Unsupervised phonemic Chinese word segmentation using Adaptor Grammars

Mark Johnson[1] and Katherine Demuth[2]

[1]Department of Computing

[2]Department of Linguistics

Macquarie University
Sydney
Australia

COLING, August 2010

MACQUARIE UNIVERSITY

Talk outline

- Adaptor grammars are a framework for expressing *non-parametric hierarchical Bayesian models*
- They can be used to define *unsupervised word segmentation* models that learn:
  - *word-internal structure:* how words are composed out of syllables, and
  - *inter-word structure:* collocational dependencies between words
- Adaptor Grammars provide state-of-the-art unsupervised segmentation results for English: *will they work for Mandarin Chinese?*
  - can Adaptor Grammars model *lexical tone*?
  - does modelling lexical tone *improve word segmentation accuracy*?

Why study computational models of language acquisition?

- Hypothesis: acquisition, comprehension and production are *computational processes*
  - computational models need not be just *descriptions* of language acquisition
  - a computational model should be able to *learn a language*
- Characterising computational models of acquisition:
  - the input (information available to learner)
  - the output (generalisations learner can make)
  - the algorithm used to map input to output
- *Bayesian inference algorithms* are optimal learners
  - computational generalisation of "ideal observer" theory
- Computational models let us study the effect of
  - changing the information in the input, and
  - altering the kinds of generalisations the learner can acquire

  in ways that would be impractical or unethical with real children
- *May be useful for designing experiments or theraputic interventions*

Unsupervised word segmentation

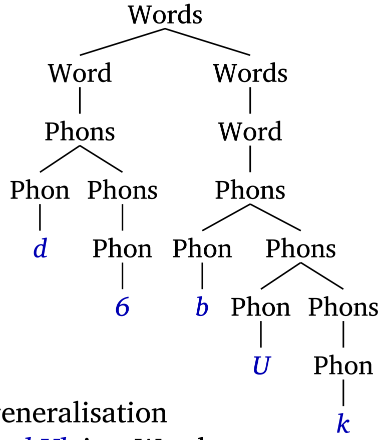
- Input: phoneme sequences with *sentence boundaries* (Brent)
  - English data produced from orthographic transcripts of child-directed speech by *looking up each word in a pronouncing dictionary*
- Task: identify *word boundaries*, and hence words, in unsegmented utterance (in ARPABET)

$$\mathrm{y}_{\vartriangle}\mathrm{u}_{\blacktriangle}\mathrm{w}_{\vartriangle}\mathrm{a}_{\vartriangle}\mathrm{n}_{\vartriangle}\mathrm{t}_{\blacktriangle}\mathrm{t}_{\vartriangle}\mathrm{u}_{\blacktriangle}\mathrm{s}_{\vartriangle}\mathrm{i}_{\blacktriangle}\mathrm{D}_{\vartriangle}\mathrm{6}_{\blacktriangle}\mathrm{b}_{\vartriangle}\mathrm{U}_{\vartriangle}\mathrm{k}$$

- Useful cues for word segmentation:
  - Phonotactics and syllable structure (Fleck)
  - Inter-word dependencies (Goldwater)

CFG models of word segmentation

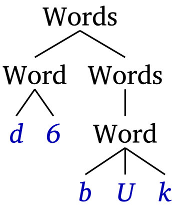
Words → Word
Words → Word Words
Word → Phons
Phons → Phon
Phons → Phon Phons
Phon → $a \mid b \mid \ldots$



- 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 *bUk* is a Word

Towards non-parametric grammars

Words → Word
Words → Word Words
Word → *all possible phoneme sequences*



- Learn probability Word → b U k
- But *infinitely many possible Word expansions*
  ⇒ this grammar is *not a PCFG*
- Given *fixed training data*, only finitely many useful rules
  ⇒ use data to choose Word rules as well as their probabilities
- Non-parametric models: parameters of model depend on data

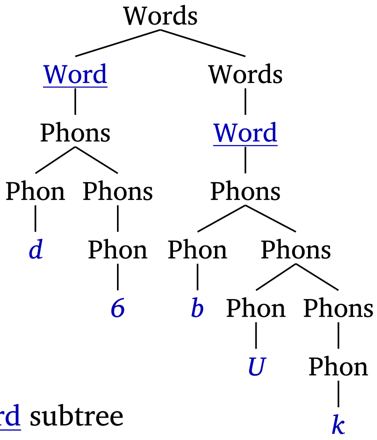
From PCFGs to Adaptor grammars

- An adaptor grammar is a PCFG where a subset of the nonterminals are *adapted*
- **Adaptor grammar generative process:**
  - to expand an *unadapted nonterminal* B: (just as in PCFG)
    - select a *rule* $B \to \beta \in R$ with prob. $\theta_{B \to \beta}$, and recursively expand nonterminals in β
  - to expand an *adapted nonterminal* B:
    - select a *previously generated subtree* T_B with prob. $\propto$ number of times T_B was generated, or
    - select a *rule* $B \to \beta \in R$ with prob. $\propto \alpha_B \, \theta_{B \to \beta}$, and recursively expand nonterminals in β

MACQUARIE UNIVERSITY

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

Properties of adaptor grammars

- Probability of regenerating an adapted subtree T_B
  $\propto$ number of times T_B was previously generated
  - adapted subtrees are *not independent*
    - an adapted subtree can be *more probable* than the rules used to construct it
  - but they are *exchangable* $\Rightarrow$ efficient sampling algorithms
  - "rich get richer" $\Rightarrow$ Zipf power-law distributions
- Each adapted nonterminal is associated with a
  Chinese Restaurant Process or *Pitman-Yor Process*
  - CFG rules define *base distribution* of CRP or PYP
- CRP/PYP parameters (e.g., α_B) can themselves be estimated (e.g., slice sampling)

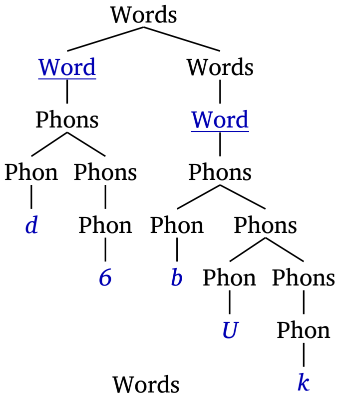
MACQUARIE UNIVERSITY

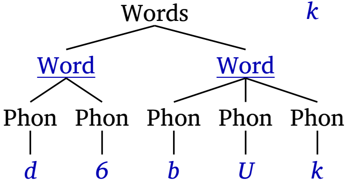
Abbreviatory notation

Words $\rightarrow$ Word
Words $\rightarrow$ Word Words
<u>Word</u> $\rightarrow$ Phons
Phons $\rightarrow$ Phon
Phons $\rightarrow$ Phon Phons

is abbreviated as

Words $\rightarrow$ Word^{+}
<u>Word</u> $\rightarrow$ Phon^{+}

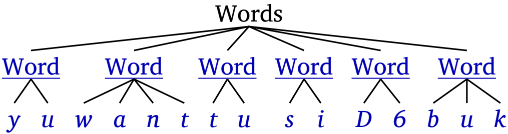




Unigram model of word segmentation

- 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

$$\text{Words} \to \underline{\text{Word}}^{+}$$
$$\underline{\text{Word}} \to \text{Phon}^{+}$$



- 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**

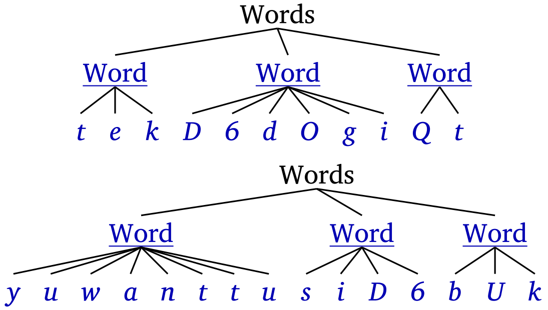
16625	Words → Word Words	9791	Words → Word
1575	Word → Phons		
4962	Phons → Phon Phons	1575	Phons → Phon
134	Phon → *D*	41	Phon → *G*
180	Phon → *A*	152	Phon → *E*
460	Word → (Phons (Phon *y*) (Phons (Phon *u*)))		
446	Word → (Phons (Phon *w*) (Phons (Phon *A*) (Phons (Phon *t*))))		
374	Word → (Phons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

MACQUARIE UNIVERSITY

Undersegmentation errors with Unigram model

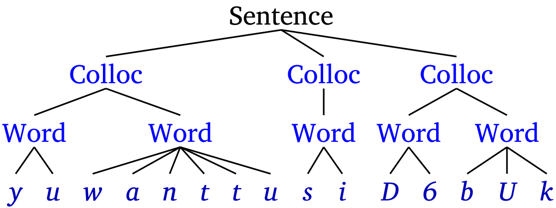
$$\text{Words} \rightarrow \underline{\text{Word}}^{+} \qquad \underline{\text{Word}} \rightarrow \text{Phon}^{+}$$

- 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)



Collocations $\Rightarrow$ Words

Sentence $\rightarrow$ Colloc^{+}
Colloc $\rightarrow$ Word^{+}
Word $\rightarrow$ Phon^{+}



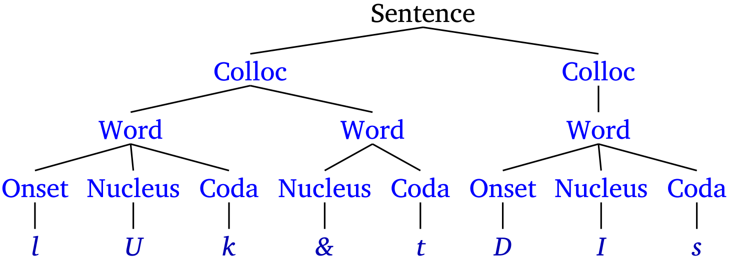
- A Colloc(ation) consists of one or more words
- Both Words and Collocs are adapted (learnt)
- Significantly improves word segmentation accuracy over unigram model (74% f-score; $\approx$ Goldwater's bigram model)

MACQUARIE UNIVERSITY

Collocations ⇒ Words ⇒ Syllables

Sentence → Colloc$^+$
Word → Syllable
Word → Syllable Syllable Syllable
Onset → Consonant$^+$
Nucleus → Vowel$^+$

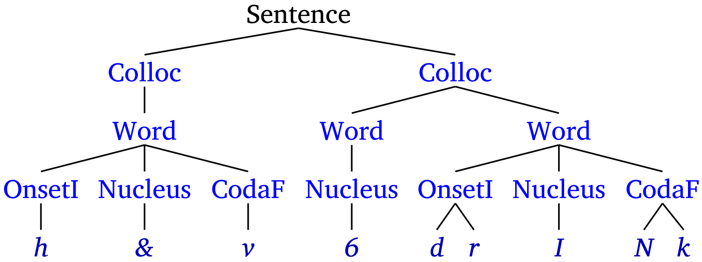
Colloc → Word$^+$
Word → Syllable Syllable
Syllable → (Onset) Rhyme
Rhyme → Nucleus (Coda)
Coda → Consonant$^+$



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

Distinguishing internal onsets/codas helps in English

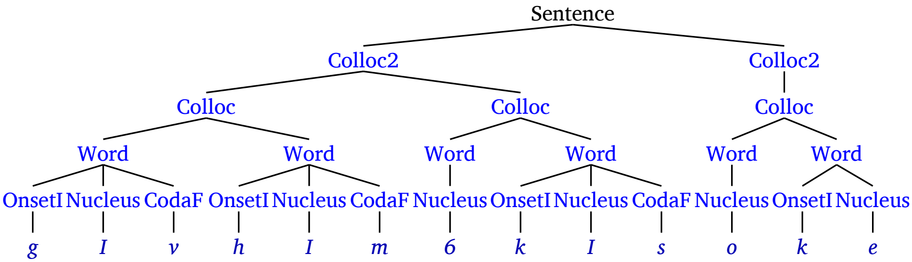
$\text{Sentence} \rightarrow \text{Colloc}^{+}$
$\underline{\text{Word}} \rightarrow \text{SyllableIF}$
$\underline{\text{Word}} \rightarrow \text{SyllableI Syllable SyllableF}$
$\underline{\text{OnsetI}} \rightarrow \text{Consonant}^{+}$
$\underline{\text{Nucleus}} \rightarrow \text{Vowel}^{+}$
$\underline{\text{Colloc}} \rightarrow \text{Word}^{+}$
$\underline{\text{Word}} \rightarrow \text{SyllableI SyllableF}$
$\text{SyllableIF} \rightarrow (\text{OnsetI})\ \text{RhymeF}$
$\text{RhymeF} \rightarrow \text{Nucleus}\ (\text{CodaF})$
$\underline{\text{CodaF}} \rightarrow \text{Consonant}^{+}$



- 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

Collocations^2 ⇒ Words ⇒ Syllables



Summary so far

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

Generalization	Accuracy
words as units (unigram)	56%
+ associations between words (collocations)	79%
+ syllable structure	87%

- *Word segmentation accuracy improves when you learn other things as well*
  - *explain away* potentially misleading generalizations

Tone in Mandarin Chinese word segmentation

- Tone in Mandarin Chinese provides an additional dimension of information to the language learner
- It is necessary in order to distinguish lexical items, but how important is it for word segmentation?
- Approach:
  - construct a pair of otherwise identical corpora, one that contains tone and one that does not
  - run identical learning algorithms on both corpora
  - compare the accuracy with which each learns word segmentation

Mandarin Chinese corpus

- Used Tardif (1993) "Beijing" corpus (in Pinyin format)
  - deleted all "Child" utterances, and utterances with codes $INTERJ, $UNINT, $VOC and $PRMPT
  - corpus contains 50,118 utterances, consisting of 187,533 word tokens

    zen3me gei3 ta1 bei1 shang4 lai2 (1.) ?
    ta1: (.) a1yi2 gei3 de (.) ta1 gei3 de .
    hen3 jian3dan1 .
- Used a Pinyin to IPA translation program to produce IPA format

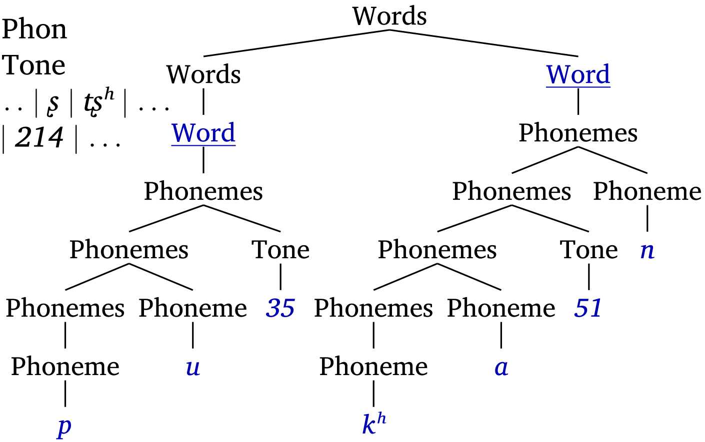
    tsən^{214}mɤ kei^{214} tʰa^{55} pei^{55} ʂɑŋ^{51} lai^{35}
    tʰa^{55} a^{55}i^{35} kei^{214} tɤ tʰa^{55} kei^{214} tɤ
    xən^{214} tɕiɛn^{214}tan^{55}
- Moved tones from end of syllable to preceding vowel

    ts ə ^{214} n m ɤ k e i ^{214} tʰ a ^{55} p e i ^{55} ʂ ɑ ^{51} ŋ l ai ^{35}
    tʰ a ^{55} a ^{55} i ^{35} k e i ^{214} t ɤ tʰ a ^{55} k e i ^{214} t ɤ
    x ə ^{214} n tɕ iɛ ^{214} n t a ^{55} n
- (Optionally delete tones)

MACQUARIE UNIVERSITY

Unigram word segmentation adaptor grammar

Words → Words Word
Words → Word
Word → Phons
Phons → Phon
Phons → Phons Phon
Phons → Phons Tone
Phon → *ai* | *o* | . . . | *ʂ* | *tʂʰ* | . . .
Tone → *35* | *55* | *214* | . . .



MACQUARIE UNIVERSITY

Collocation adaptor grammars

- Adaptor grammars with one level of collocation:

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

- Adaptor grammars with two levels of collocation:

$$\begin{array}{lll} \text{Colloc2s} \rightarrow \underline{\text{Colloc2}}^+ & \underline{\text{Colloc2}} \rightarrow \text{Collocs}^+ & \\ \text{Collocs} \rightarrow \underline{\text{Colloc}}^+ & \underline{\text{Colloc}} \rightarrow \text{Words} & \text{Words} \rightarrow \underline{\text{Word}}^+ \end{array}$$

- We experiment with up to three levels of collocation here

MACQUARIE UNIVERSITY

Syllable structure adaptor grammars

- No distinction between word-internal and word-peripheral syllables

$$\begin{array}{ll}
\underline{\text{Word}} \to \text{Syll} & \underline{\text{Word}} \to \text{Syll Syll} \\
\underline{\text{Word}} \to \text{Syll Syll Syll} & \underline{\text{Word}} \to \text{Syll Syll Syll Syll} \\
\text{Syll} \to (\underline{\text{Onset}})^? \ \underline{\text{Rhy}} & \underline{\text{Onset}} \to \text{C}^+ \\
\underline{\text{Rhy}} \to \underline{\text{Nucleus}} \ (\underline{\text{Coda}})^? & \underline{\text{Nucleus}} \to \text{V} \ (\text{V} \mid \text{Tone})^\star \\
\underline{\text{Coda}} \to \text{C}^+ & \text{C} \to \textit{ʂ} \mid \textit{tʂ}^h \mid \ldots \\
\text{V} \to \textit{ai} \mid \textit{o} \mid \ldots &
\end{array}$$

- Distinguishing word-internal and word-peripheral syllables

$$\begin{array}{ll}
\underline{\text{Word}} \to \text{SyllIF} & \underline{\text{Word}} \to \text{SyllI SyllF} \\
\underline{\text{Word}} \to \text{SyllI Syll SyllF} & \underline{\text{Word}} \to \text{SyllI Syll Syll SyllF} \\
\text{SyllIF} \to (\underline{\text{OnsetI}})^? \ \underline{\text{RhyF}} & \text{SyllI} \to (\underline{\text{OnsetI}})^? \ \underline{\text{Rhy}} \\
\text{SyllF} \to (\underline{\text{OnsetI}})^? \ \underline{\text{RhyF}} & \text{Syll} \to (\underline{\text{Onset}})^? \ \underline{\text{Rhy}} \\
\underline{\text{OnsetI}} \to \text{C}^+ & \underline{\text{RhyF}} \to \underline{\text{Nucleus}} \ (\underline{\text{CodaF}})^? \\
\underline{\text{CodaF}} \to \text{C}^+ &
\end{array}$$

MACQUARIE UNIVERSITY

Mandarin Chinese word segmentation results

- Word segmentation accuracy when input *contains tones*

	Syllables		
	None	General	Specialised
Unigram	0.57	0.50	0.50
Colloc	0.69	0.67	0.67
Colloc^2	0.72	0.75	0.75
Colloc^3	*0.64*	**0.77**	**0.77**

- Word segmentation accuracy when *tones are removed* from input

	Syllables		
	None	General	Specialised
Unigram	0.56	0.46	0.46
Colloc	0.70	0.65	0.65
Colloc^2	0.74	0.74	0.73
Colloc^3	0.75	0.76	**0.77**

MACQUARIE UNIVERSITY

Comparable English results

- English word segmentation results

	Syllables		
	None	General	Specialised
Unigram	0.56	0.46	0.46
Colloc	0.74	0.67	0.66
Colloc^2	0.79	0.84	0.84
Colloc^3	0.74	0.82	**0.87**

MACQUARIE UNIVERSITY

Discussion of Mandarin Chinese word segmentation results

- Mandarin Chinese word segmentation results broadly consistent with English results
  - unigram segmentation accuracies are similiar
  - results for other models are lower than corresponding English results
- General improvement in accuracy as number of collocation levels increases
- Caveats: the English and Mandarin Chinese corpora are not directly comparable
  - Discourse context for Mandarin Chinese corpus was far more diverse than for English corpus
  - Mandarin Chinese children were older than English children

Syllable structure and word segmentation

- Syllable structure and phonotactic constraints are very useful for English word segmentation, but are much less useful in Mandarin Chinese
  - perhaps surprising, because Mandarin Chinese has a very regular syllable structure
  - but perhaps this very predictability makes it less useful for identifying words?
  - not surprising that distinguishing word-peripheral syllables does not help, as Mandarin Chinese does not distinguish these

Tone and word segmentation

- Tones only have a small impact on segmentation accuracy
  - surprising, as they are required for lexical disambiguation
  - tones make a small improvement to simpler models (Unigram, Colloc) but no improvement with the more complex ones
    - perhaps tone is redundant given the inter-word context modelled by the Colloc$^{2-3}$ grammars?
- *Perhaps there's a better way to represent tones in the input, or use tones in the model?*
  - Neutral tones more common on function words — perhaps this can improve segmentation accuracy?
  - Tone sandhi may give information about phonological word boundaries

Conclusion and future work

- The adaptor grammar approach to word segmentation generalises to Mandarin Chinese
- Modelling inter-word dependencies (collocations) greatly improves word segmentation accuracy in Mandarin Chinese (as in English)
- Modelling syllable structure improves segmentation accuracy by a smaller amount in Mandarin Chinese (compared to English)
- Modelling tones improves segmentation accuracy of simpler models, but not of more complex models
- Future work:
  - Comparable multi-lingual corpora of infant-directed speech
  - More realistic, richer corpora (including multi-stratal input representations)
  - Model context-sensitive dependencies (e.g., phonological rules)

Interested in computational linguistics or its applications?

We're recruiting *PhD students*!.

Contact *Mark.Johnson@mq.edu.au* or *Katherine.Demuth@mq.edu.au* for more information.