From Lexicon to Grammar in Infant Word Segmentation

Constantine Lignos

University of Pennsylvania
Department of Computer and Information Science
Institute for Research in Cognitive Science

LSA Annual Meeting 2013

1/5/2013

Word segmentation is an *unsexy* problem.

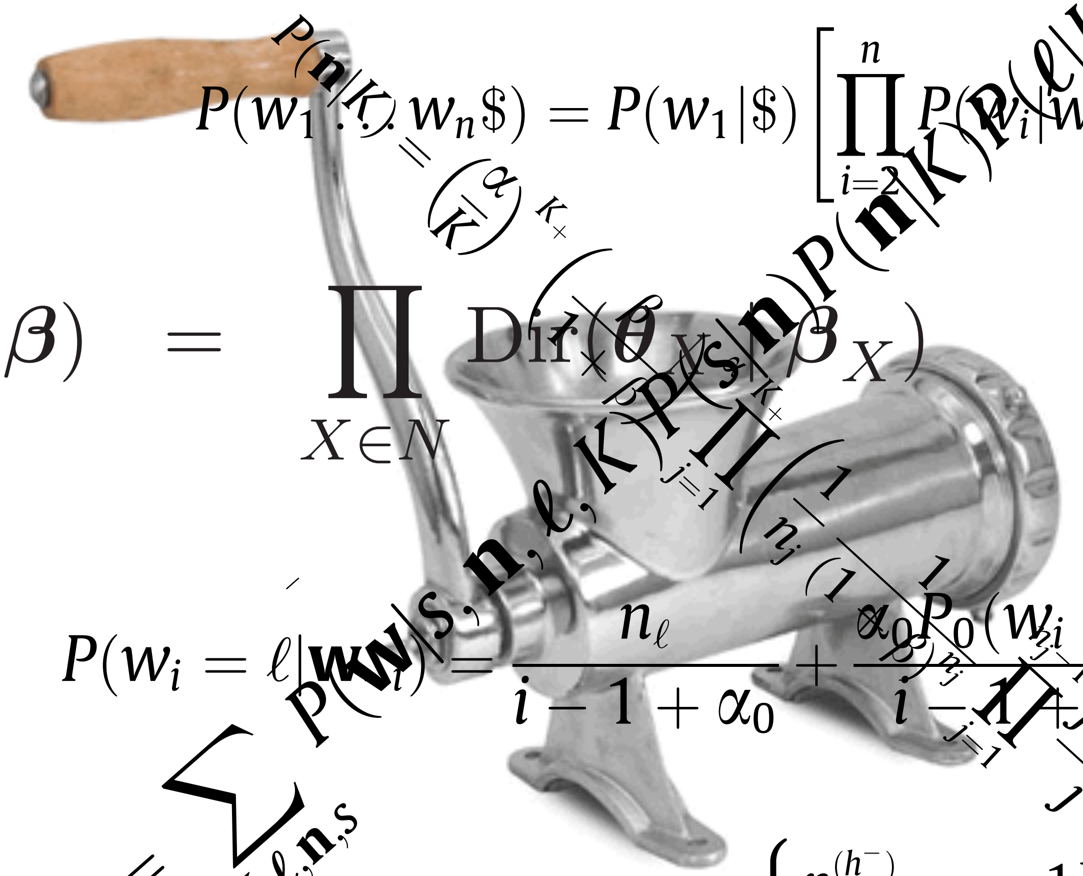
So unsexy it received little attention in early modern accounts of acquisition.

(Brown, 1973; Clark, 1974; MacWhinney, 1978)

Computational modeling of word segmentation hasn't made it any sexier or, more importantly, relevant to linguistics.

Whereareyougoing?
Howdoesabunnyrabbitwalk?
Doeshewalklikeyouordoeshegohophophop?
Whatareyoudoing?
Sweepbroom.
Isthatabroom?
Ithoughtitwasabrush.

Adam’s mother (Brown, 1973)



$$P(w_1 \ldots w_n \$) = P(w_1|\$)\left[\prod_{i=2}^{n} P(w_i|w_{i-1})\right]P(\$|w_n)$$

$$P(\mathbf{n}|K) = \left(\frac{\alpha}{K}\right)^{K_+}\left(\frac{\beta}{1-\beta}\right)^{K_+}\prod_{j=1}^{K_+}\left(\frac{1}{n_j}\frac{1}{(1-\beta)^{n_j}}\prod_{j=1}^{n_j-1}\frac{j+\frac{\alpha}{K}}{j}\right)$$

$$P(\boldsymbol{\theta} \mid \boldsymbol{\beta}) \;=\; \prod_{X \in N} \mathrm{Dir}(\boldsymbol{\theta}_X \mid \boldsymbol{\beta}_X)$$

$$P(\mathbf{w}) = \sum_{K,\ell,\mathbf{n},\varsigma} P(\mathbf{w}|\varsigma,\mathbf{n},\ell,K)P(\varsigma|\mathbf{n})P(\mathbf{n}|K)P(\ell|K)P(K)$$

$$P(w_i = \ell|\mathbf{w}_{-i}) = \frac{n_\ell}{i-1+\alpha_0} + \frac{\alpha_0 P_0(w_i = \ell)}{i-1+\alpha_0}$$

$$P(z_i = k|w_i, h^-) \propto \begin{cases} n_k^{(h^-)} & 1 \leqslant k \leqslant K\left(w_i^{(h^-)}\right) \\ \alpha_1 P_1(w_i) & k = K\left(w_i^{(h^-)}\right) + 1 \end{cases}$$

Where are you going?
How does a bunny rabbit walk?
Does he walk like you or does he go hop hop hop?
What are you doing?
Sweep broom.
Is that a broom?
I thought it was a brush.

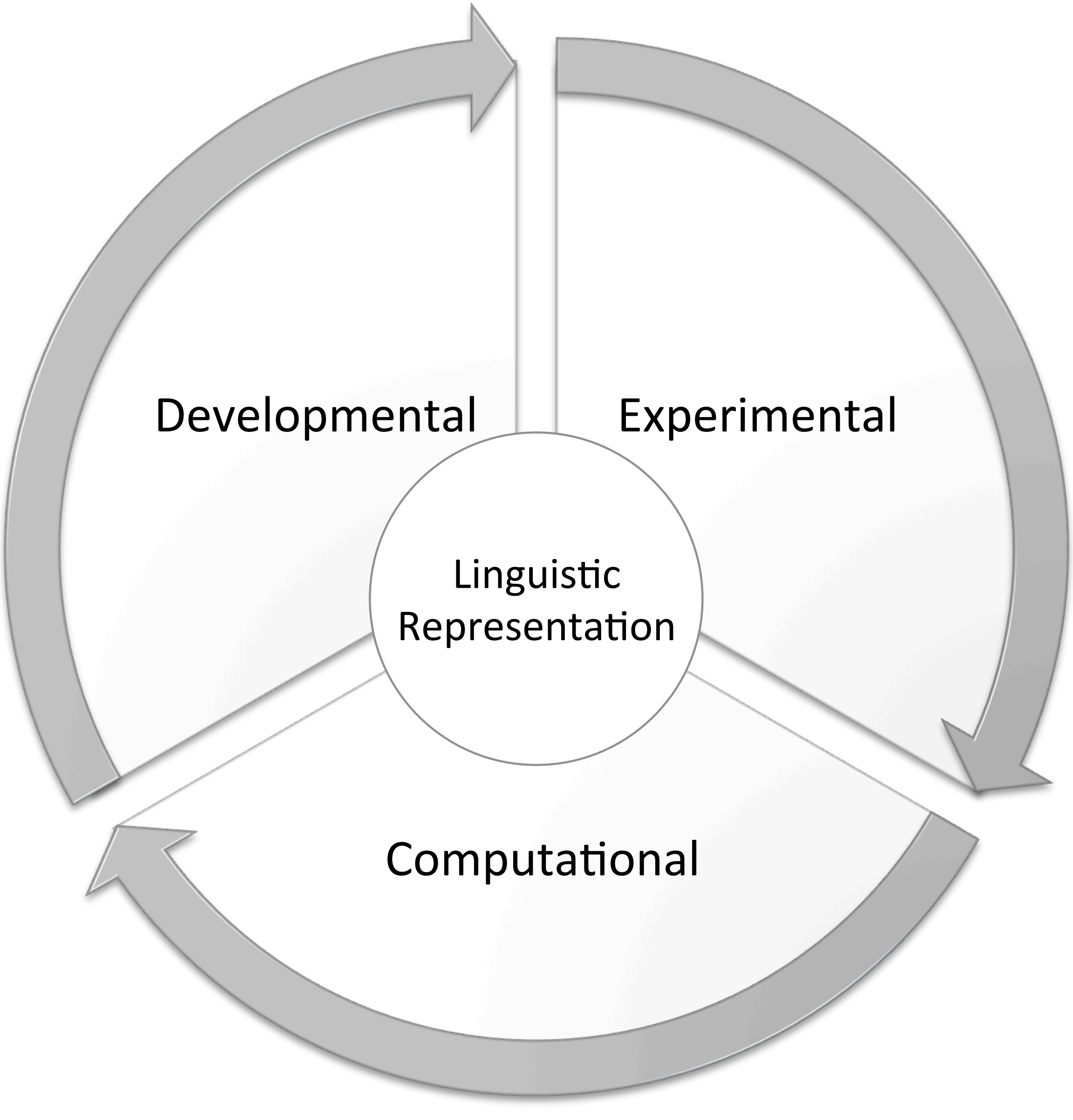
Adam’s mother (Brown, 1973)

Word segmentation is an *important* problem.

- How does an infant divide the input into reusable units?
- How does she represent those units?
- What does she know about them and when?

Not an end in itself: provides useful *units* (Peters, 1983) for learning a grammar: lexicon, morphosyntax, phonology.

Towards understanding language acquisition



Modeling of Word Segmentation

- Transitional probability-based learning a dead-end (Yang, 2004)
  - “Naturalistic” studies exist (Graf Estes et al., 2009; Hay et al., 2011; Pelucchi et al., 2009; Swingley, 2005; and many, many more…)
  - But learning *fails* in statistically natural scenarios (Johnson and Tyler, 2010; Lew-Williams et al., 2011; Lew Williams and Saffran, 2012)
- Models using lexicon, and Bayesian inference (Borschinger and Johnson, 2011; Goldwater et al., 2009; Johnson and Goldwater, 2009; Pearl et al., 2011)
  - Goal of learner is develop high-quality lexicon for segmentation, context can be crucial
  - Integrates progress in unsupervised learning (Teh, 2006)
  - High performance. Developmental impact?

Connecting to development

- Can we connect existing models to development?
  - Primarily models at *computational* level, not *algorithmic* (Marr, 1983)

  "In particular, Bayesian models often do not address how the learner might perform the computations required to achieve the optimal solution to the learning problem […] they simply state that if human behavior accords with the predictions of the model, then humans must be performing some computation (possibly a very heuristic one) that allows them to identify the same optimal solution that the model did." (Pearl and Goldwater, 2011)

- What about the *how*? Progression over time? Connecting with experimental work?

Modeling goal

Build the simplest model that:

- Aligns with infants’ capabilities
- Replicates infants’ behavior in a principled fashion
- Performs reasonably at the task

Broader picture

1. Simple approaches to segmentation can help form a lexicon (cf. “protolexicon” Ngon et al., 2013), a set of candidate words.
2. From there, the infant may begin to draw further generalizations about language.

II. An algorithm for segmentation

How do infants segment speech?

- Attending to multiple cues in the input, most popularly:
  - Bootstrapping from known words (Bortfeld et al., 2005; Dahan and Brent, 1999)
  - More easily identify novel words at beginning and ends of utterances at 8 months (Seidl & Johnson, 2006)
  - Dominant stress pattern of language (Jusczyk et al., 1999)

Infant word learning

- Infants show first clear signs of identifying words at 6 mos. (Bergelson and Swingley, 2012)
- At that age, no:
    - Stress preference (Jusczyk et al., 1993, 1999)
    - Phoneme transition preference/phonotactics (Mattys et al., 1999)
        - May not matter anyway, see Gorman (to appear)
- But:
    - Syllable as primary perceptual unit, sensitivity over time:
        - Birth: number of syllables (Bijeljac-Babic et al., 1993); 2 mos: preference for syllabifiable input (Bertoncini & Mehler, 1981), holistic syllables (Jusczyk & Derrah, 1987)
    - Ability to identify cohesive chunks (Goodsitt et al, 1993)

Overview of the proposed algorithm

- Segmenter has a lexicon of potential words it builds over time
    - Starts empty, words are added based on segmentation of each utterance
    - Each word has a score
- Operates online
    - Processes one utterance at a time
    - Cannot remember previous utterances or how it segmented them, only lexicon
- Operates left-to-right in each utterance to insert word boundaries between syllables

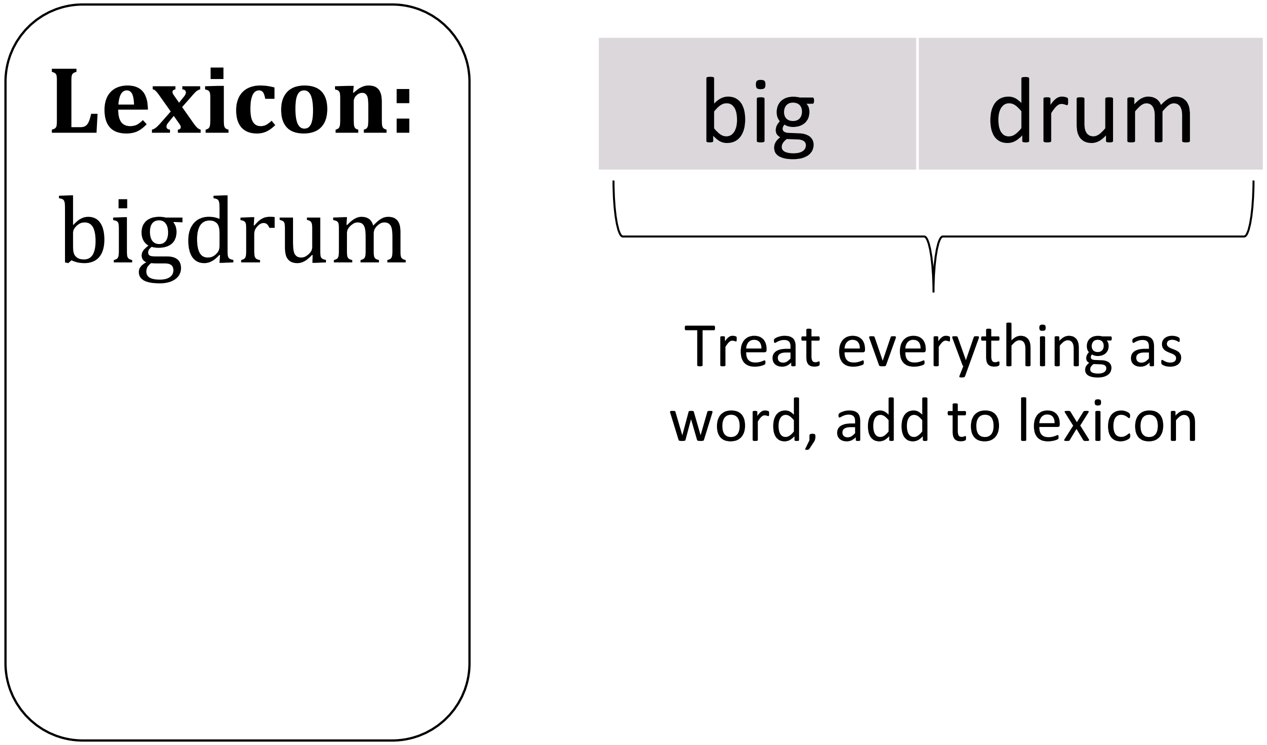
Model in a nutshell

1. Use utterance boundaries to help find initial words.
2. Bootstrap from known words.
3. Reward the words that appear to lead to better segmentations, penalize the ones that lead us astray.

- I'll work through some examples
  - Orthography for easy reading, input is syllabified phonemes

In the beginning…

- Just add whole utterances to the lexicon
- Gets words in isolation for free, but often more than one word



Lexicon:

bigdrum

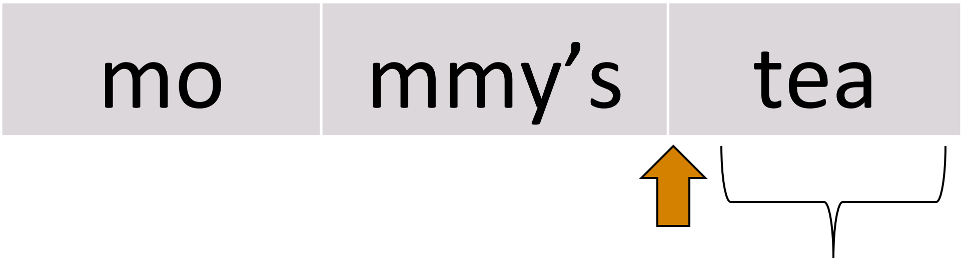
big | drum

Treat everything as word, add to lexicon

Subtractive Segmentation

- Use words in the lexicon to break up the utterance
- Increase word’s score when it is used
- Add new words to lexicon

Lexicon:
mommy’s
tea
…

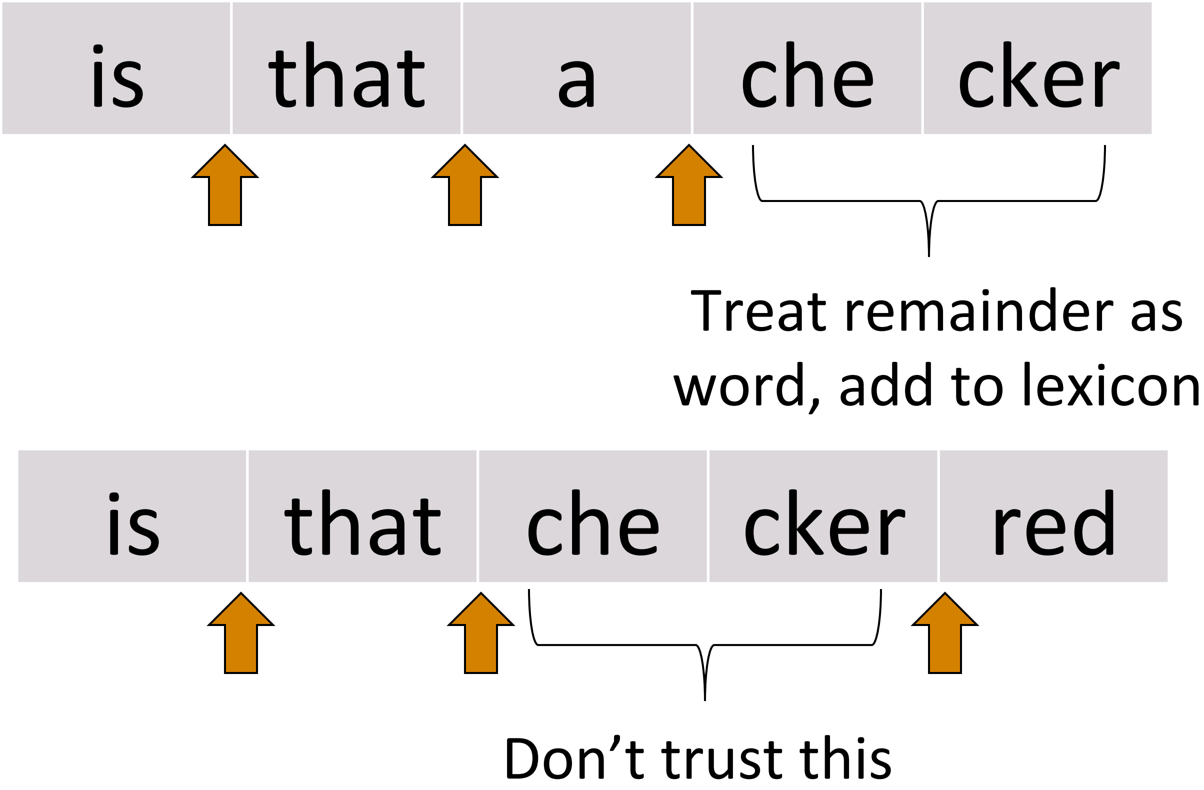


Treat remainder as word, add to lexicon

Trust

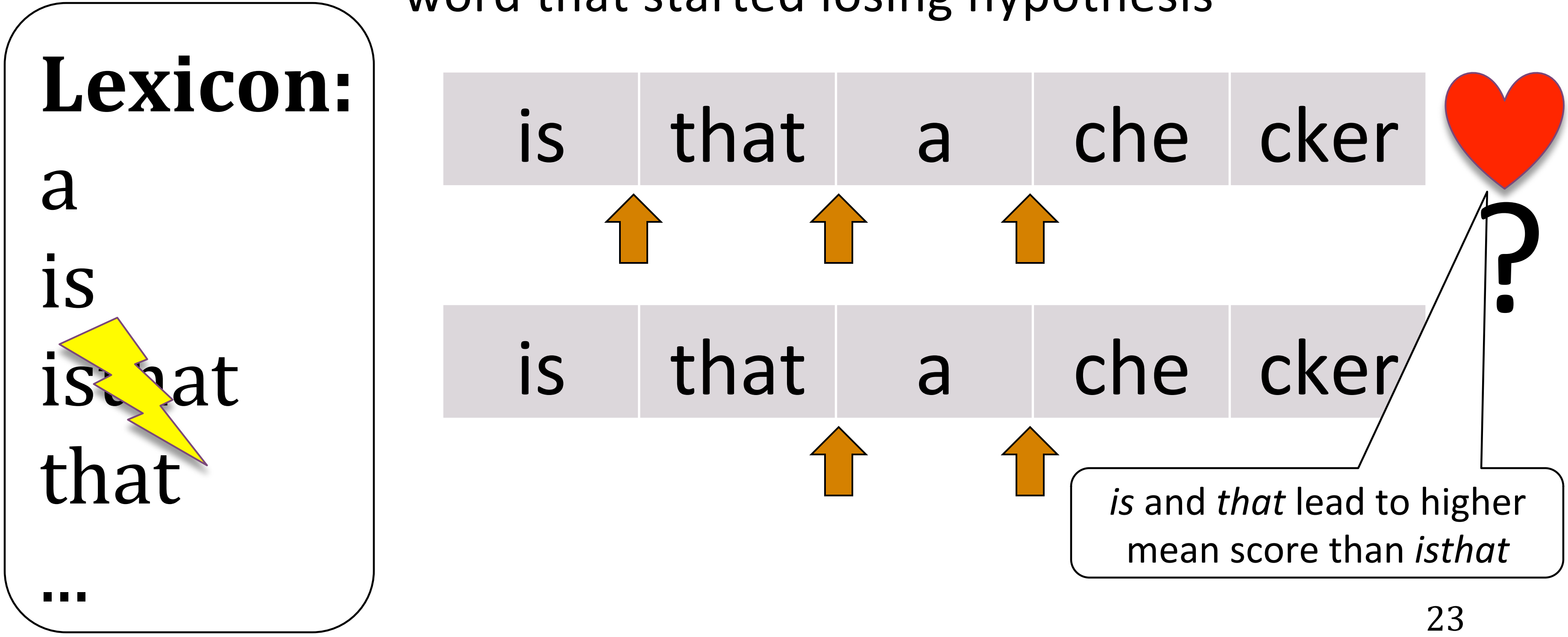
- Add new words to lexicon based on whether we *trust* them (touch an utterance boundary)

Lexicon:
a
is
that
red
checker
...



Multiple hypotheses

- For multiple possible subtractions two options:
  - Greedy approach (Lignos and Yang, 2010)
  - Pursue two hypotheses (*beam search*)
- Two hypotheses allow for *penalization*: reduce score of word that started losing hypothesis



Lexicon:
a
is
isthat
that
...

is and *that* lead to higher mean score than *isthat*

Predictions

- Default assumption of utterance = word → infants will start with oversized units and words in isolation
- Rich-get-richer scoring → As the learner is exposed to more data, learner will tend to use high-frequency elements
- Penalization → Use of collocations will decrease with time

Our evaluation corpus

- Constructed from Brown (1973) subset of CHILDES English (Adam, Eve, Sarah), ~60k utterances
- Pronunciations and stress for each word from CMUDICT, algorithmically syllabified
- Stress modified to better reflect natural speech
  - No adjacent primary stresses (Liberman & Prince, 1977; Selkirk 1984)
- Sample input:

```
B.IH0.G|D.R.AH1.M
HH.AO1.R.S
HH.UW0|IH0.Z|DH.AE1.T
```

Performance

Method	Hit	False Alarm	A’	Word F-Score
Baseline: syllable = word	1.0	1.0	0	0.753
Subtractive Segmentation	**0.992**	0.776	0.795	0.797
+Trust	0.961	0.468	0.860	0.841
+Multiple Hypotheses	0.953	**0.401**	**0.875**	**0.849**

- Performance is well above baseline
- More “infant-like” segmentation yields *better* results

Infant error patterns

- Undersegmentation at young age (Brown, 1973; Clark, 1977; Peters, 1977, 1983)
  - Function word collocations: *that-a*, *it's, isn't-it*
  - Phrases as single unit:

    look at that, what's that, oh boy, all gone, open the door
  - Function-content collocations: *picture-of*, *whole-thing*
- Oversegmentation at older ages (Peters, 1983)
  - Function word oversegmentation: *behave/be have*, *tulips/two lips, Miami/my Ami/your Ami*
  - Errors can still occur in adulthood (Cutler and Norris, 1988)

Most frequent error tokens

(Converted to orthography for easier reading)

Early		
Error	**Example**	**Freq.**
oh	over	209
a	away	184
thats-a	-	101
thank-you	-	45
some	something	39
all	always	31
any	anyone	31
it’s-a	-	30
why-don’t	-	28
don’t-know	-	26
at-the	-	24
put-the	-	24

Late		
Error	**Example**	**Freq.**
a	away	441
oh	over	194
some	something	101
any	anyone	77
all	always	67
every	everyone	60
in	inside	57
on	onto	53
-ty	pretty	41
be	become	40
more	anymore	39
huh	honey	37

III. From lexicon to grammar

We’ve got a segmentation.
What do we do with it?

Stress pattern learning

- English learning infants take primary stress to be a strong cue of word-initial position
  - Even when it leads them astray (Houston et al., 2004; Jusczyk et al., 1999)
- Identification of stress pattern in the language
  - Bias comes from content of lexicon
  - Multisyllabic words in the learner's acquired lexicon have stress-initial rate of 70.3%
  - Taking advantage of this bias in learning reduces errors by 37.0%

What does the bias mean for learning?

- Hypothesis: learner commits to bias as soon as it is reliable, thus higher initial stress rate → faster adoption of bias
- Hungarian (stress-initial by rule): word segmentation errors are rare (MacWhinney, 1976; Peters, 1983)
- English (generally stress-initial, ~80%): children develop reliable bias and apply it readily (*TARis* for *guiTAR is, NAna for baNAna, LIER for cavaLIER*)
- French (arguably no word-level stress): Delayed performance compared to other languages (Nazzi et al., 2006)

Morphological learning

- Other modeling work has identified importance of morphology in segmentation (Berg-Kirkpatrick et al., 2010; Johnson, 2008)
  - No online mechanism for using it
  - Cannot address time course of joint learning
- How can we develop an online mechanism that replicates child learning patterns?

Batch learning (Chan and Lignos, 2010)

- Iteratively select most type-frequent rule

Morpheme	Transform(s)	Brown Mean Rank	Learner Mean Rank
Present progressive (-ing)	(ø, iŋ)	2.33	2.2
Plural (-s, -es)	(ø, z/s/əz)	3	4.0
Possessive (‘s)	(ø, z/s/əz)	6.33	4.0
Third person singular (-s)	(ø, z/s/əz)	[illegible]66	4.0
Contractible copula (-’s)	(ø, z/s/əz)	[illegible]	[illegible]
Past Regular (-ed)	(ø, d/t/əd)	[illegible]	[illegible]
Contractible auxiliary (-’d, -’ll)	(ø, d/əl)	[illegible]	[illegible]

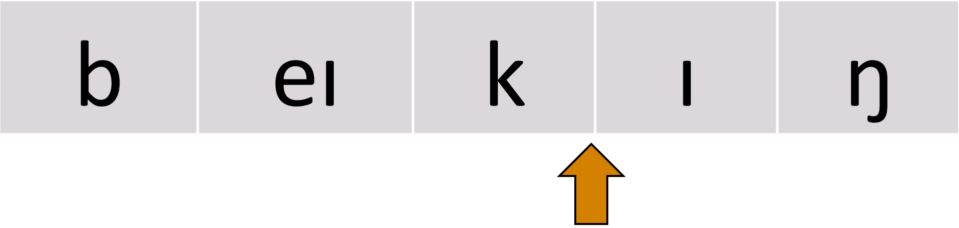
Perfect source for learning alternations, thus phonotactics (Gorman, to appear)

Progress in online learning

- Effects of aggressive decomposition during real-time processing (Lignos and Gorman, in press)
- Assume learner is eager to split the input:
    - At early stages, try all possible split points for a word (*baking)*:

  beɪkɪŋ-, beɪkɪ-ŋ, beɪk-ɪŋ, beɪ-kɪŋ, b-eɪkɪŋ

    - Use the brute force frequency estimates to estimate the frequency of morphemes
- Split words by greedy MDL (minimum description length)



b eɪ k ɪ ŋ

Split point that optimizes frequency of units

Progress in online learning

- Take the top suffixes used in the MDL segmentation, evaluate them as a possible set, and learn the highest *yield* affix
- Yield: if the stems of -*ing* words can be reused for 20 -*z* words, and 10 -*d* words, the yield of -*ing* is 30
    - Poor man's productivity measure that gets around issues of regrettable early decisions

Learned morphemes

- Learned on output of segmentation
- Initial lexicon of high-enough quality to learn common suffixes
- Acquisition order similar to Brown 1973:

Predicted order	Suffix	Example	Brown Average Rank
1	-ɪŋ	bake-ing	2.33
2	-z	happen-s	7.9
3	-d	happen-ed	9
4	-ɚ	bake-er	-
5	-t	check-ed	9
6	-ən	broke-en	-
7	-i:	smell-y	-

Conclusions

- Reward-based models that try to optimize the frequency of units match learning behavior well
  - KISS (Keep It Simple, Silly)
- Early lexicon growth helps address the chicken and egg problem of early acquisition and provides useful generalizations for word structure
- Modeling can help us close the loop between experimental and developmental findings toward a complete picture of language acquisition

Thanks to:
Charles Yang
Mitch Marcus
NSF IGERT #50504487

Constantine Lignos
lignos@cis.upenn.edu
http://www.seas.upenn.edu/~lignos

Code/data:
https://github.com/ConstantineLignos/WordSegmentation