Language Acquisition as Statistical Inference

Mark Johnson

Joint work with many people, including
Ben Börschinger, Eugene Charniak, Katherine Demuth,
Michael Frank, Sharon Goldwater, Tom Griffiths,
Bevan Jones and Ed Stabler;
thanks to Bob Berwick, Stephen Crain and Mark Steedman
for comments and suggestions

Macquarie University
Sydney, Australia

Paper and slides available from http://science.MQ.edu.au/˜mjohnson

July 2013

MACQUARIE UNIVERSITY

Main claims

- Setting grammatical parameters can be viewed as a *parametric statistical inference* problem
  - e.g., learn *whether* language has verb raising
  - if parameters are *local in the derivation tree* (e.g., lexical entries, including empty functional categories) then there is an efficient parametric statistical for identifying them
  - only requires primary linguistic data contains *positive example sentences*
- Learning a lexicon can be viewed as a *nonparametric statistical inference* problem
  - number of possible words $\Rightarrow$ number of degrees of freedom is unbounded: *learning a lexicon is a non-trivial problem!*
  - probabilistic models can integrate *multiple sources of information*, including information from the non-linguistic context
- In statistical inference usually *parameters have continuous values*, but *is this linguistically reasonable?*

Outline

Statistics and probabilistic models

Parameter-setting as parametric statistical inference

An example of syntactic parameter learning

Learning the lexicon as non-parametric inference

Synergies with syllabification

Grounded learning and learning word-topic associations

The role of social cues in word learning

Conclusions, and where do we go from here?

Statistical inference and probabilistic models

- A *statistic* is *any function of the data*
  - usually chosen to *summarise* the data
- Statistical inference usually exploits not just the occurrence of phenomena, but also their *frequency*
- *Probabilistic models* predict the frequency of phenomena
  $\Rightarrow$ very useful for statistical inference
  - inference usually involves *setting parameters* to *minimise difference* between model's expected value of a statistic and its value in data
  - statisticans have shown certain procedures are *optimal* for wide classes of inference problems
- Probabilistic extensions for virtually all theories of grammar
  - $\Rightarrow$ *no inherent conflict between grammar and statistical inference*
  - $\Rightarrow$ technically, statistical inference can be used under virtually any theory of grammar
    - *but is anything gained by doing so?*

MACQUARIE UNIVERSITY

Do “linguistic frequencies” make sense?

- Frequencies of many surface linguistic phenomena *vary dramatically with non-linguistic context*
  - arguably, word frequencies aren’t part of “knowledge of English”
- Perhaps humans only use *robust statistics*
  - e.g., closed-class words are often *orders of magnitude* more frequent than open-class words
  - e.g., the *conditional distribution of surface forms given meanings* $\mathrm{P}(\mathrm{SurfaceForm} \mid \mathrm{Meaning})$ may be robust, perhaps almost categorical (Uniqueness principle)

MACQUARIE UNIVERSITY

Why exploit frequencies when learning?

- Human learning shows frequency effects
  - usually higher frequency $\Rightarrow$ faster learning
  - $\not\Rightarrow$ statistical learning (e.g., trigger models show frequency effects)
- Frequency statistics provide *potentially valuable information*
  - parameter settings may need updating if *expected frequency is significantly higher than empirical frequency*
  - $\Rightarrow$ avoid "no negative evidence" problems
- Statistical inference seems to work better for many aspects of language than other methods
  - scales up to larger, more realistic data
  - produces more accurate results
  - more robust to noise in the input

MACQUARIE UNIVERSITY

Some theoretical results about statistical grammar inference

- *statistical learning can succeed when categorical learning fails* (e.g., PCFGs can be learnt from positive examples alone, but CFGs can't) (Horning 1969, Gold 1967)
  - statistical learning *assumes more about the input* (independent and identically-distributed)
  - and has *a weaker notion of success* (convergence in distribution)
- *learning PCFG parameters from positive examples alone is computationally intractable* (Cohen et al 2012)
  - this is a "worst-case" result, typical problems (or "real" problems) may be easy
  - *result probably generalises to Minimalist Grammars* (MGs) as well

  ⇒ MG inference algorithm sketched here will run slowly, or will converge to wrong parameter estimates, for some MGs on some data

Parametric and non-parametric inference

- A *parametric model* is one with a finite number of prespecified parameters
  - Principle-and-parameters grammars are parametric models
- *Parametric inference* is inference for the parameter values of a parametric model
- A *non-parametric model* is one which can't be defined using a bounded number of parameters
  - a lexicon is a non-parametric model if there's no universal bound on possible lexical entries (e.g., phonological forms)
- *Non-parametric inference* is inference for (some properties of) nonparametric models

MACQUARIE UNIVERSITY

Outline

Statistics and probabilistic models

Parameter-setting as parametric statistical inference

An example of syntactic parameter learning

Learning the lexicon as non-parametric inference

Synergies with syllabification

Grounded learning and learning word-topic associations

The role of social cues in word learning

Conclusions, and where do we go from here?

MACQUARIE UNIVERSITY

Statistical inference for MG parameters

- Claim: there is a *statistical algorithm for inferring parameter values of Minimalist Grammars* (MGs) from positive example sentences alone, assuming:
  - MGs are efficiently parsable
  - MG *derivations* (not parses!) have a *context-free structure*
  - parameters are associated with *subtree-local configurations* in derivations (e.g., lexical entries)
  - a probabilistic version of MG with *real-valued parameters*
- Example: learning verb-raising parameters from toy data
  - e.g., learn language has V>T movement from examples like *Sam sees often Sasha*
  - truth in advertising: this example uses an equivalent CFG instead of an MG to generate derivations
- *Not tabula rasa learning*: we estimate parameter values (e.g., that a language has V>T movement); the possible parameters and their linguistic implications are prespecified (e.g., innate)

Outline of the algorithm

- Use a "MaxEnt" probabilistic version of MGs
- Although MG *derived structures* are not context-free (because of movement) they have *context-free derivation trees* (Stabler and Keenan 2003)
- Parametric variation is *subtree-local* in derivation tree
  - e.g., availability of specific *empty functional categories* triggers different movements

⇒ The *partition function* and its derivatives can be efficiently calculated (Hunter and Dyer 2013)

⇒ Standard "hill-climbing" methods for context-free grammar parameter estimation generalise to MGs

Maximum likelihood statistical inference procedures

- If we have:
  - a probabilistic model P that depends on parameter values w, and
  - data D we want to use to infer w

  the *Principle of Maximum Likelihood* is: *select the w that makes the probability of the data $\mathrm{P}(D)$ as large as possible*
- Maximum likelihood inference is *asymptotically optimal* in several ways
- Maximising likelihood is an *optimisation problem*
- *Calculating $\mathrm{P}(D)$* (or something related to it) is necessary
  - need the *derivative of the partition function* for hill-climbing search

MACQUARIE UNIVERSITY

Maximum Likelihood and the Subset Principle

- The Maximum Likelihood Principle entails a probabilistic version of the Subset Principle (Berwick 1985)
- Maximum Likelihood Principle: select parameter weights w to make the probability of data $\mathrm{P}(D)$ as large as possible
- $\mathrm{P}(D)$ is the *product* of the probabilities of the sentences in D
  - $\Rightarrow$ w assigns each sentence in D relatively large probability
  - $\Rightarrow$ w generates at least the sentences in D
- Probabilities of all sentences must *sum to 1*
  - $\Rightarrow$ can assign higher probability to sentences in D if w generates fewer sentences outside of D
    - e.g., if w generates 100 sentences, then each can have prob. 0.01 if w generates 1,000 sentences, then each can have prob. 0.001

$\Rightarrow$ Maximum likelihood estimation selects w so sentences in D have high prob., and few sentences not in D have high prob.

MACQUARIE UNIVERSITY

The utility of continuous-valued parameters

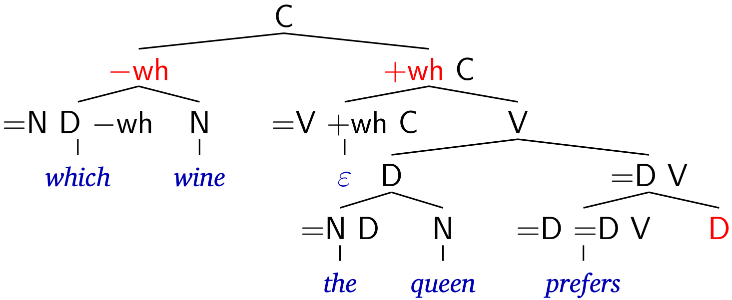
- Standardly, linguistic parameters are *discrete* (e.g., Boolean)
- Most statistical inference procedures use *continuous* parameters
- In the models presented here, parameters and lexical entries are associated with *real-valued weights*
  - E.g., if $w_{V>T} \ll 0$ then a derivation containing V-to-T movement will be much less likely than one that does not
  - E.g., if $w_{will:V} \ll 0$ then a derivation containing the word *will* with syntactic category V will be much less likely
- Continuous parameter values and probability models:
  - can represent *partial or incomplete knowledge* with intermediate values (e.g., when learner isn't sure)
  - define a *gradient* that enables *incremental "hill climbing" search*
  - but also might allow *"zombie" parameter settings* that don't correspond to possible human languages

MACQUARIE UNIVERSITY

Derivations in Minimalist Grammars

- Grammar has two fundamental operations: *external merge* (head-complement combination) and *internal merge* (movement)
- Both operations are driven by *feature checking*
  - derivation terminates when all formal features have been *checked* or cancelled
- MG as formalised by Stabler and Keenan (2003):
  - the *string and derived tree languages* MGs generate are *not context-free*, but
  - MG derivations are specified by a *derivation tree*, which abstracts over surface order to reflect the structure of internal and external merges, and
  - the *possible derivation trees* have a *context-free structure* (c.f. TAG)

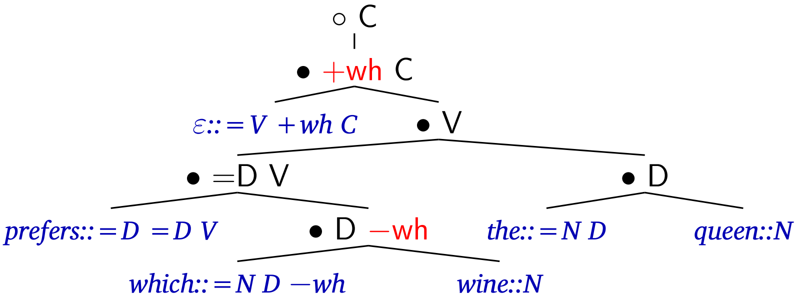
Example MG derived tree



which wine the queen prefers

MACQUARIE UNIVERSITY

Example MG derivation tree



which wine the queen prefers

Calculating the probability $\mathrm{P}(D)$ of data D

- If data D is a sequence of independently generated sentences $D = (s_1, \ldots, s_n)$, then:

$$\mathrm{P}(D) \;=\; \mathrm{P}(s_1) \times \ldots \times \mathrm{P}(s_n)$$

- If a sentence s is ambiguous with derivations $\tau_1, \ldots, \tau_m$ then:

$$\mathrm{P}(s) \;=\; \mathrm{P}(\tau_1) + \ldots + \mathrm{P}(\tau_m)$$

- These are standard formal language theory assumptions
  - which does not mean they are correct!
  - Luong et al (2013) shows learning can improve by modeling dependencies between s_i and s_{i+1}
- Key issue: *how do we define the probability $\mathrm{P}(\tau)$ of derivation τ?*
- If s is very ambiguous (as is typical during learning), need to *calculate $\mathrm{P}(s)$ without enumerating all its derivations*

Parsing Minimalist Grammars

- For Maximum Likelihood inference we need to calculate the MG derivations of the sentences in the training data D
- Stabler (2012) describes several algorithms for parsing with MGs
  - MGs can be translated to equivalent Multiple CFGs (MCFGs)
  - while MCFGs are strictly more expressive than CFGs, for any given sentence there is a CFG that generates an equivalent set of parses (Ljunglöf 2012)
  - ⇒ CFG methods for efficient parsing (Lari and Young 1990) should generalise to MGs

MaxEnt probability distributions on MG derivations

- Associate each parameter π with a function from derivations τ to the number of times some configuration appears in τ
  - e.g., $+\mathrm{wh}(\tau)$ is the number of WH-movements in τ
  - same as *constraints* in Optimality Theory
- Each parameter π has a *real-valued weight* w_π
- The probability $\mathrm{P}(\tau)$ of derivation τ is:

$$\mathrm{P}(\tau) \;=\; \frac{1}{Z}\exp\left(\sum_{\pi} w_\pi\, \pi(\tau)\right)$$

  where $\pi(\tau)$ is the number of times the configuration π occurs in τ
- w_π generalises a conventional binary parameter value:
  - if $w_\pi > 0$ then each occurence of π *increases* $\mathrm{P}(\tau)$
  - if $w_\pi < 0$ then each occurence of π *decreases* $\mathrm{P}(\tau)$
- Essentially the same as Abney (1996) and Harmonic Grammar (Smolensky et al 1993)

The importance of the partition function Z

- Probability $\mathrm{P}(\tau)$ of a derivation τ:

$$\mathrm{P}(\tau) \;=\; \frac{1}{Z} \exp\left(\sum_{\pi} w_{\pi}\, \pi(\tau)\right)$$

- The *partition function* Z is crucial for statistical inference
  - inference algorithms for learning w_π without Z are more heuristic
- Calculating Z naively involves *summing over all possible derivations of all possible strings*, but this is usually *infeasable*
- But if the possible derivations τ have a context-free structure and the π configurations are "local", it is possible to calculate Z without exhaustive enumeration

MACQUARIE UNIVERSITY

Calculating the partition function Z for MGs

- Hunter and Dyer (2013) observe that the partition function Z for MGs can be *efficiently calculated* generalising the techniques of Nederhof and Satta (2008) if:
  - the parameters π are *functions of local subtrees of the derivation tree* τ, and
  - the possible MG derivations have a *context-free structure*
- Stabler (2012) suggests that *empty functional categories control parametric variation* in MGs
  - e.g., if lexicon contains "ε::=V +wh C" then language has WH-movement
  - the number of occurences of each empty functional category is a function of local subtrees

$\Rightarrow$ If we define a parameter π_λ for each lexical entry λ where:

  - $\pi_\lambda(\tau)$ = number of times λ occurs in derivation τ
  - then the partition function Z can be efficiently calculated.

MACQUARIE UNIVERSITY

Outline

Statistics and probabilistic models

Parameter-setting as parametric statistical inference

An example of syntactic parameter learning

Learning the lexicon as non-parametric inference

Synergies with syllabification

Grounded learning and learning word-topic associations

The role of social cues in word learning

Conclusions, and where do we go from here?

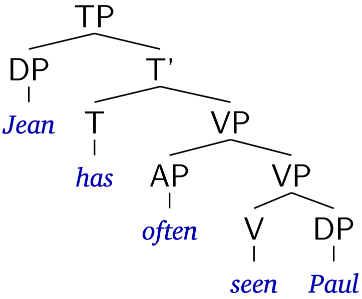
MACQUARIE UNIVERSITY

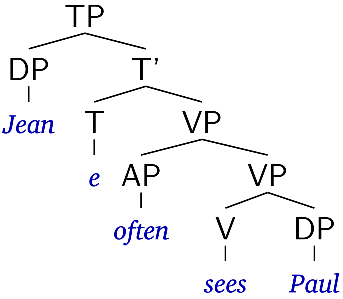
A "toy" example

- Involves verb movement and inversion (Pollock 1989)
- 3 different sets of about 25 input sentences
  - ("English") *Sam often sees Sasha, Q will Sam see Sasha, . . .*
  - ("French") *Sam sees often Sasha, Sam will often see Sasha, . . .*
  - ("German") *Sees Sam often Sasha, Will Sam Sasha see, . . .*
- *Syntactic parameters*: V>T, T>C, T>Q, XP>SpecCP, $\mathrm{V}_{\mathrm{init}}$, $\mathrm{V}_{\mathrm{fin}}$
- *Lexical parameters* associating all words with all categories (e.g., *will*:I, *will*:V, *will*:D)
- Hand-written CFG instead of MG; parameters associated with CF rules rather than empty categories
  - grammar inspired by MG analyses
  - *calculates same parameter functions* π as MG would
  - could use a MG parser if one were available

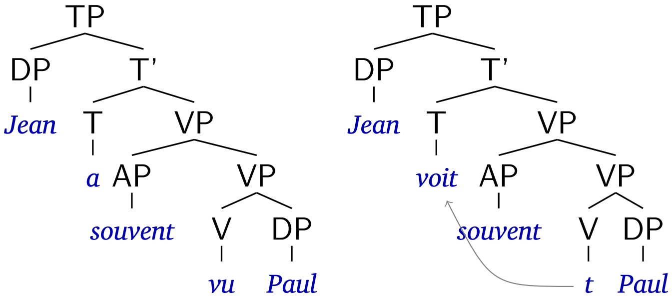
MACQUARIE UNIVERSITY

“English”: no V-to-T movement

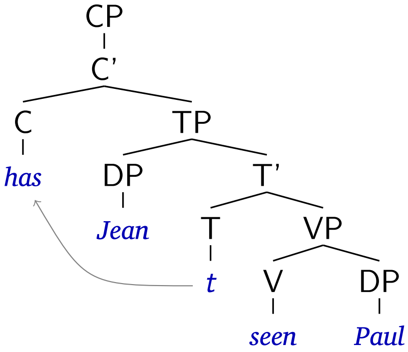




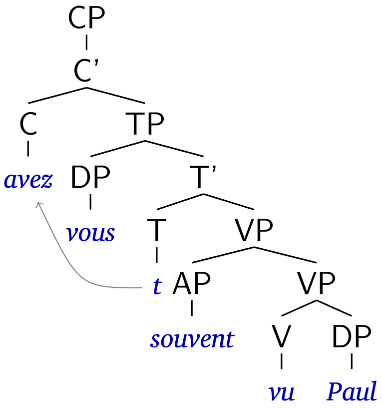
“French”: V-to-T movement

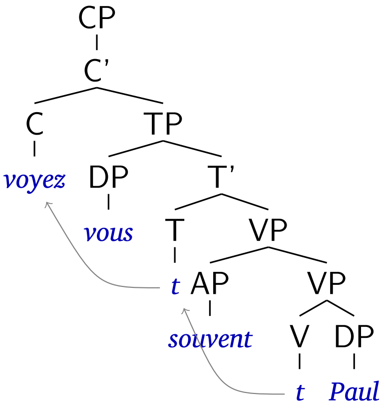


“English”: T-to-C movement in questions

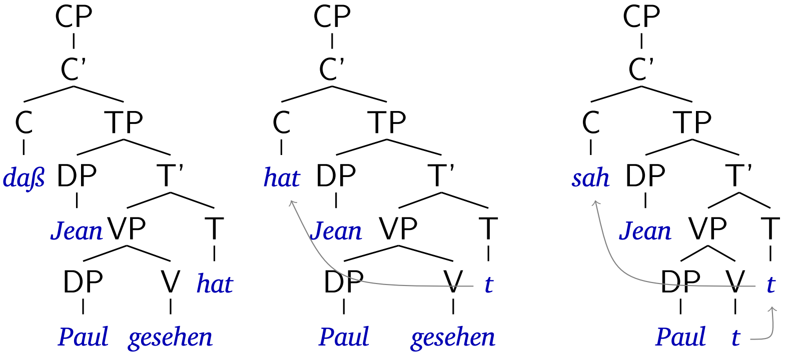


“French”: T-to-C movement in questions

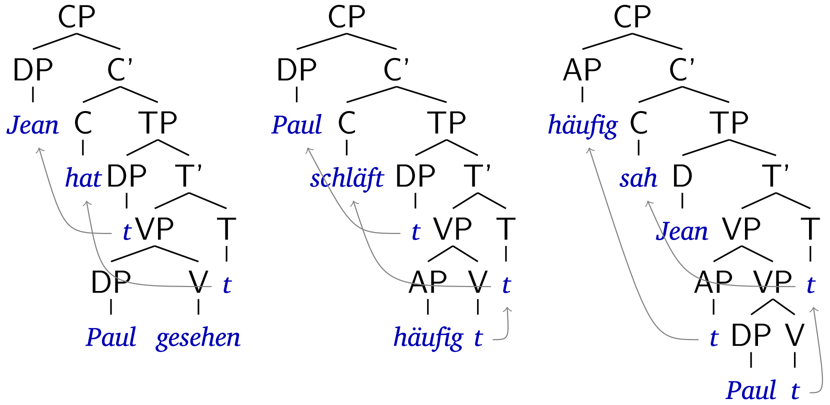




“German”: V-to-T and T-to-C movement



"German": V-to-T, T-to-C and XP-to-SpecCP movement



Input to parameter inference procedure

- A CFG designed to mimic MG derivations, with parameters associated with rules
- About 25 sentences, such as:
  - ("English") *Sam often sees Sasha, Q will Sam see Sasha*
  - ("French") *Sam sees often Sasha, Q see Sam Sasha*
  - ("German") *Sam sees Sasha, sees Sam Sasha, will Sam Sasha see*
- Identifying parameter values is easy if we know lexical categories
- Identifying lexical entries is easy if we know parameter values
- Learning both jointly faces a "chicken-and-egg" problem

Algorithm for statistical parameter estimation

- Parameter estimation algorithm:

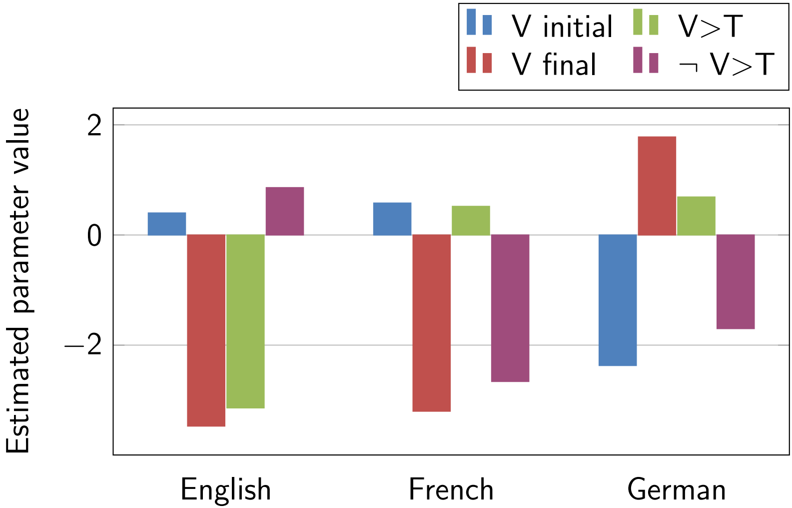
  Initialise parameter weights somehow
  Repeat until converged:
      calculate likelihood and its derivatives
      update parameter weights to increase likelihood

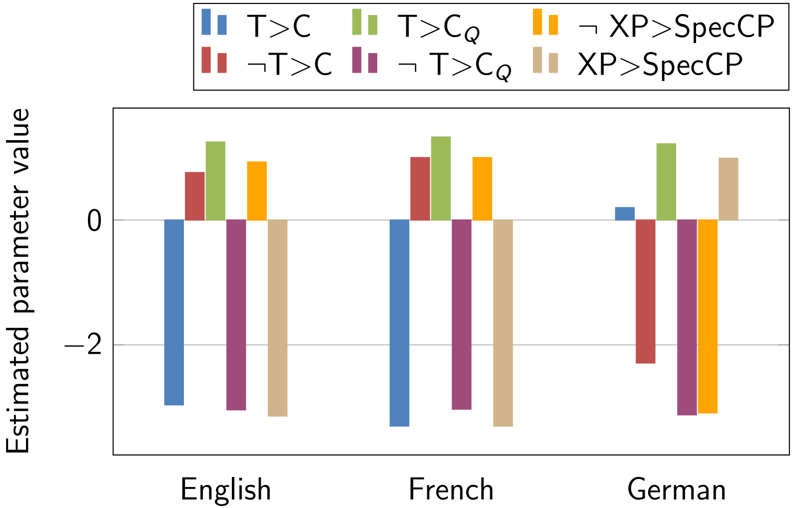
- Very simple parameter weights updates suffice
- Computationally most complex part of procedure is *parsing the data* to calculate likelihood and its derivatives

  $\Rightarrow$ learning is a by-product of parsing

- Straight-forward to develop *incremental on-line* versions of this algorithm (e.g., stochastic gradient ascent)
  - an advantage of explicit probabilistic models is that there are standard techniques for developing algorithms with various properties

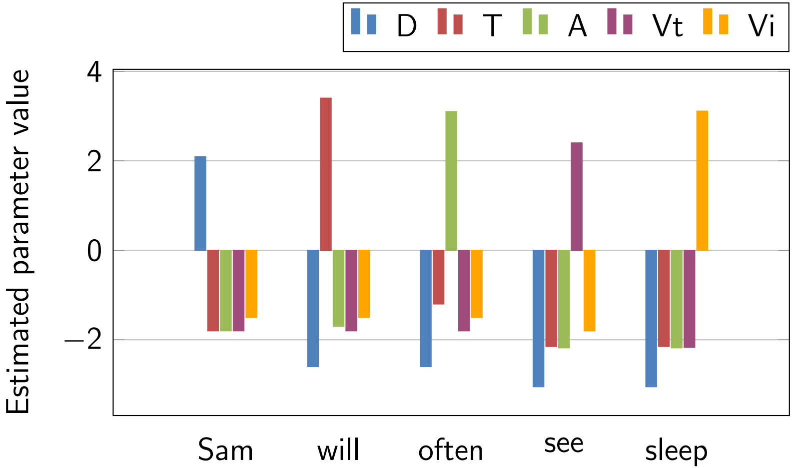
MACQUARIE UNIVERSITY


V initial
V final
V>T
¬ V>T
Estimated parameter value
2
0
−2
English
French
German

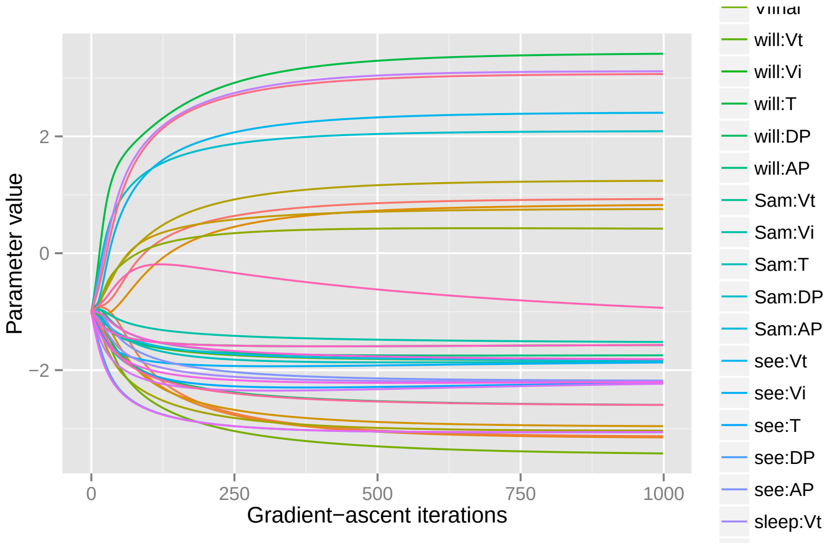


MACQUARIE UNIVERSITY

Lexical parameters for English

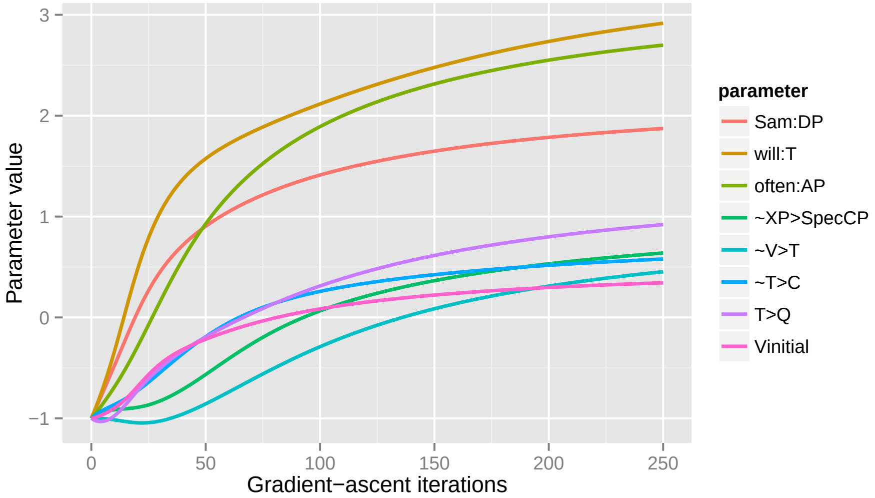


Learning English parameters



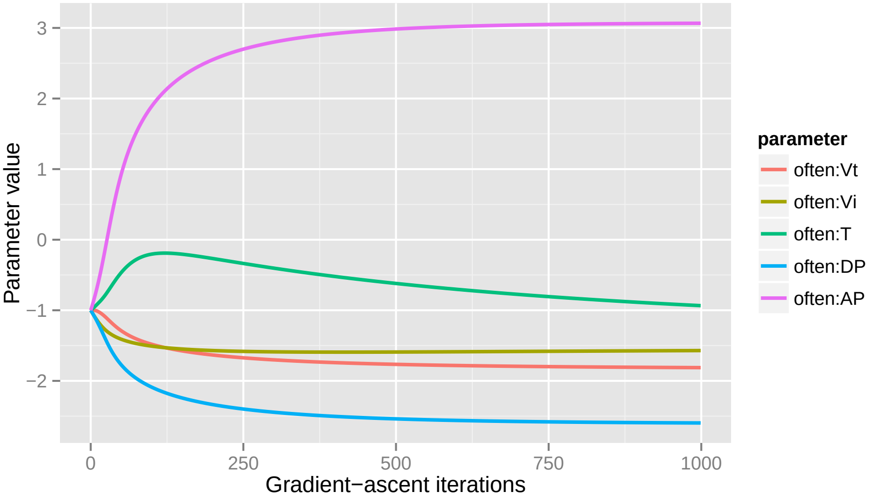
MACQUARIE UNIVERSITY

Learning English lexical and syntactic parameters



MACQUARIE UNIVERSITY

Learning "often" in English



Relation to other work

- Many other "toy" parameter-learning systems:
  - E.g., Yang (2002) describes an error-driven learner with templates triggering parameter value updates
  - we *jointly learn lexical categories and syntactic parameters*
- Error-driven learners like Yang's can be viewed as an approximation to the algorithm proposed here:
  - on-line error-driven parameter updates are a stochastic approximation to gradient-based hill-climbing
  - MG parsing is approximated with template matching

Relation to Harmonic Grammar and Optimality Theory

- Harmonic Grammars are MaxEnt models that associate weights with configurations much as we do here (Smolensky et al 1993)
  - because no constraints are placed on possible parameters or derivations, little detail about computation for parameter estimation
- Optimality Theory can be viewed as a discretised version of Harmonic Grammar in which *all parameter weights must be negative*
- MaxEnt models like these are widely used in phonology (Goldwater and Johnson 2003, Hayes and Wilson 2008)

Outline

Statistics and probabilistic models

Parameter-setting as parametric statistical inference

An example of syntactic parameter learning

Learning the lexicon as non-parametric inference

Synergies with syllabification

Grounded learning and learning word-topic associations

The role of social cues in word learning

Conclusions, and where do we go from here?

MACQUARIE UNIVERSITY

Why is learning the lexicon interesting?

- Tens/hundreds of thousands of *arbitrary form-meaning pairs* (as well as non-predictable syntactic properties)

$\Rightarrow$ Orders of magnitude more lexical parameters than syntactic parameters (?)

  - perhaps learning a language is mainly learning its vocabulary?

- Lexicalising some aspect of grammar doesn't automatically make it easy to learn

MACQUARIE UNIVERSITY

Why non-parametric models of the lexicon?

- A *parametric model* has a *pre-specified, finite set of parameters*
- A *non-parametric model* is one that can't be described using a finite pre-specified set of parameters
- We represent lexical forms as (structured) *sequences of phonemes* (ignore meanings for simplicity)
- While every lexicon is finite, there is no *universal* bound on the possible lexical forms (and meanings)
- *Bayesian non-parametric inference* can perform inference about *models that have unboundedly many parameters*
  - mathematically, our models have *a parameter for every possible lexical form*
  - although the models can't be directly represented, we can still make inferences about them
  - obvious idea: lexical entries we currently have no evidence for aren't explicitly represented

MACQUARIE UNIVERSITY

Language acquisition as Bayesian inference

$$\underbrace{\mathrm{P}(\text{Grammar} \mid \text{Data})}_{\text{Posterior}} \propto \underbrace{\mathrm{P}(\text{Data} \mid \text{Grammar})}_{\text{Likelihood}} \underbrace{\mathrm{P}(\text{Grammar})}_{\text{Prior}}$$

- Likelihood measures *how well grammar describes data*
- Prior expresses knowledge of grammar before data is seen
  - can be very specific (e.g., Universal Grammar)
  - can be very general (e.g., third factors, prefer shorter grammars)
- Prior can also express *markedness preferences* ("soft universals")
- Posterior is a *product* of both likelihood and prior
  - a grammar must do well on both to have high posterior probability
- Priors are especially important in non-parametric inference
  - "flat" priors are unavailable
- Bayesian inference is almost same as Minimum Description Length (MDL)

Nonparametric Bayesian inference with adaptor grammars

- Many of our models have a very similar structure:
  - a *generator* specifies possible entities (e.g., lexical entries)
  - the entries are *composed* to form the observed data
  - each entry's frequency is *adapted* based on the data
- "Adaptor grammars" are a framework for nonparametric Bayesian inference that uses probabilistic context-free grammars to specify the *possible entities* and the *way they combine*

Unsupervised word segmentation: a simplified lexical acquisition problem

- Input: phoneme sequences with *sentence boundaries* (Brent)
- Task: identify *word boundaries*, and hence *words*

j △ u ▲ w △ ɑ △ n △ t ▲ t △ u ▲ s △ i ▲ ð △ ə ▲ b △ ʊ △ k

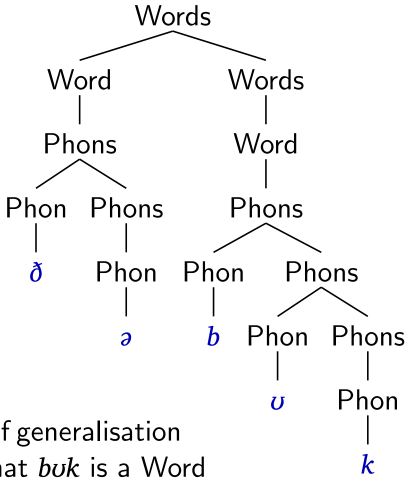
ju wɑnt tu si ðə bʊk

"you want to see the book"

- Ignoring phonology and morphology, this involves learning the pronunciations of the lexical items in the language

An attempt at a PCFG for word segmentation

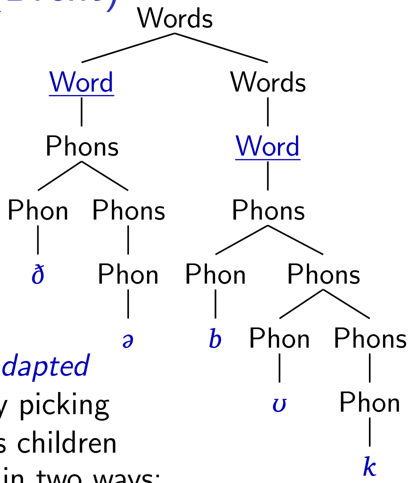
Words $\rightarrow$ Word
Words $\rightarrow$ Word Words
Word $\rightarrow$ Phons
Phons $\rightarrow$ Phon
Phons $\rightarrow$ Phon Phons
Phon $\rightarrow$ $a \mid b \mid \ldots$



- CFG trees can *describe* segmentation, but
- PCFGs *can't distinguish* good segmentations from bad ones
  - PCFG rules are *too small* a unit of generalisation
  - need to learn e.g., probability that *bʊk* is a Word

Unigram adaptor grammar (Brent)

Words → Word
Words → Word Words
<u>Word</u> → Phons
Phons → Phon
Phons → Phon Phons



- The trees generated are defined by CFG rules as in a CFG
- A subset of the nonterminals are *adapted*
- *Unadapted nonterminals* expand by picking a rule and recursively expanding its children
- *Adapted nonterminals* can expand in two ways:
  - by picking a rule and recursively expanding its children, or
  - by generating a previously generated tree (with probability proportional to the number of times previously generated)

Monte Carlo sampling from the posterior

- The number of possible words grows quickly as a function of the size of the training data
  - explicitly representing all possible words becomes infeasible
  - problem becomes worse as lexical entries become more complicated

⇒ *Sample* from the posterior distribution instead of explicitly representing all possible words and their parameters

- *Gibbs sampler* for adaptor grammars:
  - Initialise parses for each sentence (e.g., randomly) and extract a lexicon
  - Repeat until converged:

    Pick a sentence at random
    Remove the lexical entries in its parse from the lexicon
    Parse sentence using lexical entries learnt from other sentences
    Add lexical entries from this sentence to lexicon

- There are *incremental on-line* sampling algorithms as well (Börschinger and Johnson 2012)

Adaptor grammar learnt from Brent corpus

- **Initial grammar**

1	Words → Word Words	1	Words → Word
1	Word → Phon		
1	Phons → Phon Phons	1	Phons → Phon
1	Phon → D	1	Phon → G
1	Phon → A	1	Phon → E

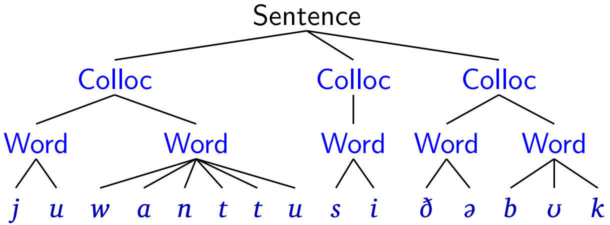
- **A grammar learnt from Brent corpus**

16625	Words → Word Words	9791	Words → Word
1575	Word → Phons		
4962	Phons → Phon Phons	1575	Phons → Phon
134	Phon → D	41	Phon → G
180	Phon → A	152	Phon → E
460	Word → (Phons (Phon y) (Phons (Phon u)))		
446	Word → (Phons (Phon w) (Phons (Phon A) (Phons (Phon t)))		
374	Word → (Phons (Phon D) (Phons (Phon 6)))		
372	Word → (Phons (Phon $\&$) (Phons (Phon n) (Phons (Phon d)))		

MACQUARIE UNIVERSITY

Collocations capture distributional properties

$$\begin{aligned}
\text{Sentence} &\rightarrow \text{Colloc}^{+} \\
\underline{\text{Colloc}} &\rightarrow \text{Word}^{+} \\
\underline{\text{Word}} &\rightarrow \text{Phon}^{+}
\end{aligned}$$



- A Colloc(ation) consists of one or more words
- Both Words and Collocs are adapted (learnt)
- Significantly improves word segmentation accuracy over unigram model (76% f-score; $\approx$ Goldwater's bigram model)

Outline

Statistics and probabilistic models

Parameter-setting as parametric statistical inference

An example of syntactic parameter learning

Learning the lexicon as non-parametric inference

Synergies with syllabification

Grounded learning and learning word-topic associations

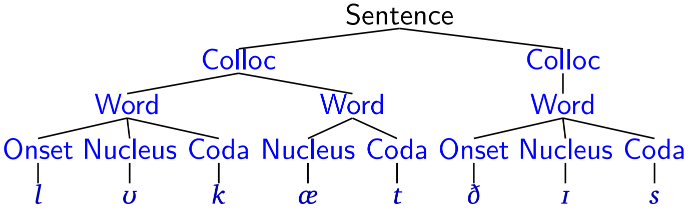
The role of social cues in word learning

Conclusions, and where do we go from here?

MACQUARIE UNIVERSITY

Jointly learning words and syllables

Sentence $\rightarrow$ Colloc^{+}	Colloc $\rightarrow$ Word^{+}
Word $\rightarrow$ Syllable$^{\{1:3\}}$	Syllable $\rightarrow$ (Onset) Rhyme
Onset $\rightarrow$ Consonant^{+}	Rhyme $\rightarrow$ Nucleus (Coda)
Nucleus $\rightarrow$ Vowel^{+}	Coda $\rightarrow$ Consonant^{+}



- Rudimentary syllable model (improved model does better)
- With 2 Collocation levels, f-score = 84%

MACQUARIE UNIVERSITY

Distinguishing internal onsets/codas helps

Sentence $\rightarrow$ Colloc^{+}

Word $\rightarrow$ SyllableIF

Word $\rightarrow$ SyllableI Syllable SyllableF

OnsetI $\rightarrow$ Consonant^{+}

Nucleus $\rightarrow$ Vowel^{+}

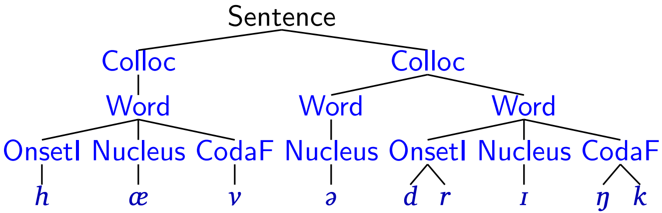
Colloc $\rightarrow$ Word^{+}

Word $\rightarrow$ SyllableI SyllableF

SyllableIF $\rightarrow$ (OnsetI) RhymeF

RhymeF $\rightarrow$ Nucleus (CodaF)

CodaF $\rightarrow$ Consonant^{+}



- With 2 Collocation levels, not distinguishing initial/final clusters, f-score = 84%
- With 3 Collocation levels, distinguishing initial/final clusters, f-score = 87%

MACQUARIE UNIVERSITY

Summary of word segmentation

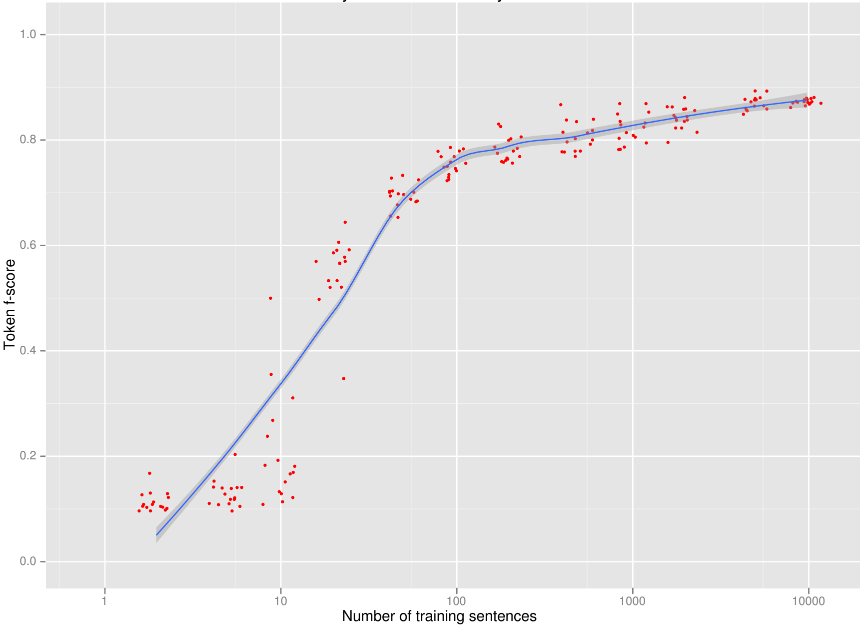
- Word segmentation accuracy depends on the kinds of generalisations learnt.

Generalization	**Accuracy**
words as units (unigram)	56%
+ associations between words (collocations)	76%
+ syllable structure	84%
+ interaction between segmentation and syllable structure	87%

- *Synergies in learning words and syllable structure*
  - joint inference permits the learner to *explain away* potentially misleading generalizations
- In recent work we've also included stress in English

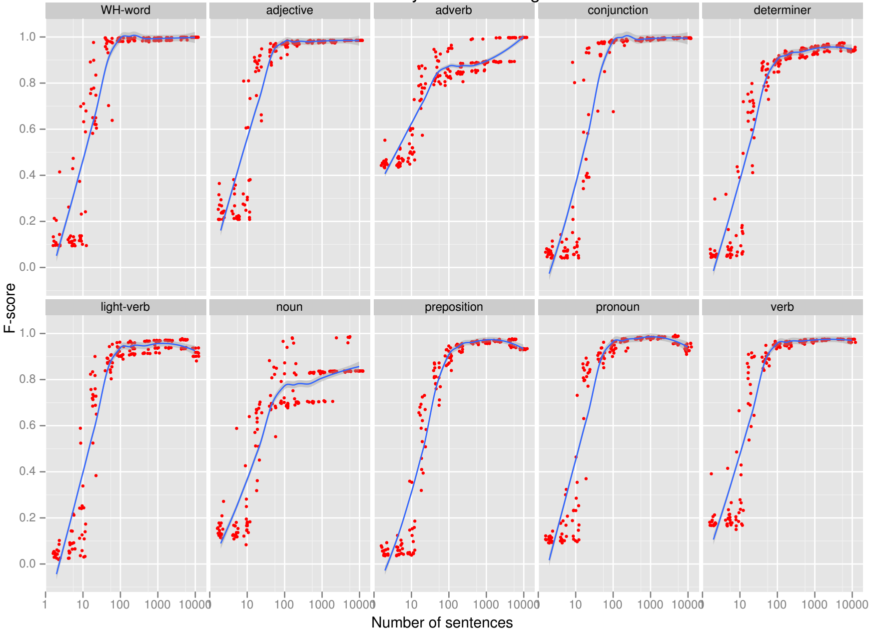
MACQUARIE UNIVERSITY

Accuracy of Collocation + Syllable model

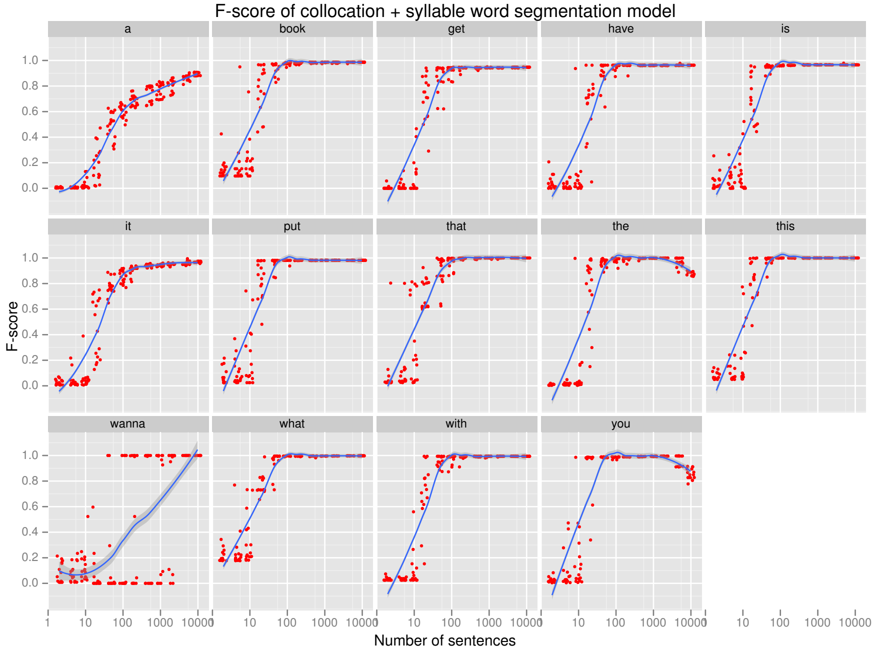


MACQUARIE UNIVERSITY

F-score of collocation + syllable word segmentation model



MACQUARIE UNIVERSITY



MACQUARIE UNIVERSITY

Outline

Statistics and probabilistic models

Parameter-setting as parametric statistical inference

An example of syntactic parameter learning

Learning the lexicon as non-parametric inference

Synergies with syllabification

Grounded learning and learning word-topic associations

The role of social cues in word learning

Conclusions, and where do we go from here?

MACQUARIE UNIVERSITY

Mapping words to referents

- Input to learner:
  - word sequence: *Is that the pig?*
  - objects in nonlinguistic context: DOG, PIG
- Learning objectives:
  - identify utterance topic: PIG
  - identify word-topic mapping: *pig* $\rightsquigarrow$ PIG

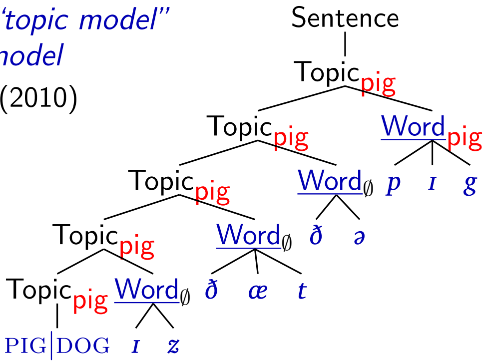
MACQUARIE UNIVERSITY

AGs for joint segmentation and topic-mapping

- Combine topic-model PCFG with word segmentation AGs
- Input consists of unsegmented phonemic forms prefixed with possible topics:

PIG|DOG *ɪ z ð æ t ð ə p ɪ g*

- E.g., combination of *Frank "topic model"* and *unigram segmentation model*
  - equivalent to Jones et al (2010)
- Easy to define *other combinations of topic models and segmentation models*



Does non-linguistic context help segmentation?

Model		word segmentation
segmentation	**topics**	**token f-score**
unigram	not used	0.533
unigram	any number	0.537
unigram	one per sentence	0.547
collocation	not used	0.695
collocation	any number	0.726
collocation	one per sentence	0.719
collocation	one per collocation	**0.750**

- Not much improvement with unigram model
  - consistent with results from Jones et al (2010)
- Larger improvement with collocation model
  - most gain with *one topical word per topical collocation* (this constraint cannot be imposed on unigram model)

MACQUARIE UNIVERSITY

Does better segmentation help topic identification?

- Task: identify object (if any) *this sentence* is about

Model		sentence referent	
segmentation	**topics**	**accuracy**	**f-score**
unigram	not used	0.709	0
unigram	any number	0.702	0.355
unigram	one per sentence	0.503	0.495
collocation	not used	0.709	0
collocation	any number	0.728	0.280
collocation	one per sentence	0.440	0.493
collocation	one per collocation	**0.839**	**0.747**

- The collocation grammar with *one topical word per topical collocation* is the only model clearly better than baseline

MACQUARIE UNIVERSITY

Does better segmentation help learning word-to-referent mappings?

- Task: identify *head nouns* of NPs referring to topical objects (e.g. *pɪg* ⇝ PIG in input PIG | DOG *ɪ z ð æ t ð ə p ɪ g*)

Model		**topical word**
segmentation	**topics**	**f-score**
unigram	not used	0
unigram	any number	0.149
unigram	one per sentence	0.147
collocation	not used	0
collocation	any number	0.220
collocation	one per sentence	0.321
collocation	one per collocation	**0.636**

- The collocation grammar with one topical word per topical collocation is best at identifying head nouns of referring NPs

MACQUARIE UNIVERSITY

Summary of grounded learning and word segmentation

- *Word to object mapping is learnt more accurately when words are segmented more accurately*
  - improving segmentation accuracy improves topic detection and acquisition of topical words
- *Word segmentation accuracy improves when exploiting non-linguistic context information*
  - incorporating word-topic mapping improves segmentation accuracy (at least with collocation grammars)

$\Rightarrow$ *There are synergies a learner can exploit when learning word segmentation and word-object mappings*

Outline

Statistics and probabilistic models

Parameter-setting as parametric statistical inference

An example of syntactic parameter learning

Learning the lexicon as non-parametric inference

Synergies with syllabification

Grounded learning and learning word-topic associations

The role of social cues in word learning

Conclusions, and where do we go from here?

Why study social cues?

- Everyone agrees social interactions are important for children's early language acquisition
  - e.g. children who engage in more joint attention with caregivers (e.g., looking at toys together) learn words faster (Carpenter 1998)
- *Can computational models exploit social cues?*
  - we show this by building models that can exploit social cues, and show they *learn better on data with social cues than on data with social cues removed*
- Many different social cues could be relevant: *can our models learn the importance of different social cues?*
  - our models estimate *probability of each cue occuring with "topical objects"* and *probability of each cue occuring with "non-topical objects"*
  - they do this in an unsupervised way, i.e., they are not told which objects are topical

Exploiting social cues for learning word topics

- Frank et al (2012) corpus of 4,763 utterances containing:
  - the orthographic words uttered by the care-giver,
  - a set of *available topics* (i.e., objects in the non-linguistic objects),
  - the values of the social cues, and
  - a set of *intended topics*, which the care-giver refers to.
- Social cues annotated in corpus:

Social cue	**Value**
child.eyes	objects child is looking at
child.hands	objects child is touching
mom.eyes	objects care-giver is looking at
mom.hands	objects care-giver is touching
mom.point	objects care-giver is pointing to

- Frank et al (2012) give extensive information on corpus

Exploiting social cues in word learning

- In the four different models we tried, *social cues* improved the accuracy of:
  - recovering the *utterance topic*
  - identifying the *word(s) referring to the topic*, and
  - *learning a lexicon* (word $\rightsquigarrow$ topic mapping)
- *kideyes* was the most important social cue for each of these tasks in all of the models
- Social cues don't seem to improve word segmentation
- Luong, Frank and Johnson (2013) extend the model to capture *topic continuity across sentences*
  - further improves model's accuracy

MACQUARIE UNIVERSITY

Outline

Statistics and probabilistic models

Parameter-setting as parametric statistical inference

An example of syntactic parameter learning

Learning the lexicon as non-parametric inference

Synergies with syllabification

Grounded learning and learning word-topic associations

The role of social cues in word learning

Conclusions, and where do we go from here?

Statistical inference for syntactic parameters

- *No inherent contradiction between probabilistic models, statistical inference and grammars*
- Statistical inference can be used to *set real-valued parameters* (learn empty functional categories) in Minimalist Grammars (MGs)
  - parameters are local in context-free derivation structures
    $\Rightarrow$ efficient computation
  - can solve "chicken-and-egg" learning problems
  - does not need negative evidence
- Not a *tabula rasa* learner
  - depends on a rich inventory of prespecified parameters

Nonparametric Bayesian inference for the lexicon

- Capable of learning word pronunciations from unsegmented input
- Easy to extend model so it exploits additional information:
  - distributional information (collocations)
  - syllable structure and stress information
  - "topic" information about the relationship between words and objects in the non-linguistic context
  - social cues (e.g., which object the care-giver is looking at)
  - intersentential "topic" dependencies
- Produces state-of-the-art accuracies on a variety of languages

MACQUARIE UNIVERSITY

Future directions in word learning

- Word learning models are now well developed:
  - compare models' predictions to real developmental profiles
  - what combination of information sources produces most realistic results?
- Extend models to handle more realistic input:
  - phonological variation between underlying and surface form (Elsner et al 2012, Börschinger et al 2013)
  - continuous (acoustic?) features as input
- Extend models to incorporate morphology, syntax and semantics:
  - not difficult in principle
  - are syntactic theories capable of handling child-directed speech?
  - will we need "disfluency" models?

MACQUARIE UNIVERSITY

Future directions in syntactic parameter acquisition

- *Are real-valued parameters linguistically reasonable?*
- Does algorithm "scale up" to realistic grammars and corpora?
  - parsing and inference components use efficient dynamic programming algorithms
  - many informal proposals, but no "universal" MGs (perhaps start with well-understood families like Romance?)
  - generally disappointing results scaling up PCFGs (de Marken 1995)
  - but our grammars lack so much (e.g., LF movement, binding)
- Exploit semantic information in the non-linguistic context
  - e.g., learn from surface forms paired with their logical form semantics (Kwiatkowski et al 2012)
  - but what is the "language of thought"?
- Use a nonparametric Bayesian model to *learn the empty functional categories of a language* (c.f., Bisk and Hockenmaier 2013)

MACQUARIE UNIVERSITY

Why probabilistic models?

- Probabilistic models are a *computational level* description
  - they define the relevant variables and dependencies between them
- Models are stated at a *higher level of abstraction* than algorithms:
  - ⇒ easier to see how to incorporate additional dependencies (e.g., non-linguistic context)
- There are standard ways of constructing inference algorithms for probabilistic models:
  - usually multiple algorithms for same model with different properties (e.g., incremental, on-line)
- My opinion: *it's premature to focus on algorithms*
  - identify relevant variables and their dependencies first!
  - *optimal inference procedures* let us explore consequences of a model *without committing to any particular algorithm*

MACQUARIE UNIVERSITY

How might statistics change linguistics?

- Few examples where probabilistic models/statistical inference provides crucial insights
  - role of negative evidence in learning
  - statistical inference compatible with conventional parameter setting
- Non-parametric inference can learn which parameters are relevant
  - needs a generative model or “grammar” of possible parameters
  - but probability theory is generally agnostic as to parameters
- Probabilistic models have more relevance to psycholinguistics and language acquisition
  - these are *computational* processes
  - explicit computational models can make predictions about the *time course* of these processes

MACQUARIE UNIVERSITY

This research was supported by Australian Reseach Council Discovery Projects DP110102506 and DP110102593.

Paper and slides available from http://science.MQ.edu.au/˜mjohnson

Interested in computational linguistics and its relationship to linguistics, language acquisition or neurolinguistics? *We're recruiting PhD students!*

Contact me or anyone from Macquarie University for more information.

MACQUARIE UNIVERSITY