Synergies in Language Acquisition

Mark Johnson

joint work with Benjamin Börschinger, Katherine Demuth, Michael Frank, Sharon Goldwater, Tom Griffiths and Bevan Jones

Macquarie University
Sydney, Australia

Outline

How can computational models help us understand language acquisition?

- Most computational linguistics research focuses on parsing or learning *algorithms*
- A *computational model* (Marr 1982) of acquisition specifies:
  - the input (information available to learner)
  - the output (generalisations learner can make)
  - a model that relates input to output
- This talk compares:
  - *staged learning*, which learns one kind of thing at a time, and
  - *joint learning*, which learns several kinds of things simultaneously,

  and demonstrates *synergies in acquisition* that only joint learners exploit
- We do this by *comparing models that differ solely in the kinds of generalisations they can form*

Bayesian learning as an "ideal observer" theory of learning

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

- 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., 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
- Posterior is a *distribution* over grammars
  - captures *learner's uncertainty* about which grammar is correct

The acquisition of the lexicon as non-parametric inference

- What has to be learned in order to learn a word?
  - pronunciation (sequence of phonemes)
  - syntactic properties
  - semantic properties (what kinds of things it can refer to)

  There are *unboundedly many* different possible pronunciations (and possible meanings?)
- **Parametric inference:** learn values of a *finite number* of parameters
- **Non-parametric inference:**
  - possibly infinite number of parameters
  - learn which parameters are relevant as well as their values
- *Adaptor grammars* use a grammar to generate parameters for learning (e.g., possible lexical items)
  - builds on *non-parametric hierarchical Bayesian inference*

Outline

Unsupervised word segmentation

- 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"

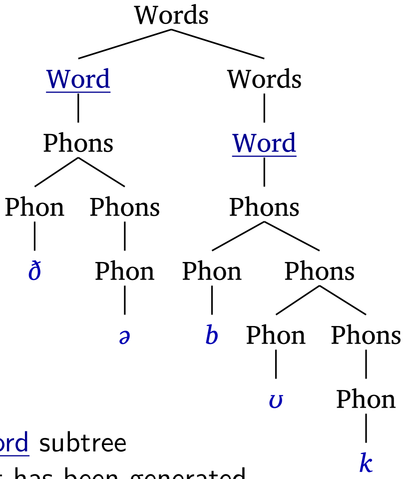
- Ignoring phonology and morphology, this involves learning the pronunciations of the lexical items in the language

Adaptor grammars as non-parametric hierachical Bayesian models

- The trees generated by an adaptor grammar 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, as in a PCFG
- *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)
- *Adaptor Grammars generalise from types rather than tokens* at all levels

Unigram adaptor grammar (Brent)

Words → Word
Words → Word Words
Word → Phons
Phons → Phon
Phons → Phon Phons



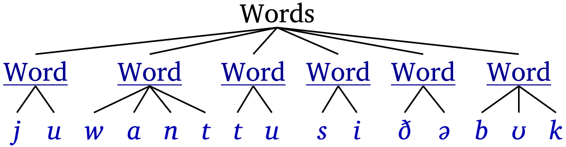
- Word nonterminal is adapted

⇒ To generate a Word:
  - select a previously generated Word subtree with prob. $\propto$ number of times it has been generated
  - expand using Word → Phons rule with prob. $\propto \alpha_{\text{Word}}$ and recursively expand Phons

Unigram model of word segmentation

- Unigram "bag of words" model (Brent):
  - generate a *dictionary*, i.e., a set of words, where each word is a random sequence of phonemes
    - Bayesian prior prefers smaller dictionaries
  - generate each utterance by choosing each word at random from dictionary
- Brent's unigram model as an adaptor grammar:

Words $\to$ Word$^+$
<u>Word</u> $\to$ Phoneme$^+$



- Accuracy of word segmentation learnt: *56% token f-score* (same as Brent model)
- But we can construct many more word segmentation models using AGs

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*

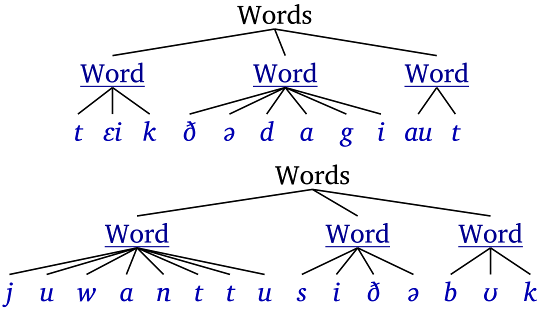
- **A grammar learnt from Brent corpus**

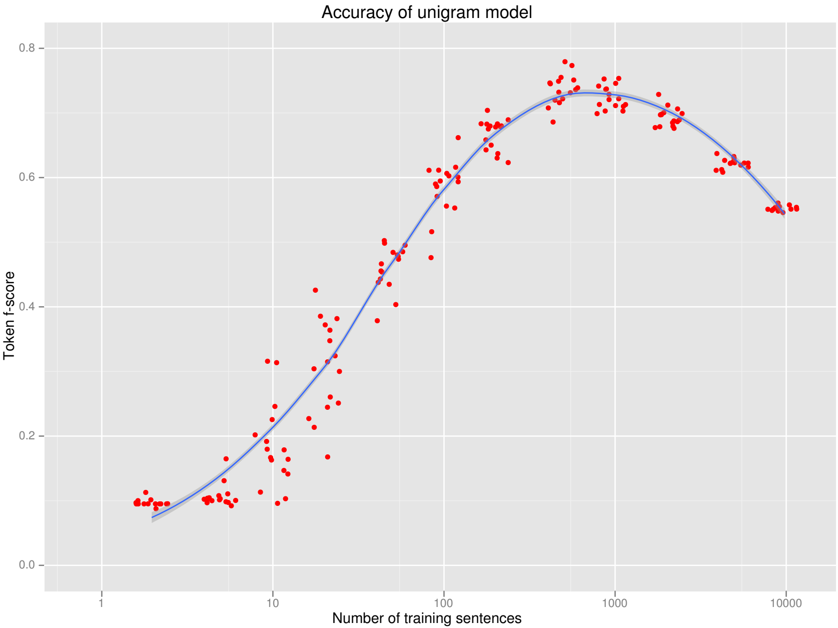
16625	Words → Word Words	9791	Words → Word
1575	Word → Phons		
4962	Phons → Phon Phons	1575	Phons → Phon
134	Phon → *D*	41	Phon → *G*
180	Phon → *A*	152	Phon → *E*
460	Word → (Phons (Phon *y*) (Phons (Phon *u*)))		
446	Word → (Phons (Phon *w*) (Phons (Phon *A*) (Phons (Phon *t*)))		
374	Word → (Phons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

Undersegmentation errors with Unigram model

$$\text{Words} \rightarrow \underline{\text{Word}}^{+} \qquad \underline{\text{Word}} \rightarrow \text{Phon}^{+}$$

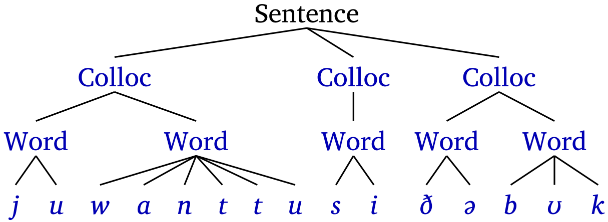
- Unigram word segmentation model assumes each word is generated independently
- But there are strong inter-word dependencies (collocations)
- Unigram model can only capture such dependencies by analyzing collocations as words (Goldwater 2006)




Accuracy of unigram model
Token f-score
0.0
0.2
0.4
0.6
0.8
1
10
100
1000
10000
Number of training sentences

Collocations ⇒ Words

$$\text{Sentence} \to \text{Colloc}^+$$
$$\underline{\text{Colloc}} \to \text{Word}^+$$
$$\underline{\text{Word}} \to \text{Phon}^+$$



- 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; ≈ Goldwater's bigram model)

Outline

Two hypotheses about language acquisition

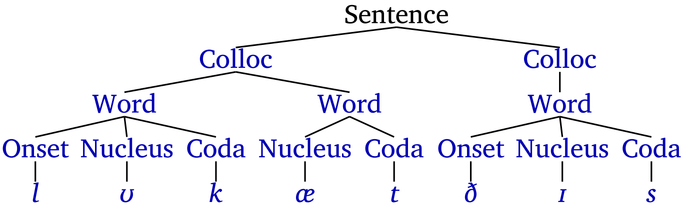
1. Pre-programmed *staged acquisition* of linguistic components
   - Conventional view of *lexical acquisition*, e.g., Kuhl (2004)
     - child first learns the phoneme inventory, which it then uses to learn
     - phonotactic cues for word segmentation, which are used to learn
     - phonological forms of words in the lexicon, ...
2. *Interactive acquisition* of all linguistic components together
   - corresponds to *joint inference* for all components of language
   - stages in language acquisition might be due to:
     - child's input may contain more information about some components
     - some components of language may be learnable with less data

Synergies: an advantage of interactive learning

- An *interactive learner* can take advantage of *synergies in acquisition*
  - partial knowledge of component A provides information about component B
  - partial knowledge of component B provides information about component A
- A staged learner can only take advantage of one of these dependencies
- An interactive or *joint learner* can benefit from a positive feedback cycle between A and B
- Are there synergies in *learning how to segment words* and *identifying the referents of words*?

Jointly learning words and syllables

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

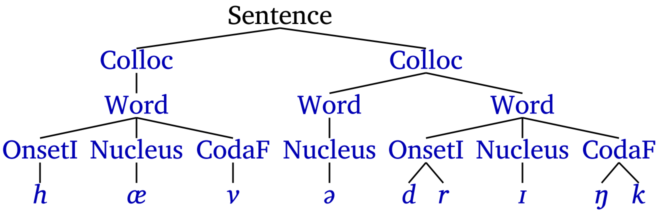


- Rudimentary syllable model (an improved model might do better)
- With 2 Collocation levels, f-score = 84%

Distinguishing internal onsets/codas helps

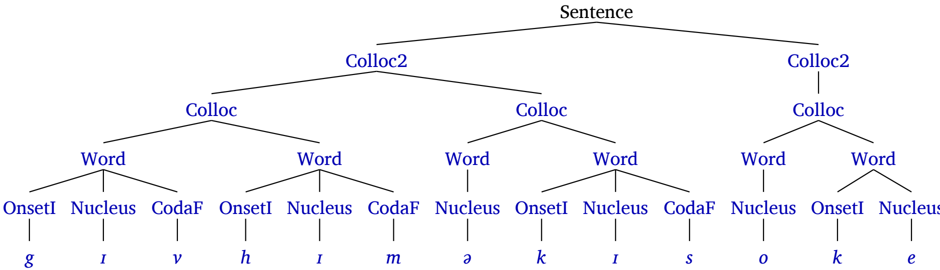
Sentence → Colloc$^+$
Word → SyllableIF
Word → SyllableI Syllable SyllableF
OnsetI → Consonant$^+$
Nucleus → Vowel$^+$

Colloc → Word$^+$
Word → SyllableI SyllableF
SyllableIF → (OnsetI) RhymeF
RhymeF → Nucleus (CodaF)
CodaF → Consonant$^+$



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

Collocations^2 ⇒ Words ⇒ Syllables

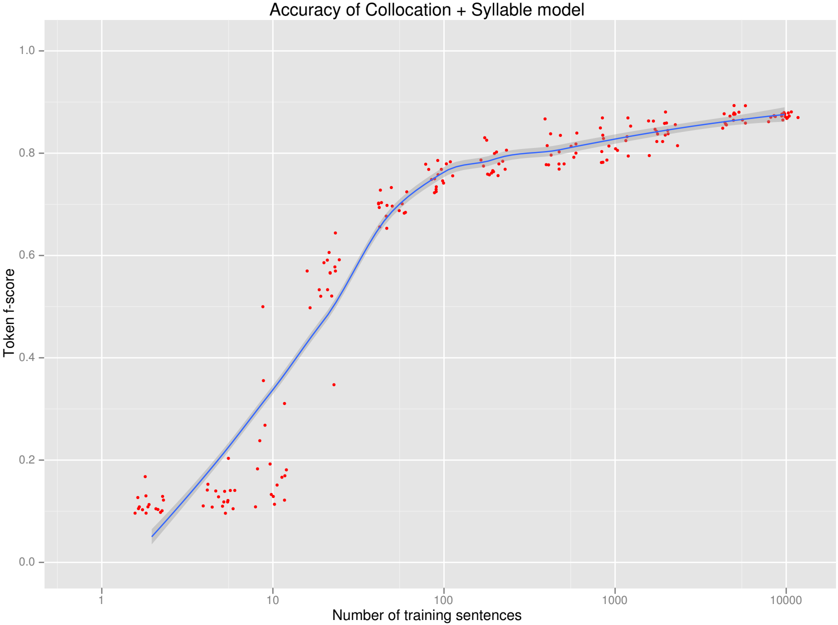


Interaction between syllable phonotactics and segmentation

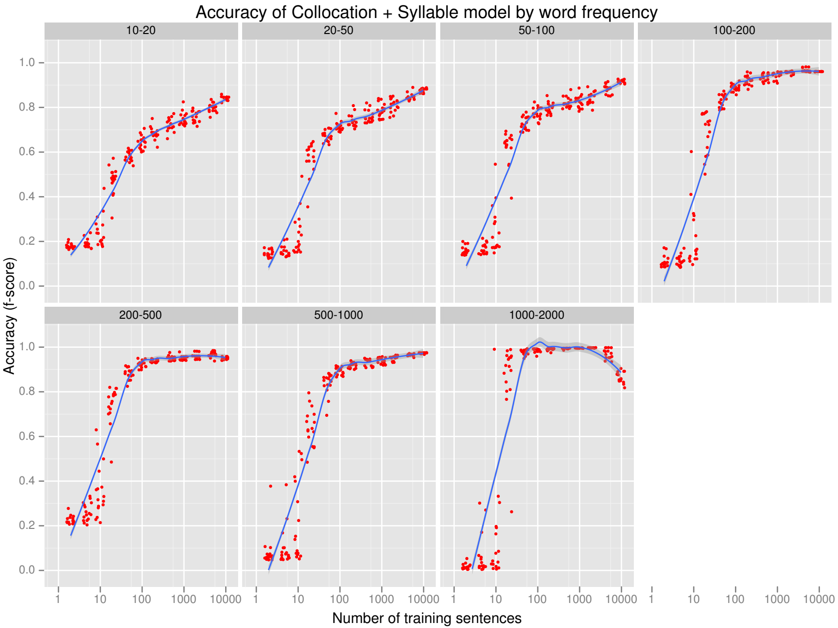
- Word segmentation accuracy depends on the kinds of generalisations the model can learn

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

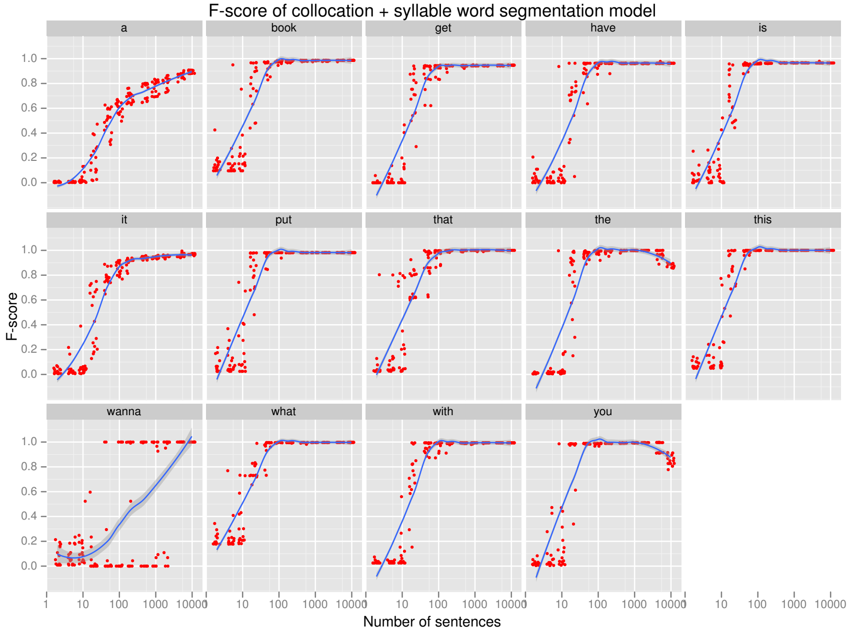
- *Synergies in learning words and syllable structure*
  - joint inference permits the learner to *explain away* potentially misleading generalizations

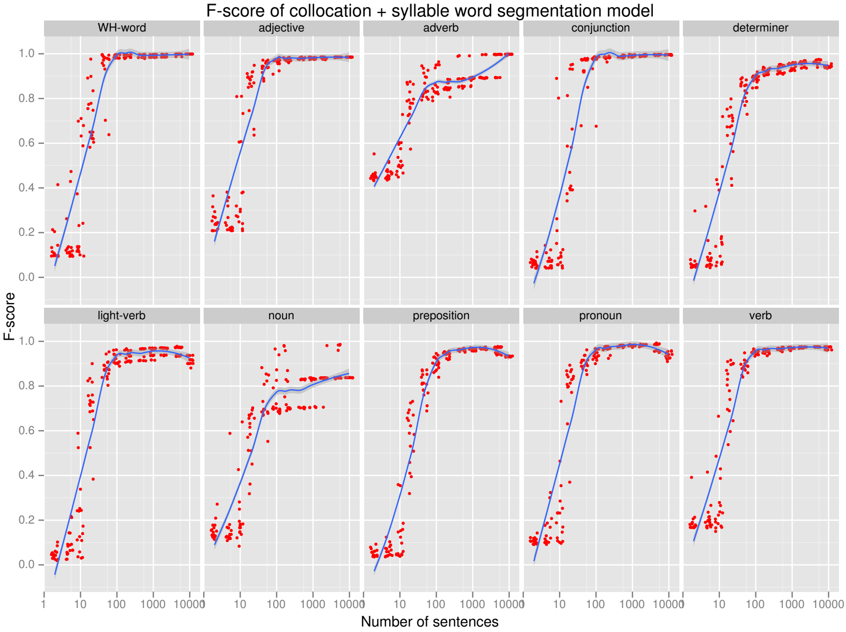


Accuracy of Collocation + Syllable model
Token f-score
0.0
0.2
0.4
0.6
0.8
1.0
1
10
100
1000
10000
Number of training sentences



Accuracy of Collocation + Syllable model by word frequency
10-20
20-50
50-100
100-200
200-500
500-1000
1000-2000
Accuracy (f-score)
0.0
0.2
0.4
0.6
0.8
1.0
1
10
100
1000
10000
Number of training sentences


F-score of collocation + syllable word segmentation model
a
book
get
have
is
it
put
that
the
this
wanna
what
with
you
F-score
1.0
0.8
0.6
0.4
0.2
0.0
1
10
100
1000
10000
Number of sentences


F-score of collocation + syllable word segmentation model
WH-word
adjective
adverb
conjunction
determiner
light-verb
noun
preposition
pronoun
verb
F-score
Number of sentences
0.0
0.2
0.4
0.6
0.8
1.0
1
10
100
1000
10000

Outline

Exploiting stress in word segmentation

- Stress is the “accentuation of syllables within words”
  - 2-syllable words with initial stress: *GIant, PICture, HEAting*
  - 2-syllable words with final stress: *toDAY, aHEAD, aLLOW*
- English has a *strong preference for initial stress* (Cutler 1987)
  - 50% of tokens / 85% of types have initial stress
  - but: 50% of tokens / 5% of types are unstressed
- Strong evidence that English-speaking children use stress for word segmentation
- Data preparation: stress marked on vowel nucleus

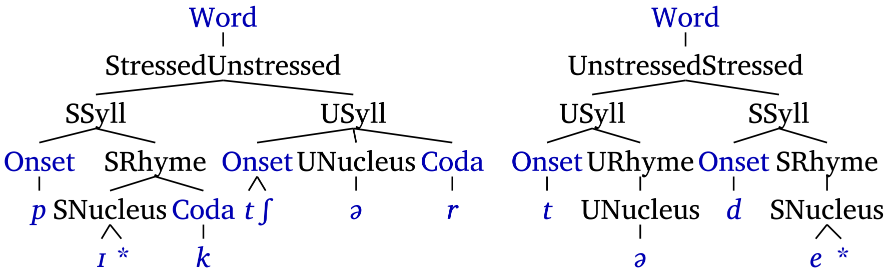
$$j \,{}_{\vartriangle}\, u \,{}_{\blacktriangle}\, w \,{}_{\vartriangle}\, a^* \,{}_{\vartriangle}\, n \,{}_{\vartriangle}\, t \,{}_{\blacktriangle}\, t \,{}_{\vartriangle}\, u \,{}_{\blacktriangle}\, s \,{}_{\vartriangle}\, i^* \,{}_{\blacktriangle}\, ð \,{}_{\vartriangle}\, ə \,{}_{\blacktriangle}\, b \,{}_{\vartriangle}\, ʊ^* \,{}_{\vartriangle}\, k$$

“you want to see the book”

  - c.f. Johnson and Demuth (2010) tone annotation in Chinese
  - function words are unstressed (contra Yang and others)

Learning stress patterns with AGs

- Grammar can represent all possible stress patterns (up to 4 syllables)
- Stress pattern probabilities *learned jointly with phonotactics and segmentation*



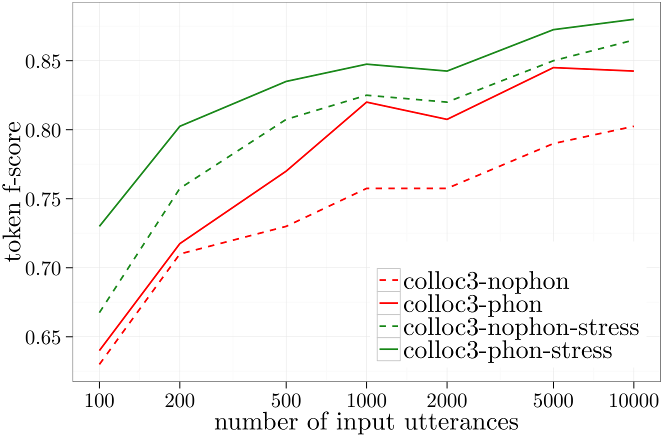
Stress and phonotactics in word segmentation

- Models differ only in *kinds of generalisations* they can form
  - phonotactic models learn generalisations about word edges
  - stress models learn probability of strong/weak sequences

Model	Accuracy
collocations + syllable structure	0.81
+ phonotactic cues	0.85
+ stress	0.86
+ both	**0.88**

- Token f-score on the *Alex portion of the Providence corpus*
- *Both phonotactics and stress are useful cues for word segmentation*
- Performance improves when both are used ⇒ complementary cues for word segmentation

Stress and phonotactics over time



- Joint stress+phonotactic model is best with small data
- Models with either eventually catch up

More on learning stress

- Probability of *initial stress* and *unstressed* word rules rapidly converges on their *type frequencies* in the data
- Consistently underestimates probability of *stress-second patterns* (true type frequency = 0.07, estimated type frequency = 0.04)
  - stress-second is also problematic for English children
- Probability of word rules with more than one stress approaches zero as data grows

  $\Rightarrow$ *Unique stress constraint* (Yang 2004) can be acquired

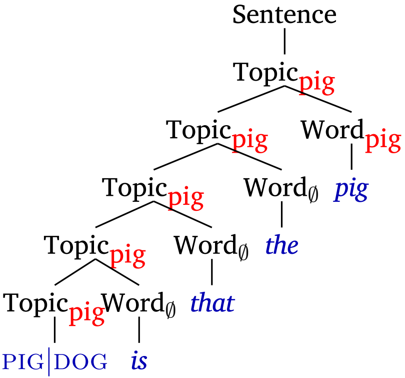
Outline

Prior work: mapping words to topics

- 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* $\mapsto$ PIG

Frank et al (2009) "topic models" as PCFGs

- Prefix sentences with *possible topic marker*, e.g., PIG|DOG
- PCFG rules *choose a topic* from topic marker and *propagate it through sentence*
- Each word is either generated from sentence topic or null topic $\emptyset$



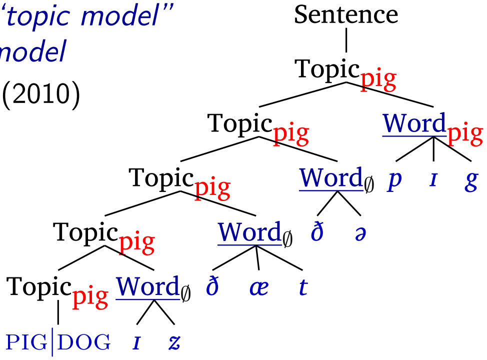
- Grammar can require *at most one topical word per sentence*
- Bayesian inference for PCFG rules and trees corresponds to Bayesian inference for word and sentence topics using topic model (Johnson 2010)

AGs for joint segmentation and topic-mapping

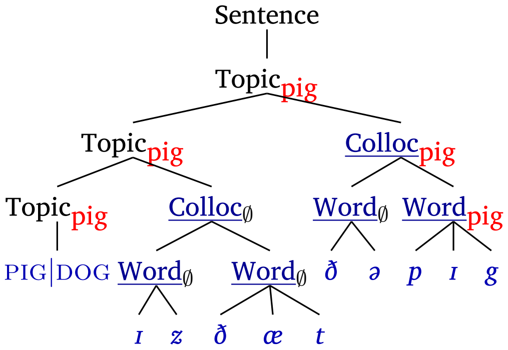
- Combine topic-model PCFG with word segmentation AGs
- Input consists of unsegmented phonemic forms prefixed with possible topics:

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*



Collocation topic model AG



- Collocations are either "topical" or not
- Easy to modify this grammar so
  - at most one topical word per sentence, or
  - at most *one topical word per topical collocation*

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)

Does better segmentation help topic identification?

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

Model		sentence topic	
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**

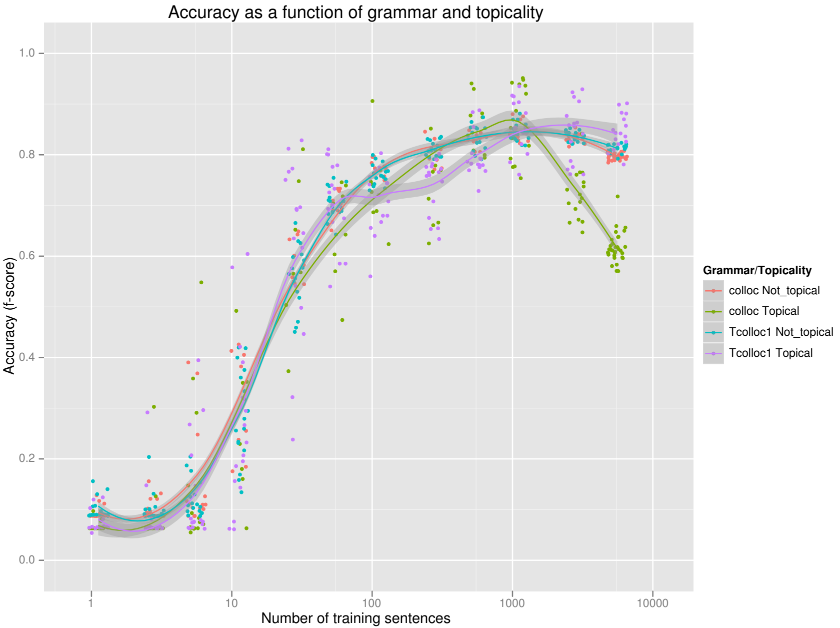
- The collocation grammar with *one topical word per topical collocation* is the only model clearly better than baseline

Does better segmentation help learning word-topic mappings?

- Task: identify *head nouns* of NPs referring to topical objects (e.g. *pɪg* $\mapsto$ 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 topical NPs



Accuracy as a function of grammar and topicality
Accuracy (f-score)
0.0
0.2
0.4
0.6
0.8
1.0
1
10
100
1000
10000
Number of training sentences
Grammar/Topicality
colloc Not_topical
colloc Topical
Tcolloc1 Not_topical
Tcolloc1 Topical

Summary of jointly learning word segmentation and word-to-topic mappings

- *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

Social cues and word-topic mapping

- Social interactions are important for early language acquisition
- *Can computational models exploit social cues?*
  - we show this by building models that can exploit social cues, and show they *learns better word-topic mappings on data with social cues than when social cues are removed*
  - no evidence that social cues improve word-segmentation accuracy
- Our models learn *relative importance of different social cues*
  - 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
  - ablation tests show that *eye-gaze* is the most important social cue for learning word-topic mappings

Function words in word segmentation

- Some psychologists believe children exploit function words in early language acquisition
  - function words often are high frequency and phonologically simple $\Rightarrow$ easy to learn?
  - function words typically appear in phrase-peripheral positions $\Rightarrow$ provide "anchors" for word and phrase segmentation
- Modify word segmentation grammar to optionally generate *sequences of mono-syllabic "function words" at collocation edges*
  - *improves word segmentation f-score from 0.87 to 0.92*
- Model can learn directionality of function word attachment
  - Bayes factor hypothesis test *overwhelmingly prefers left to right function word attachment in English*

Jointly learning word segmentation and phonological alternation

- Word segmentation models so far don't account for phonological alternations (e.g., final devoicing, /t/-deletion, etc.)
  - fundamental operation in CFG is string concatenation
  - no principled reason why adaptor grammars can't be combined with phonological operations
- Börschinger, Johnson and Demuth (2013) generalises Goldwater's bigram word segmentation model to allow word-final /t/-deletion
  - applied to Buckeye corpus of adult speech
- Current work: incorporate a MaxEnt "Harmony theory" model of phonological alternation

On-line learning using particle filters

- The Adaptor Grammar software uses a batch algorithm that repeatedly parses the data
  - in principle, all the algorithm requires is a source of random samples from the training data
  - $\Rightarrow$ Adaptor Grammars can be learned on-line
- *Particle filters* are a standard technique for on-line Bayesian inference
  - a particle filter updates multiple analyses in parallel
- Börschinger and Johnson (2011, 2012) explore on-line particle filter algorithms for Goldwater's bigram model
  - a particle filter needs tens of thousands of particles to approach the Metropolis-within-Gibbs algorithm used for Adaptor Grammars here
  - adding a *rejuvenation step* reduces the number of particles needed dramatically

Synergy failure: morphology and word segmentation

- We haven't found a synergy between morphology and word segmentation
  - failure to find $\not\Rightarrow$ non-existence
- Why might we not have found any synergy?
  - *no synergies exist:*
    - morphological acquisition is largely independent of word segmentation
  - *wrong data:*
    - child-directed English doesn't contain enough inflectional morphology to be useful
    - $\Rightarrow$ study languages with richer inflectional morphology
  - *wrong models:*
    - our models didn't learn morpho-phonology, which plays a big role in English
    - $\Rightarrow$ extend MaxEnt Harmony-theory models of word segmentation and phonology to include morphology

Outline

Conclusions and future work

- *Joint learning* often uses information in the input more effectively than staged learning
  - Learning syllable structure and word segmentation
  - Learning word-topic associations and word segmentation
- *Do children exploit such synergies in language acquisition?*
- Adaptor grammars are a flexible framework for stating non-parametric hierarchical Bayesian models
  - the accuracies obtained here are the best reported in the literature
- Future work: make the models more realistic
  - extend expressive power of AGs (e.g., incorporating MaxEnt/Harmony-theory components)
  - richer data (e.g., more non-linguistic context)
  - more realistic data (e.g., phonological variation)
  - *cross-linguistic research* (we've applied our models to French, Sesotho and Chinese)

How specific should our computational models be?

- **Marr's (1982) three levels of computational models:**
  - *computational level* (inputs, outputs and relation between them)
  - *algorithmic level* (steps involved in mapping from input to output)
  - *implementation level* (physical processes involved)
- Algorithmic-level models are extremely popular, but I think we should focus on computational-level models first
  - we know almost nothing about *how hierarchical structures are represented and manipulated in the brain*
  - ⇒ we know almost nothing about which data structures and operations are neurologically plausible
  - current models only explain a tiny fraction of language processing or acquisition
  - typically computational models can be extended, while algorithms need to be completely changed
  - ⇒ *today's computational models have a greater chance of being relevant than today's algorithms*

Why a child's learning algorithm may be nothing like our algorithms

- Enormous differences in "hardware" $\Rightarrow$ different feasible algorithms
- As scientists we need *generic learning algorithms*, but a child only needs a *specific learning algorithm*
  - as scientists we want to study the effects of different modelling assumptions on learning
  - $\Rightarrow$ we need generic algorithms that work for a range of different models, so we can compare them
  - a child only needs an algorithm that works for whatever model they have
  - $\Rightarrow$ the child's algorithm might be specialised to their model, and need not work at all for other kinds of models
- The field of *machine learning* has developed many generic learning algorithms: Expectation-maximisation, variational Bayes, Markov chain Monte Carlo, Gibbs samplers, particle filters, ...

The future of Bayesian models of language acquisition

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

- So far our grammars and priors don't encode much linguistic knowledge, but in principle they can!
  - how do we represent this knowledge?
  - how can we learn efficiently using this knowledge?
- Should permit us to *empirically investigate effects of specific universals on the course of language acquisition*
- My guess: the interaction between innate knowledge and learning will be *richer and more interesting* than either the rationalists or empiricists currently imagine!