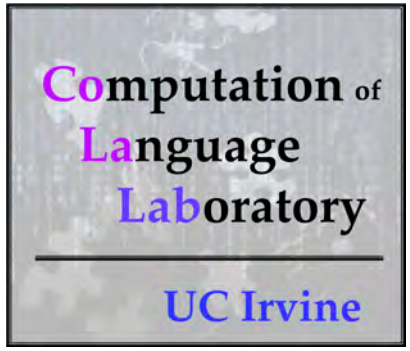
Understanding the "Less is More" Effect in Language Development: A Look at Word Segmentation

Galia Barsever

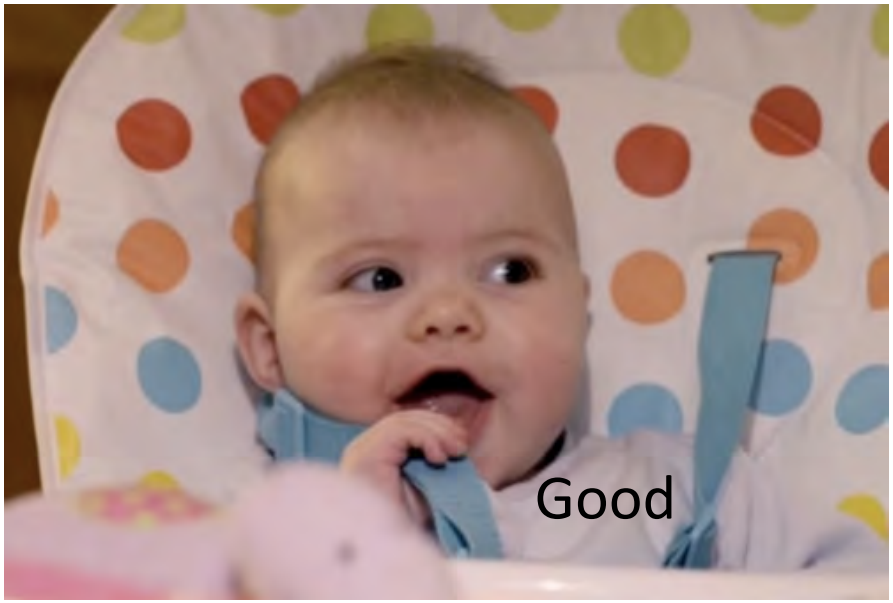
Department of Cognitive Sciences

University of California, Irvine



Language development

- Acquiring language is a hard task.
- Young children are good at it.
- Adults are usually bad at it.



Word Segmentation

- One of the first problems an infant must solve when learning language.

Word Segmentation

- Given a corpus of fluent speech or text (no utterance-internal word boundaries), we want to identify the words

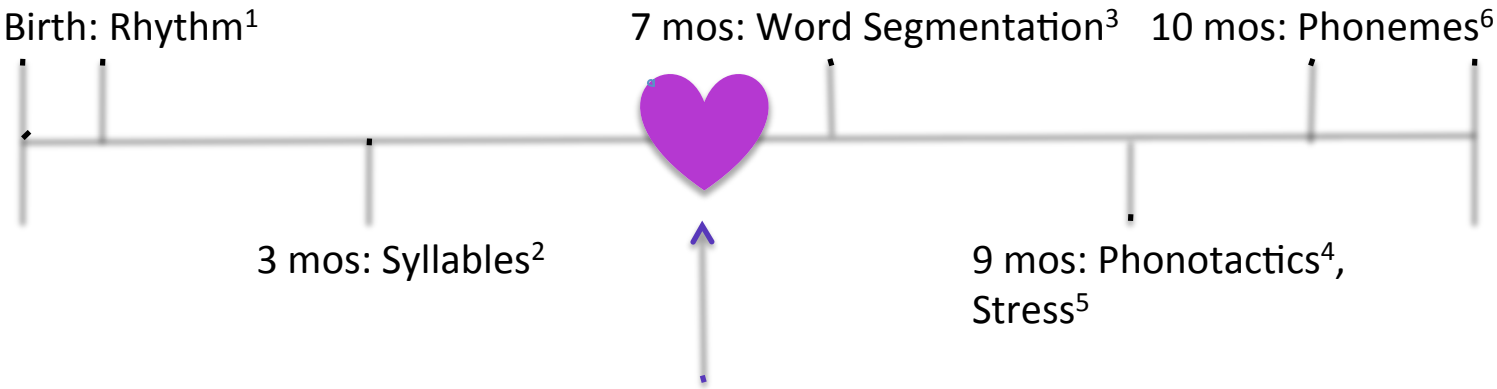
whatsthat
thedoggie
yeah
wheresthedoggie

whats that
the doggie
yeah
wheres the doggie

Word Segmentation

- When does word segmentation happen?
- Very early according to time course
  - Interesting because children are not very cognitively developed early on.

Infant Language Tasks



Our data comes from child-directed speech to infants about this age

[1]Nazzi et al. 1998
[2]Eimas 1999
[3]Jusczyk et al. 1999
[4]Echols et al. 1997
[5]Jusczyk et al. 1993
[6] Werker & Tees 1984

Why Use Computational Models?

- In digital children…
  - We have control of the input representation
  - We have control of what strategy they're using to update their beliefs

…among other things

Why Use Computational Models?

- Can't control real kids
  - For example: the word "flower"

Two ways to represent input:

- In syllables: flow-er → two separate parts
- In phonemes: /f l a ʊ ə r/ → six separate parts

  - Can't dictate to children which way to see things

The "Less is More" hypothesis

- Less cognitive resources are paradoxically better for language learning
  - The very idea of "Less is More" is counter-intuitive.
  - And yet, babies are better than adults at learning language

- How could less information processing capabilities yield better results?

The "Less is More" hypothesis

- What support do we have?
  - It works for adults (sometimes) for second language learning (Cochran et al. 1999, Kersten et al. 2001 but see Perfors 2011).
  - Less input led to better learning!

Word Segmentation: Potential Strategies

- Infants make use of many different cues
  - Many require you to already know words
  - Example: stress patterns (EMphasis vs. emPHAsis)

- Statistical information may provide initial bootstrapping
  - Used very early (Thiessen and Saffran, 2003)
  - Language independent, so it doesn't require children to know words already.

Ideal vs. Constrained

- We have two basic types of models we work with:
  - Ideal learners have perfect memory and process information in a batch all once. They have enough processing resources to search all potential segmentations and select the optimal segmentation.
  - Constrained learners do not process in a batch, and select segmentation probabilistically and may not pick best one all of the time.

Making Model More Like Kids

- Constrained versus Ideal learners:
  - Constrained learners have limited processing and decaying memory.
  - Can only remember the decisions made about segmentation very recently, not enough processing power to remember every decision.
    - Ideal learners have amazing memory abilities, but the decayed learners are more like kids.

Two Potential Statistical Strategies

- »Bayesian inference
- »PHOCUS

Bayesian strategy	PHOCUS strategy
Process incrementally or Batch	Process incrementally
Transitional probability: Over units that make up words and Over sequences of words	Transitional probability: Over units that make up words
Constrained and Ideal	Constrained
Build lexicon	Build lexicon

Lexicons

- Both our strategies use lexicons
  - Meaning, it takes words it thinks it knows and puts it in a vocabulary to compare against future possible segmentations

Previous work on Bayesian segmentation (using phonemes as the basic unit)

- Goldwater, Griffiths, & Johnson (2009) [Bayesian learning – yay! It works great!] ← Ideal learners
- Pearl, Goldwater, Steyvers (2011) ← Constrained learners
  - Found "Less is More" effect, but not really strong

Bayesian Segmentation

$$P(h|d) \propto P(d|h) * P(h)$$

Posterior Probability

Likelihood

Prior

A formal way of representing belief update

Takes a prior belief and multiplies that by the likelihood of that belief.

What is Bayesian Segmentation?

- Data: unsegmented corpus (transcriptions)
- Hypotheses: sequences of word tokens
  - Example: Data to be segmented: **Ilikekitties**
    - » Hypothesis 1: Ilike kitties
    - » Hypothesis 2: I like ki tties
    - » ~~Hypothesis 3: I like doggies~~
- Optimal solution is segmentation with highest posterior probability.

Making Model More Like Kids

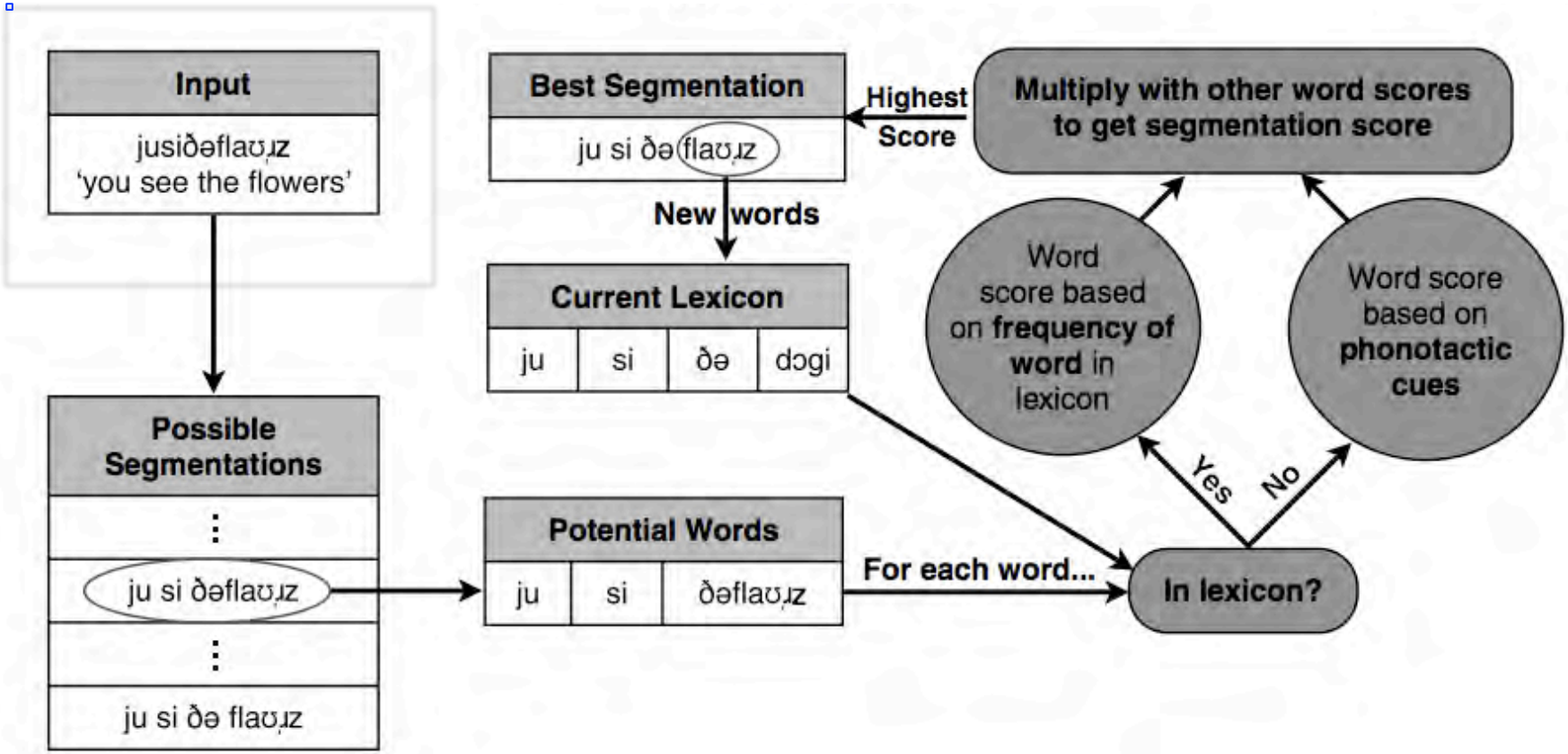
- Even if we have a Bayesian strategy, there are a lot of options when implementing it
  - Syllables versus phonemes as unit of representation in input
  - Dependent or independent on units next to them
    - » We are talking about syllable-based and dependent (both more likely strategies)

Another Strategy

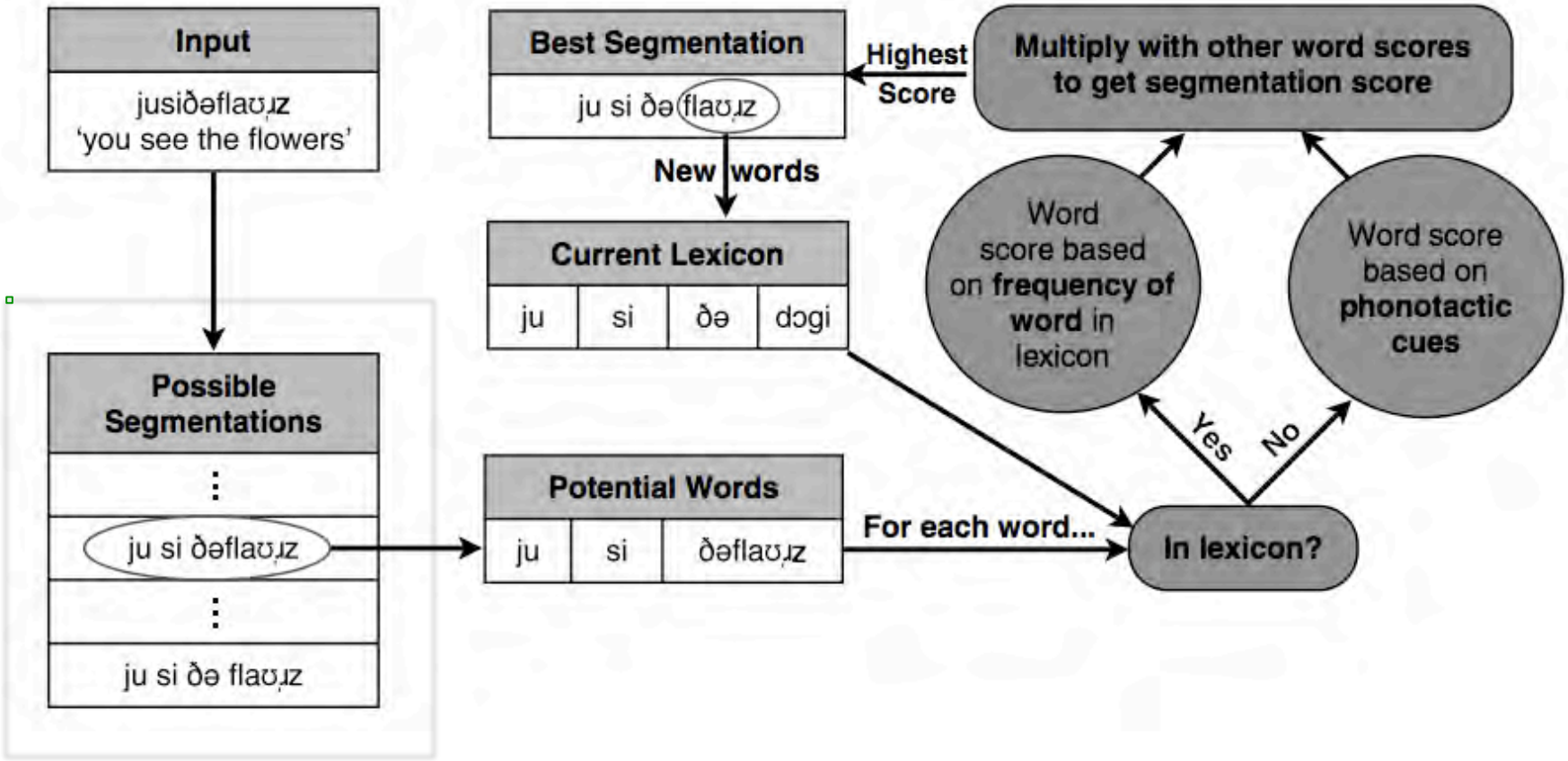
- The one used by PHOCUS (Phonotactic Cue Segmenter).
  - Initially used to run over phonemes, now running over syllables
  - Also a constrained statistical learner, but doesn't use Bayesian inference.

PHOCUS

For each utterance encountered…

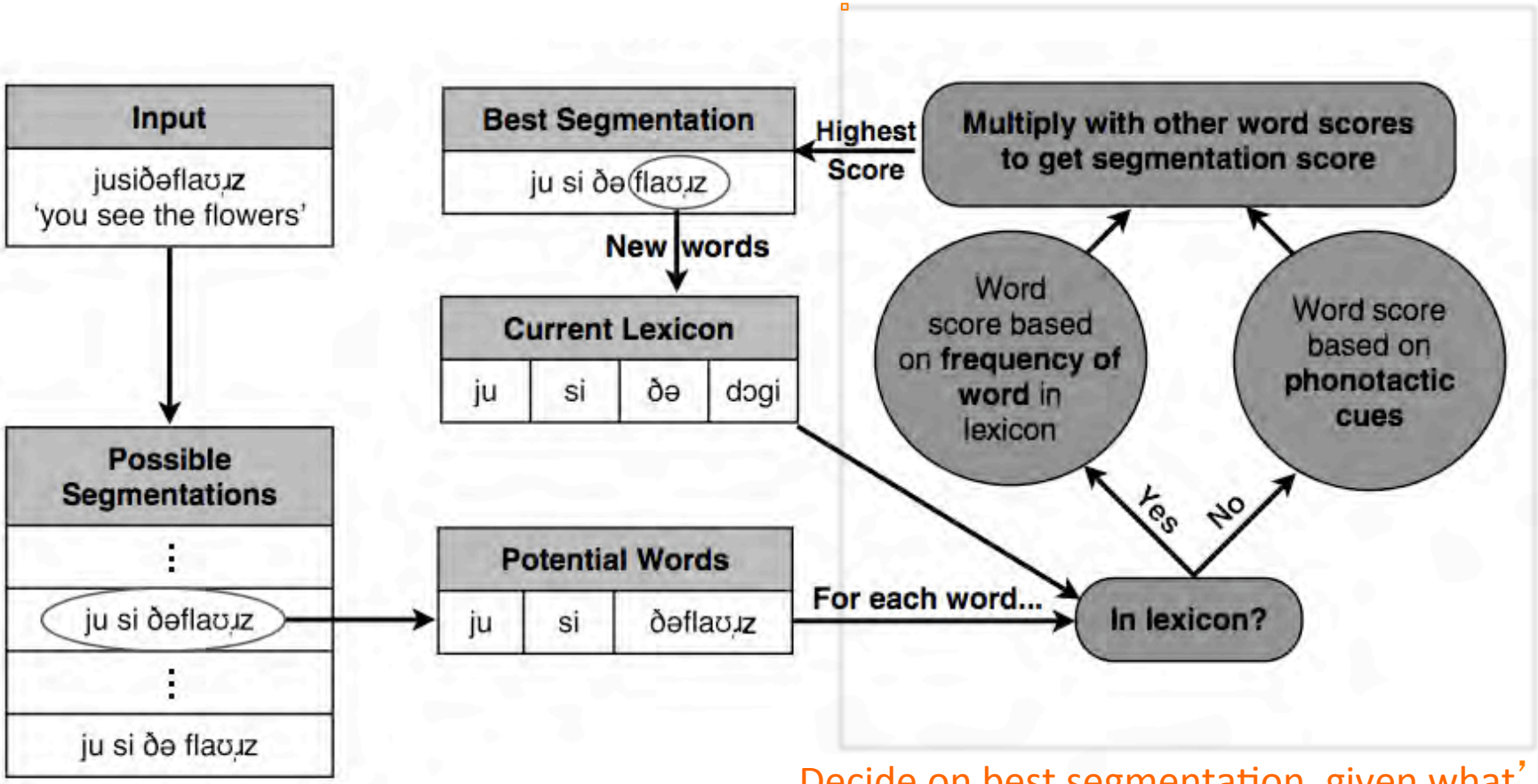


PHOCUS



Look at possible segmentations…

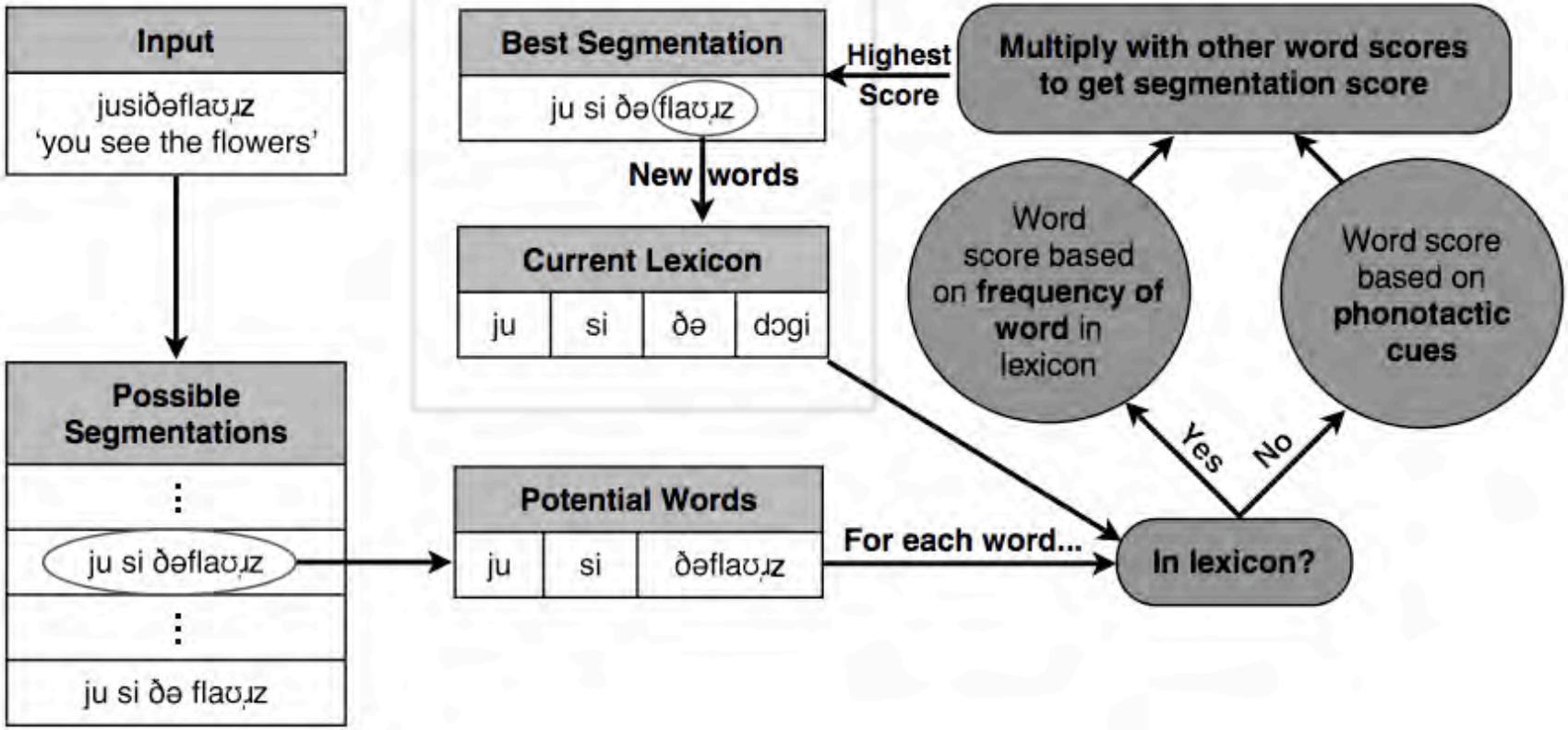
PHOCUS



Decide on best segmentation, given what's already in the lexicon (both the words and frequency of those words)…

PHOCUS

Once you have the best segmentation, update the lexicon with those new words



Comparing Strategies

- What happens if PHOCUS shows stronger or as good performance as Bayesian?
  - Using constrained learners is what adds the "less is more" boost.

Comparing Strategies

- What happens if PHOCUS shows weaker performance than Bayesian?
  - The Bayesian segmentation strategy is what adds to the "less is more" effect.

Results for English Data

F-score

Ideal Best Constrained PHOCUS

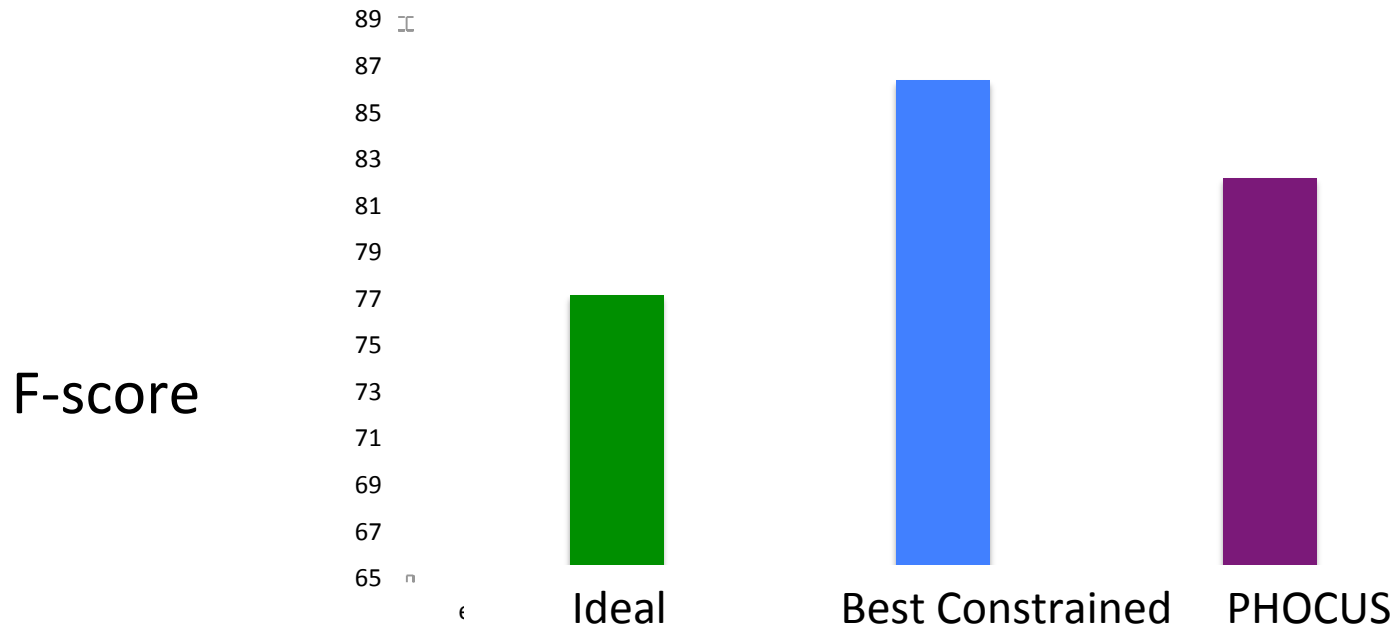
Looking at Precision and Recall:
- Precision: #words correct/#words guessed
- Recall: #words correct/#true words

Number that combines both: F-score is the harmonic mean of both

$$F = 2 * \frac{precision * recall}{precision + recall}$$

Results for English Data

F-score over Ideal-Bayesian, Constrained-Bayesian, and PHOCUS over syllables



What did we find?

The PHOCUS model does pretty much the same as Bayesian constrained (though not quite as good)

What does this mean?

The "less is more" effect is (probably) mostly a result of a constrained learner.

Open Questions

- We see this effect, but how does this work over different languages?
  - What kind of cross-linguistic differences do we get when we look at syllables vs. phonemes as unit of representation? Is PHOCUS still almost as good?

- Is it important to have a strategy that is building a lexicon as long as you have constrained memory?

Thanks

- Thank you to Dr. Lisa Pearl and Lawrence Phillips!