(t|w)$?
  - need additional *linking assumptions* about the relationship between parses and strings
- . . . but no one really knows!

Outline

Why Statistics?

Learning probabilistic context-free grammars

Factoring learning into simpler components

The Janus-faced nature of computational linguistics

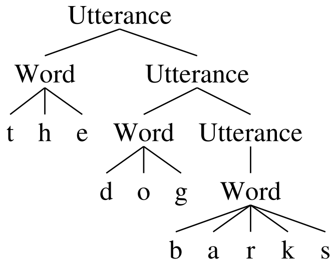
Conclusion

Factoring the language learning problem

- Factor the language learning problem into linguistically simpler components
- Focus on components that might be less dependent on context and semantics (e.g., word segmentation, phonology)
- Identify relevant information sources (including prior knowledge, e.g., UG) by comparing models
- Combine components to produce more ambitious learners
- PCFG-like grammars are a natural way to formulate many of these components

Joint work with Sharon Goldwater and Tom Griffiths

Word Segmentation



Data = t h e d o g b a r k s

Utterance → Word Utterance

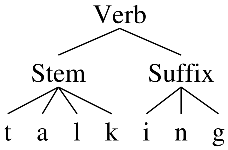
Utterance → Word

Word → w $\qquad w \in \Sigma^\star$

- Algorithms for word segmentation from this information already exists (e.g., Elman, Brent)
- Likely that children perform some word segmentation before they know the meanings of words

Concatenative morphology

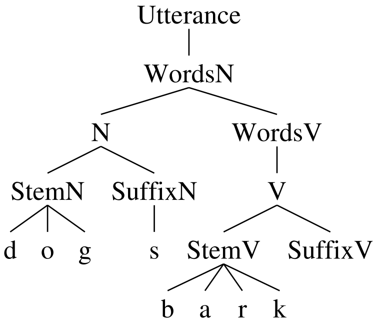
Data = t a l k i n g



$$
\begin{array}{ll}
\text{Verb} \longrightarrow \text{Stem Suffix} & \\
\text{Stem} \longrightarrow w & w \in \Sigma^{\star} \\
\text{Suffix} \longrightarrow w & w \in \Sigma^{\star}
\end{array}
$$

- Morphological alternation provides primary evidence for phonological generalizations (“trucks” /s/ vs. “cars” /z/)
- Morphemes may also provide clues for word segmentation
- Algorithms for doing this already exist (e.g., Goldsmith)

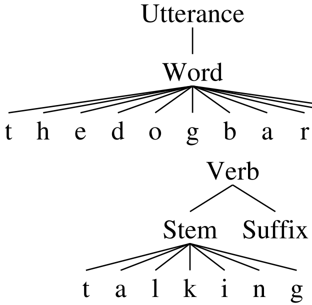
PCFG components can be integrated



$$
\begin{array}{ll}
\text{Utterance} \rightarrow \text{Words}_S & S \in \mathcal{S} \\
\text{Words}_S \rightarrow S \ \text{Words}_T & T \in \mathcal{S} \\
S \rightarrow \text{Stem}_S \ \text{Suffix}_S & \\
\text{Stem}_S \rightarrow t & t \in \Sigma^\star \\
\text{Suffix}_S \rightarrow f & f \in \Sigma^\star
\end{array}
$$

Problems with maximum likelihood estimation

- Maximum likelihood picks model that best fits the data
- *Saturated models* exactly mimic the training data
  $\Rightarrow$ highest likelihood
- Need a different estimation framework



Bayesian estimation

$$\underbrace{\text{P(Hypothesis}|\text{Data)}}_{\text{Posterior}} \propto \underbrace{\text{P(Data}|\text{Hypothesis)}}_{\text{Likelihood}} \underbrace{\text{P(Hypothesis)}}_{\text{Prior}}$$

- Priors can be sensitive to linguistic structure (e.g., a word should contain a vowel)
- Priors can encode linguistic universals and markedness preferences (e.g., complex clusters appear at word onsets)
- Priors can prefer *sparse solutions*
- The choice of the prior is as much a linguistic issue as the design of the grammar!

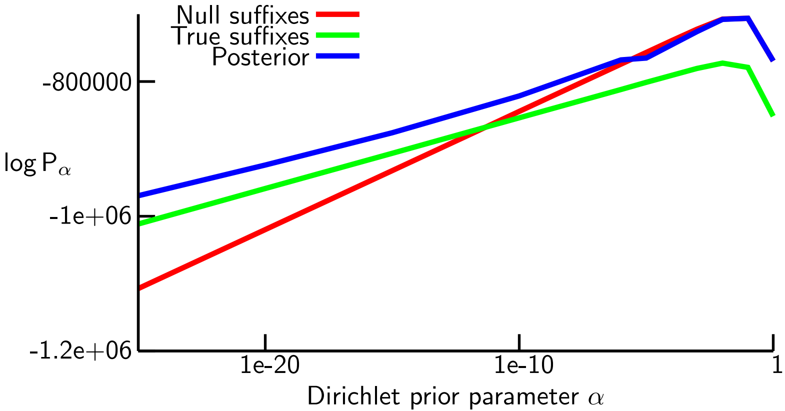
Morphological segmentation experiment

- Trained on orthographic verbs from U Penn. Wall Street Journal treebank
- Dirichlet prior prefers sparse solutions (sparser solutions as $\alpha \to 0$)
- Gibbs Sampler used to sample from posterior distribution of parses
    - reanalyses each word based on a grammar estimated from the parses of the other words

Posterior samples from WSJ verb tokens

$\alpha = 0.1$	$\alpha = 10^{-5}$		$\alpha = 10^{-10}$		$\alpha = 10^{-15}$	
expect	expect		expect		expect	
expects	expects		expects		expects	
expected	expected		expected		expected	
expecting	expect	ing	expect	ing	expect	ing
include	include		include		include	
includes	includes		includ	es	includ	es
included	included		includ	ed	includ	ed
including	including		including		including	
add	add		add		add	
adds	adds		adds		add	s
added	added		add	ed	added	
adding	adding		add	ing	add	ing
continue	continue		continue		continue	
continues	continues		continue	s	continue	s
continued	continued		continu	ed	continu	ed
continuing	continuing		continu	ing	continu	ing
report	report		report		report	

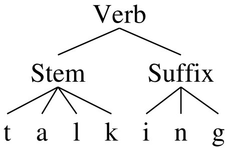
Log posterior of models on token data



- Correct solution is nowhere near as likely as posterior

⇒ model is wrong!

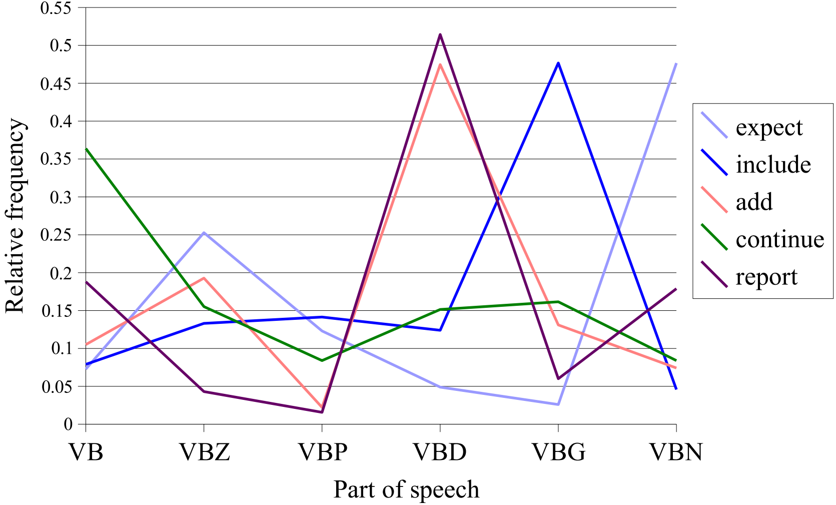
Independence assumption in PCFG model



$$\mathrm{P(Word)} = \mathrm{P(Stem)P(Suffix)}$$

- Model expects relative frequency of each suffix *to be the same for all stems*

Relative frequencies of inflected verb forms



Types and tokens

- A word *type* is a distinct word shape
- A word *token* is an occurrence of a word

$$\begin{aligned}
\text{Data} &= \text{``the cat chased the other cat''} \\
\text{Tokens} &= \text{``the'' 2, ``cat'' 2, ``chased'' 1, ``other'' 1} \\
\text{Types} &= \text{``the'' 1, ``cat'' 1, ``chased'' 1, ``other'' 1}
\end{aligned}$$

- Using word types instead of word tokens effectively normalizes for frequency variations

Posterior samples from WSJ verb *types*

$\alpha = 0.1$		$\alpha = 10^{-5}$		$\alpha = 10^{-10}$		$\alpha = 10^{-15}$	
expect		expect		expect		exp	ect
expects		expect	s	expect	s	exp	ects
expected		expect	ed	expect	ed	exp	ected
expect	ing	expect	ing	expect	ing	exp	ecting
include		includ	e	includ	e	includ	e
include	s	includ	es	includ	es	includ	es
included		includ	ed	includ	ed	includ	ed
including		includ	ing	includ	ing	includ	ing
add		add		add		add	
adds		add	s	add	s	add	s
add	ed	add	ed	add	ed	add	ed
adding		add	ing	add	ing	add	ing
continue		continu	e	continu	e	continu	e
continue	s	continu	es	continu	es	continu	es
continu	ed	continu	ed	continu	ed	continu	ed
continuing		continu	ing	continu	ing	continu	ing
report		report		repo	rt	rep	ort

Summary so far

- Unsupervised learning is difficult on real data!
- There's a lot to learn from simple problems
  - need models that require all stems in same class to have same suffixes but permit suffix frequencies to vary with the stem
- Related problems arise in other linguistic domains as well
  - Many verbs share the same subcategorization frames, but subcategorization frame frequencies depend on head verb.
- Hopefully we can combine these simple learners to study their interaction in more complex domains

Outline

Why Statistics?

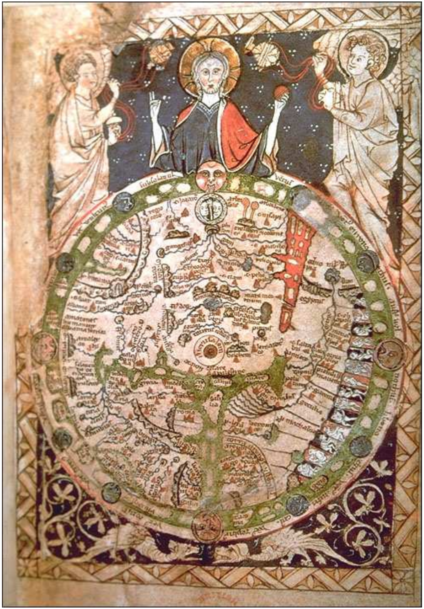
Learning probabilistic context-free grammars

Factoring learning into simpler components

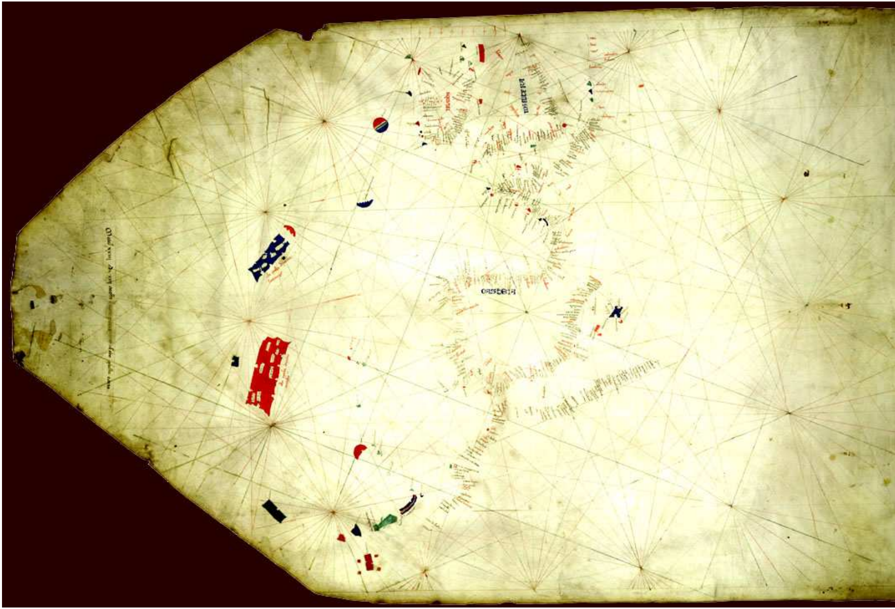
The Janus-faced nature of computational linguistics

Conclusion

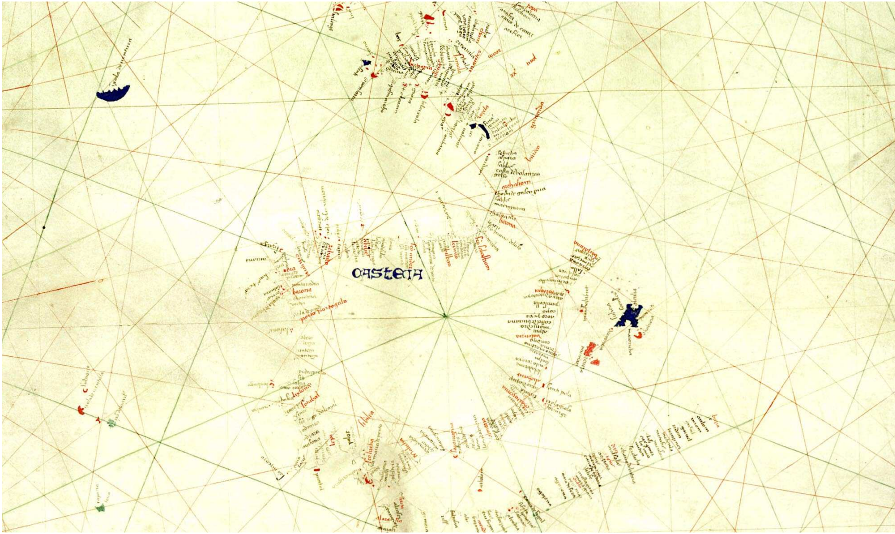
Psalter Mappa Mundi (1225?)



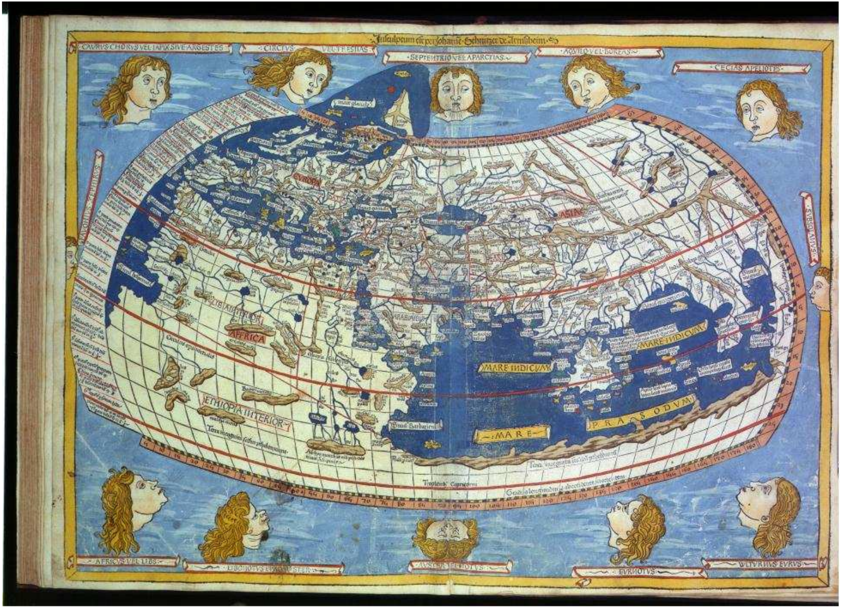
Portolan chart circa 1424



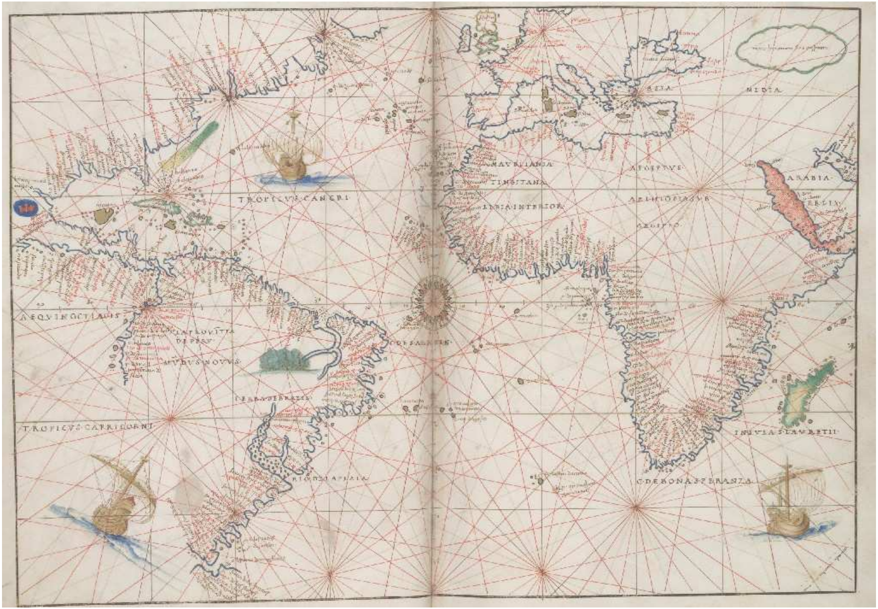
Portolan chart circa 1424 (center)



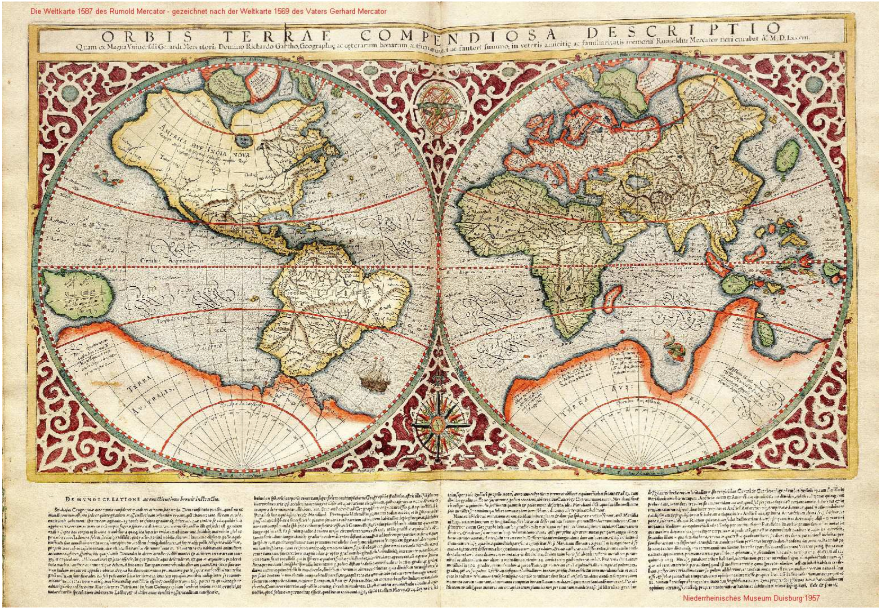
Waldseemüller 1507, after Ptolemy



Battista Agnese portolan chart circa 1550



Mercator 1569


Die Weltkarte 1587 des Rumold Mercator - gezeichnet nach der Weltkarte 1569 des Vaters Gerhard Mercator

ORBIS TERRAE COMPENDIOSA DESCRIPTIO
Quam ex Magna Vniuersali Gerardi Mercatoris Domino Richardo Gartho, Geographiæ ac cæterarum bonarum artium amatori ac fautori summo, in veteris amicitiæ ac familiaritatis memoriã Rumoldus Mercator fieri curabat Aº. M.D.LXXXVII.
AMERICA SIVE INDIA NOVA
Circulus Aequinoctialis
Tropicus Capricorni
TERRA AVSTRALIS
DE MVNDI CREATIONE ac constitutione breuis instructio.

Niederrheinisches Museum Duisburg 1957

… back to computational linguistics

- Be wary of analogies from the history of science!
  - we only remember the successes
- May wind up learning something very different to what you hoped
- Cartography and geography benefited from both the academic and Portolan traditions
- Geography turned out to be about brute empirical facts
  - but geology and plate tectonics
- Mathematics (geometry and trigonometry) turned out to be essential
- Even wrong ideas can be very important
  - the cosmographic tradition survives in celestial navigation

Outline

Why Statistics?

Learning probabilistic context-free grammars

Factoring learning into simpler components

The Janus-faced nature of computational linguistics

Conclusion

Conclusion

- Statistical methods have both engineering and scientific applications
- Inference plays a central role in linguistic theory
- Uncertain information $\Rightarrow$ statistical inference
- The statistical component of a model may require as much linguistic insight as the structural component
- Factoring the learning problem into linguistically simpler pieces may be a good way to proceed
- Who knows what the future will bring?

Thank you

“I ask you to look both ways. For the road to a knowledge of the stars leads through the atom; and important knowledge of the atom has been reached through the stars.”

— Sir Arthur Eddington

“Everything should be made as simple as possible, but not one bit simpler.”

— Albert Einstein

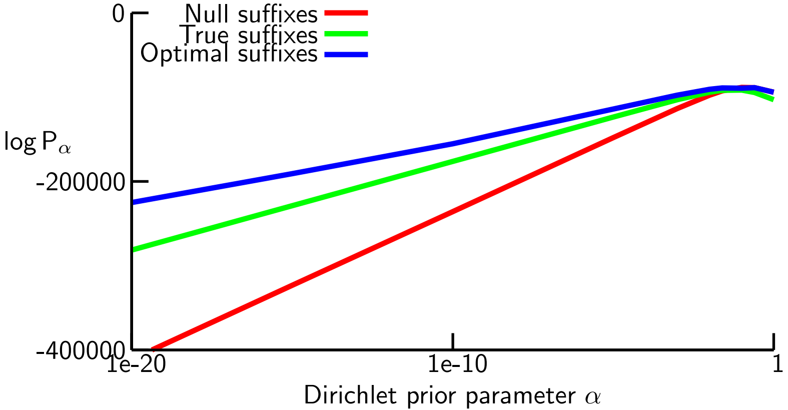
“Something unknown is doing we don’t know what.”

— Sir Arthur Eddington

“You can observe a lot just by watching.”

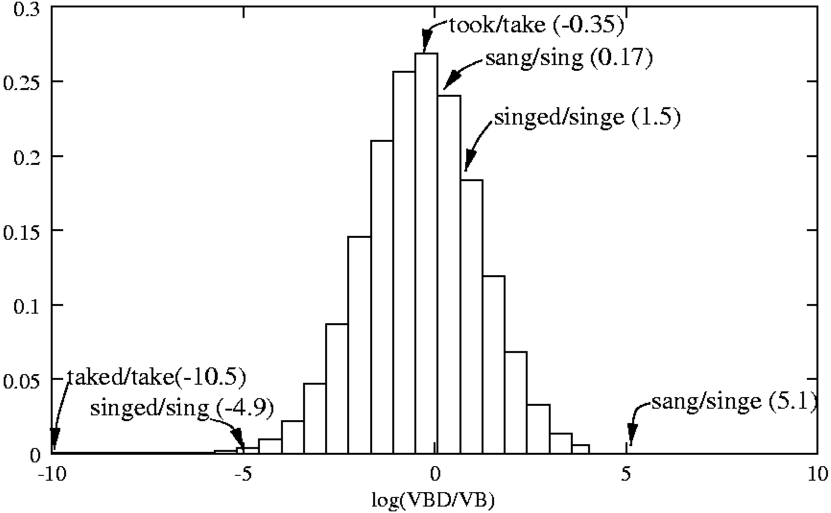
— Yogi Berra

Log posterior of models on type data



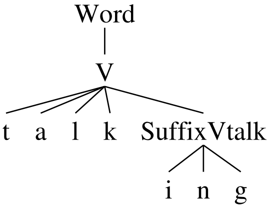
- Correct solution is close to optimal for $\alpha = 10^{-3}$

Morpheme frequencies provide useful information



Yarowsky and Wicentowski (2000) “Minimally supervised Morphological Analysis by Multimodal Alignment”

Suffix frequency depends on stem



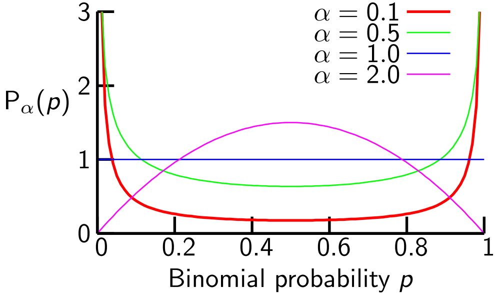
$$\begin{array}{ll} \text{Word} \rightarrow S & S \in \mathcal{S} \\ S \rightarrow t\ \text{Suffix}_{S,t} & t \in \Sigma^\star \\ \text{Suffix}_{S,t} \rightarrow f & f \in \Sigma^\star \end{array}$$

- Suffix distributions $\text{Suffix}_{S,t} \rightarrow f$ depend on the stem t
- Prior constrains suffix distributions $\text{Suffix}_{S,t} \rightarrow f$ for stems t in the same class to be similar
- Model is *saturated* and *not context-free*

Dirichlet priors and sparse solutions

- The expansions of a nonterminal in a PCFG are distributed according to a multinomial
- Dirichlet priors are a standard prior over multinomial distributions

$$\mathrm{P}(p_1, \ldots, p_n) \quad \propto \quad \prod_{i=1}^{n} p_i^{\alpha - 1} \qquad \alpha > 0$$



Estimation procedures

- Dirichlet prior prefers sparse solutions ⇒ MAP grammar may be undefined even though MAP parses are defined
- Markov Chain Monte Carlo techniques can sample from the posterior distribution over grammars and parses
- *Gibbs sampling:*
  1. Construct a corpus of (word,tree) pairs by randomly assigning trees to each word in the data
  2. Repeat:
     - 2.1 Pick a word w and its tree from the corpus at random
     - 2.2 Estimate a grammar from the trees assigned to the other words in the corpus
     - 2.3 Parse w with this grammar, producing a distribution over trees
     - 2.4 Select a tree t from this distribution, and add (w, t) to the corpus

Outline

Why Statistics?

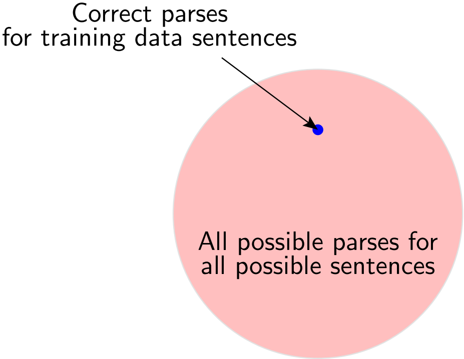
Learning probabilistic context-free grammars

Factoring learning into simpler components

The Janus-faced nature of computational linguistics

Conclusion

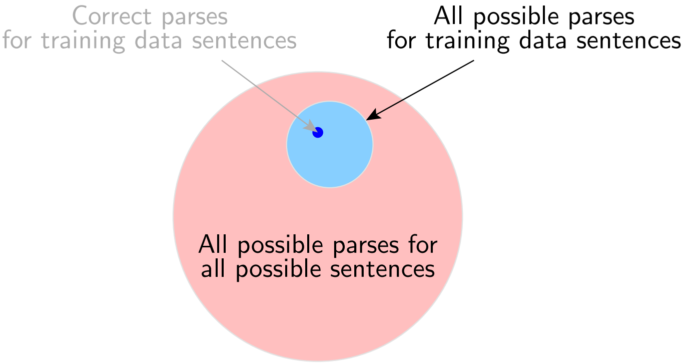
Maximum likelihood estimation from visible data



- Standard maximum likelihood estimation makes the treebank trees t and strings w as likely as possible relative to all other possible trees and strings

$$\widehat{g} \; = \; \arg\max_{g} \; = \; \mathsf{P}_g(w, t) \; = \; \arg\max_{g} \mathsf{P}_g(t|w) \, \mathsf{P}_g(w)$$

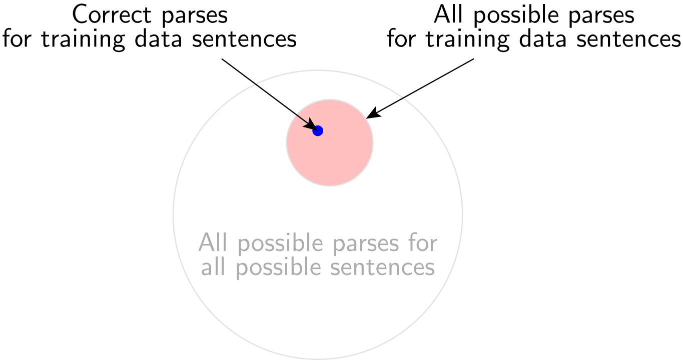
Maximum likelihood estimation from hidden data



- Maximum likelihood estimation maximizes the probability of the words w of the training data, relative to all other possible word strings

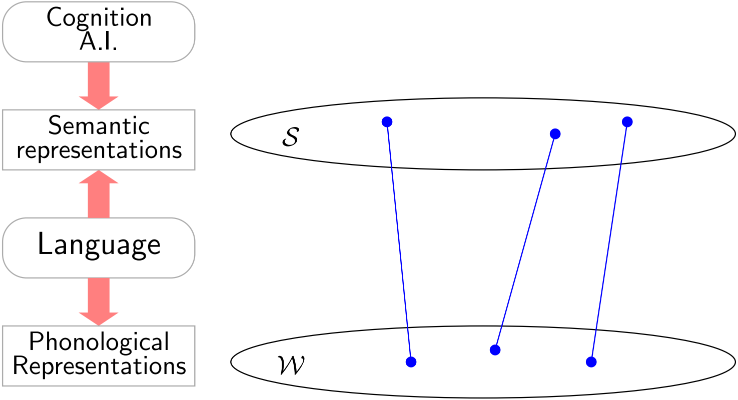
$$\widehat{g} \;=\; \arg\max_{g} \mathrm{P}_{g}(w) \;=\; \arg\max_{g} \sum_{t} \mathrm{P}_{g}(t, w)$$

Conditional MLE from visible data



- Conditional MLE maximizes the *conditional probability* $\mathsf{P}_g(t|w)$ of the training trees t relative to the training words w
- *learns nothing* from the distribution $\mathsf{P}_g(w)$ of words

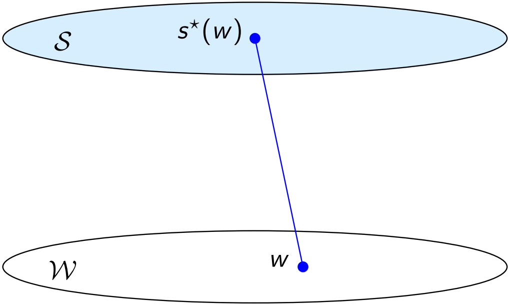
Language as a transduction from form to meaning



- Grammar generates a phonological form w from a semantic representation s

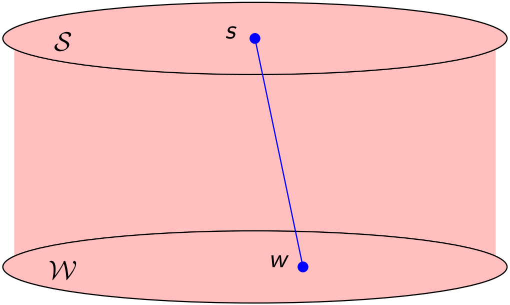
$$\mathrm{P}(w, s) \;=\; \underbrace{\mathrm{P}_g(w|s)}_{\text{“language”}} \; \underbrace{\mathrm{P}_c(s)}_{\text{“cognition”}}$$

Interpretation is finding the most likely meaning $s^\star$



$$s^\star(w) \;=\; \arg\max_{s\in\mathcal{S}} \mathrm{P}(s|w) \;=\; \arg\max_{s\in\mathcal{S}} \mathrm{P}_g(w|s)\mathrm{P}_c(s)$$

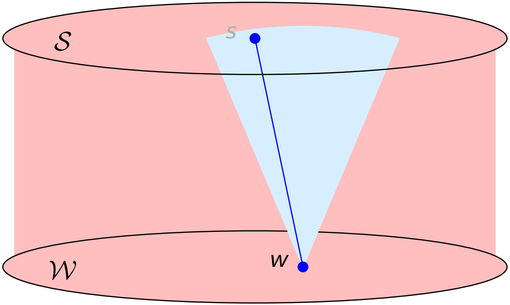
Maximum likelihood estimate $\widehat{g}$ from visible data



- Training data consists of phonology/semantic pairs (w, s)
- Maximum likelihood estimate of grammar $\widehat{g}$ makes (w, s) as likely as possible relative to all other possible pairs $(w', s'), w' \in \mathcal{W}, s' \in \mathcal{S}$

$$\widehat{g} \;=\; \arg\max_{g} \mathrm{P}(w, s) \;=\; \arg\max_{g} \mathrm{P}(w|s)$$

MLE $\widehat{g}$ from hidden data



- Training data consists of phonological strings w alone
- MLE makes w as likely as possible relative to other strings

$$\widehat{g} \;=\; \arg\max_{g} \mathrm{P}(w) \;=\; \arg\max_{g} \sum_{s \in \mathcal{S}} \mathrm{P}_{g}(w|s)\mathrm{P}_{c}(s)$$

⇒ *It may be possible to learn g from strings alone*

- The cognitive model P_c can in principle be learnt the same way