Modeling Infants' Learning of Word Segmentation and Structure

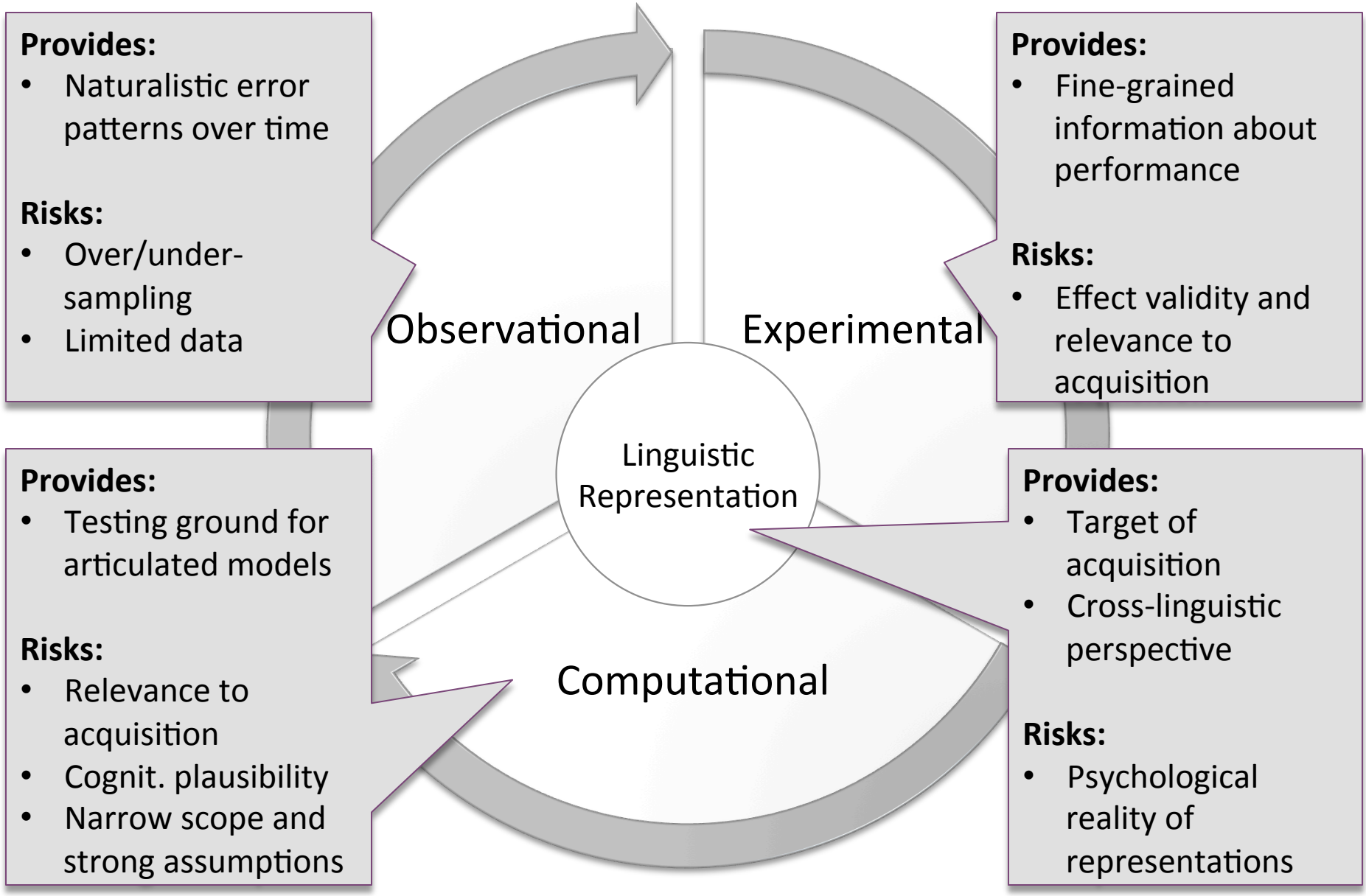
Constantine Lignos

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

University of Edinburgh Institute for Language, Cognition and Computation Seminar

11/5/2012

Towards understanding language acquisition



I. Infant word segmentation

Word segmentation (WS)

- Approximation: utterances of sounds as input, goal is to identify the words that produced the input
- But word segmentation is not an end in itself: provides useful *units* (Peters, 1983) for learning and understanding
  - Lexicon, Morphosyntax, Phonology
- My interest: how can we explain the behavior of children as they learn to segment words?

The finding that launched a thousand studies

*bidaku**padotigolabu**bidaku*
(Saffran et al., 1996)

- Three syllable "words" like *bidaku*
- Transitional probability (TP) manipulation
  - Word-internal: 1.0
  - Between-word: 0.33
- Infants discriminate between "words" and their scrambled form or sequences that cross word boundaries
- Claim: infants use TP minima to perform segmentation

Questions begged

1. How can TPs be used as a part of a model of WS?
2. Does the use of TPs align with other experimental/ developmental findings?

Often replicated, but never validated

- Many replications of findings related to TP-based segmentation, cross-domain and cross-species (Hauser et al., 2001; Saffran et al., 1999, and many, *many*, others)
- But at every possibility, other cues override TPs
  - Stress, when infants reach 9 months (Thiessen and Saffran, 2003)
  - Prosodic boundaries imply word boundaries (Shukla et al., 2011)
- TP learning fails in more naturalistic circumstances (*pace* Graf Estes et. al., 2009; Hay et al., 2011; Pelucchi et al., 2009)
  - Presence of words in isolation (Lew-Williams et. al, 2011)
  - Learning words of varying length (Johnson and Tyler, 2010; Lew Williams and Saffran, 2012)

Modeling of WS

- TP minima by themselves are a dead-end (Yang, 2004)
- 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 these 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?

A better way?

- It turns out simple strategies work well (Gambell and Yang, 2006; Yang, 2004)
- Can we model learning over naturalistic development, e.g., past tense debate?

Our modeling goal:

Build the simplest model that:

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

II. An algorithm for segmentation

How do infants segment speech?

- Possible strategy: identification of words in isolation (Peters, 1983; Pinker et al., 1984)
  - Unlikely to be sufficient (Aslin et al., 1996), but probably helpful (Brent and Siskind, 2001)
- 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)
- 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)

Modeling assumptions

- In modeling, assumptions needed to help isolate phenomena at a particular level
  - With goal to relax assumptions as more is known about solution
- Learner is given canonical syllabified input
  - As standard in experimental work and corpus analysis
  - Convergence toward the syllable as the primary unit of speech perception of young infants (Mehler et al., 1990)
- Able to map acoustic signal to strong/weak stress on syllables (Johnson & Jusczyk, 2001)

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…

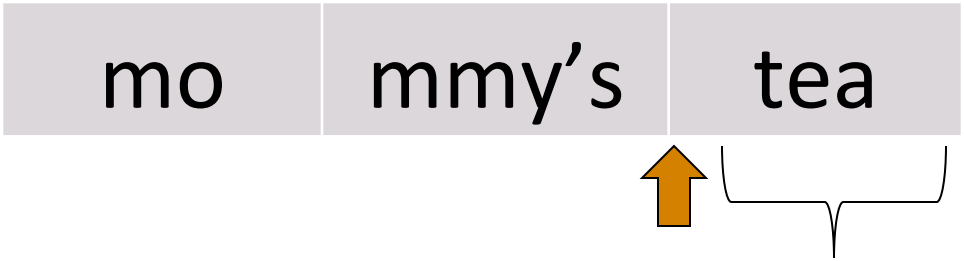
- Just add whole utterances to the lexicon
- Gets words in isolation for free, but often more than one word

Lexicon:
bigdrum

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
…

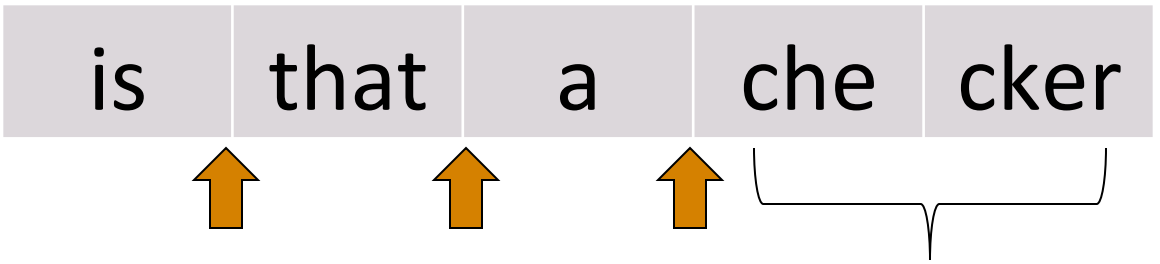


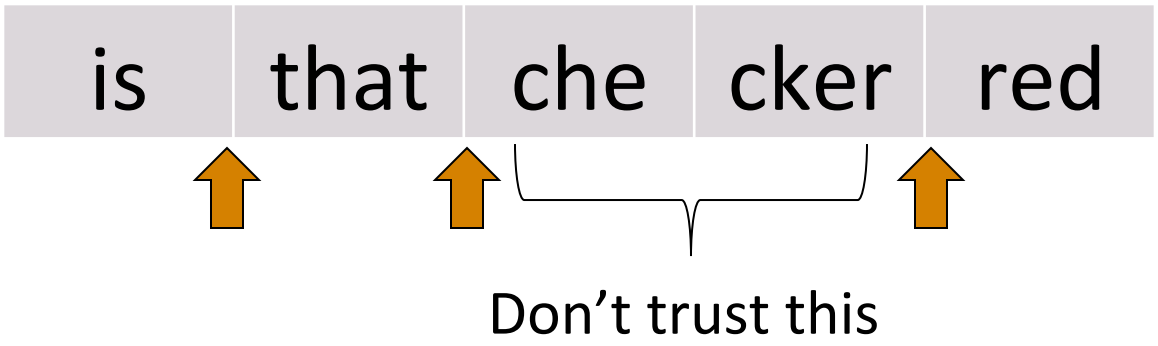
Treat remainder as word, add to lexicon

Trust

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

Lexicon:
a
is
that
red
checker
…

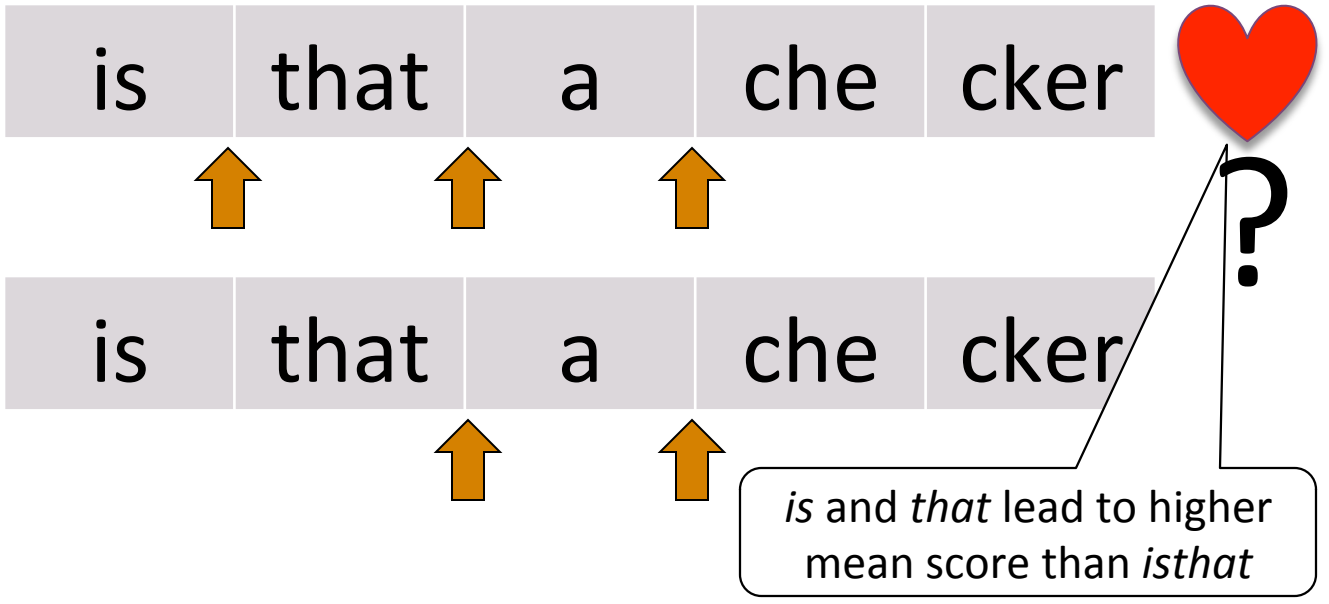




Multiple hypotheses

- 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
…



Scoring hypotheses

- Prefer the hypothesis that uses the higher-scoring words
  - Winner is rewarded, word scores will go up: “rich get richer”
- Geometric mean of scores of words used as metric:

$$\arg\max_{H}\left(\prod_{w_i \in H} score(w_i)\right)^{\frac{1}{n}}$$

  - Useful for compound splitting (Koehn and Knight, 2003; Lignos, 2010)
  - Doesn’t have bias for fewer or more words, but seeks a higher average score
  - Post-hoc testing shows this outperforms other obvious metrics (MDL/MLE, other means)

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

III. Results and analysis

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

Evaluation

- A' calculated over syllable boundaries
  - Balance of hit rate and false alarm rate for discriminating word boundaries
  - Trapezoidal approximation of area under ROC curve
- F-score used for word token identification
  - Balance of precision (how often a word identified in an utterance is correct) and recall (how many correct words were found)
  - Ex: *is that a lady* segmented as *isthat a lady*
    - Precision 2/3: *a, lady*; *isthat*
    - Recall 2/4: *a, lady; is, that*
- Compared against baseline strategies

Evaluation

- Evaluated segmenter in three forms:
  - Subtractive segmentation
  - Subtractive segmentation with *trust*, only adding words to the lexicon if they touch an utterance boundary
  - Subtractive segmentation with trust and *multiple hypotheses*, considering two hypothetical segmentations and penalizing the loser
- Errors computed over learning corpus from first utterance
  - Worst-case evaluation for online learner
- Word boundaries taken as orthographic boundaries in the input
  - Aligns with morphological and phonological definitions of word

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**

- Trust and multiple hypotheses significantly reduce FA rate
- Perfect memory is not crucial: evaluating A’ and F-score with imperfect memory/syllable identification yield similar 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

- Divided into early (first 10k utterances) and late (last 10k utterances) stages of learning
- Coded most frequent incorrect words in output as:
  - Function: Overuse of function word (away→ a way)
  - Function collocation: Two function words (that's a → that'sa)
  - Content collocation: Content and content/function word (a ball → aball)
  - Other
- Distribution changes across time (Chi-squared $p < .0001$)

Time	Function	Func. Colloc.	Cont. Colloc.	Other
Early	44.2%	37.0%	8.5%	10.4%
Late	70.6%	1.0%	1.2%	27.3%

Most frequent error tokens

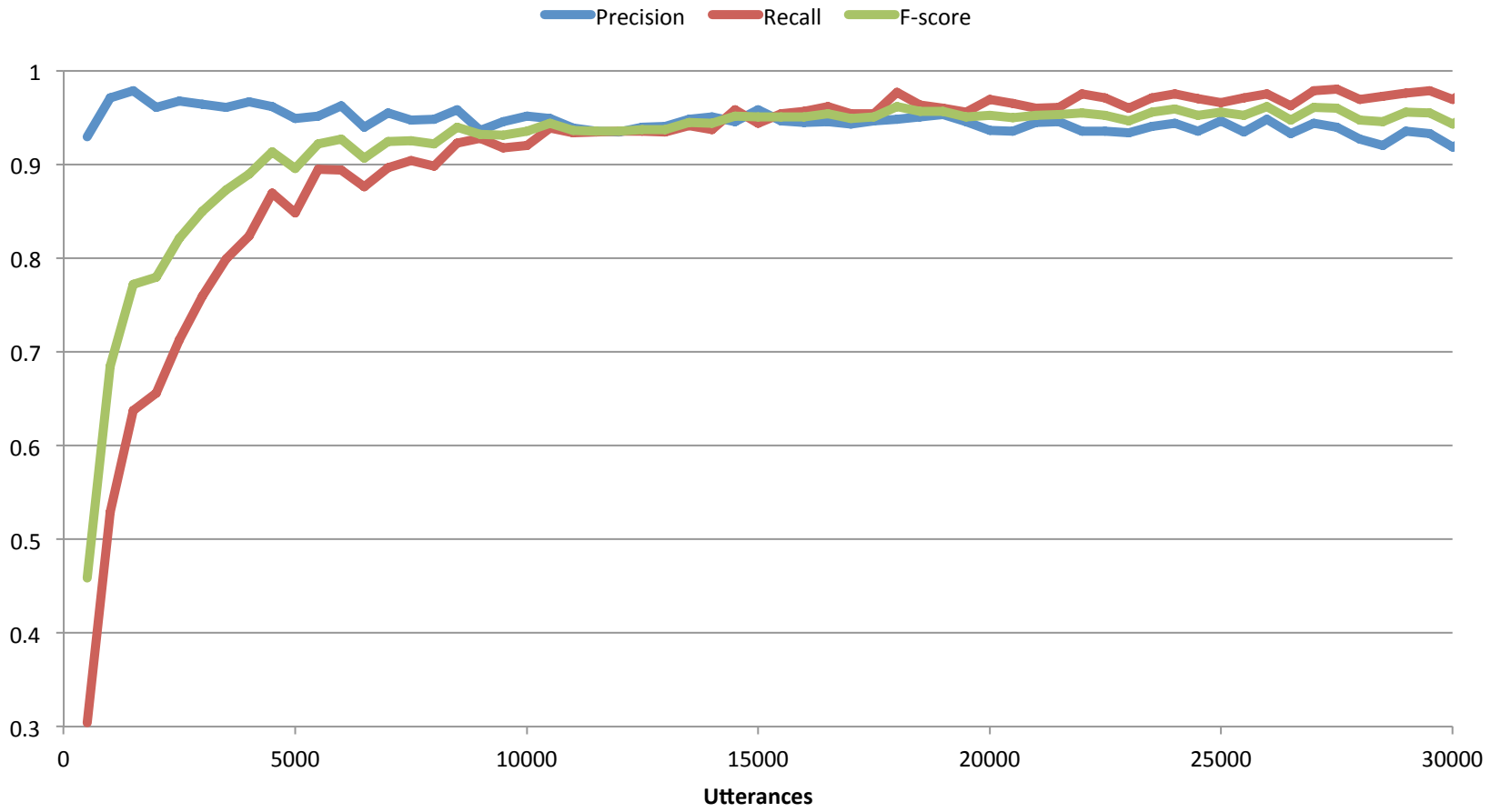
(Converted to orthography for easier reading)

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

Learning curve

- Learner starts undersegmenting, as it learns achieves balance with slight oversegmentation



Stress pattern learning

- Identification of stress pattern in the language
  - 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%
- How does this bias affect the learner?
  - 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*)
  - French (arguably no word-level stress): Delayed performance compared to other languages (Nazzi et al., 2006)

Moving forward

- What are the units to focus on?
  - What's the correct segmentation for:
    - Spanish verbs with clitics: *Dámelo (give me it)*
  - If morpheme discovery is a part of the problem, gold standard question becomes less relevant.
- Feedback with other levels
  - There are joint learning models, but they have serious limitations

IV. Preliminary extensions

Morphological learning

- Other modeling work has identified importance of morphology in segmentation (Berg-Kirkpatrick et al., 2010; Johnson, 2008)
  - But no online mechanism for using it
- 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)	9.66	4.0
Contractible copula (-’s)	(ø, z/s/əz)	12.66	4.0
Past Regular (-ed)	(ø, d/t/əd)	9	6.2
Contractible auxiliary (-’d, -’ll)	(ø, d/əl)	14	8.5

A first attempt at an online learner

- 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-, bakin-g, baki-ng, bak-ing, ba-king, b-aking

  - Use the brute force frequency estimates to estimate the frequency of morphemes
  - Split words by greedy MDL (minimum description length)
- After initial splits, to gauge which morphemes make good stems
  - Take the top *N* apparent affixes, use them to segment words, mark the highest yield affix as good

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	-

Improving evaluation of word segmentation

- Development of comprehensive signal detection metrics for (online!) segmenter performance
  - Boundary discrimination and bias, word-level metrics
  - Some accommodation for clitics, morphological complexity
  - Framework for running all segmenters of note
- Application of mixed-effects modeling and residualization to get a fined-grained view of segmenter performance
  - How much of learning success is edge-effects, words in isolation, frequency, etc.
- Comparison of word learning order to age of acquisition data
  - CDI norms, etc.

Conclusions

- Reward-based model can lead to the changes in unit size seen in children and later promote language-specific segmentation strategies
- 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