

Psych 215: Language Sciences (Language Acquisition)

Lecture 7 Word Segmentation

Computational Problem

Divide spoken speech into words

húwzəfréjdəvðəbíg bæ'dwə'lf

Computational Problem

Divide spoken speech into words

húwzəfréjdəvðəbíg bæ'dwə'lf

↓
húwz əfréjd əv ðə bíg bæ'd wə'lf
who's afraid of the big bad wolf



Word Segmentation

“One task faced by all language learners is the segmentation of fluent speech into words. This process is particularly difficult because word boundaries in fluent speech are marked inconsistently by discrete acoustic events such as pauses...it is not clear what information is used by infants to discover word boundaries...there is no invariant cue to word boundaries present in all languages.”

- Saffran, Aslin, & Newport (1996)

Statistical Information Available

Maybe infants are sensitive to the statistical patterns contained in sequences of sounds.

“Over a corpus of speech there are measurable statistical regularities that distinguish recurring sound sequences that comprise words from the more accidental sound sequences that occur across word boundaries.” - Saffran, Aslin, & Newport (1996)

who's afraid of the big bad wolf

Statistical Information Available

Maybe infants are sensitive to the statistical patterns contained in sequences of sounds.

“Over a corpus of speech there are measurable statistical regularities that distinguish recurring sound sequences that comprise words from the more accidental sound sequences that occur across word boundaries.” - Saffran, Aslin, & Newport (1996)

Statistical regularity: *a + afraid* is a common sound sequence

who's **afraid** of the big bad wolf

Statistical Information Available

Maybe infants are sensitive to the statistical patterns contained in sequences of sounds.

“Over a corpus of speech there are measurable statistical regularities that distinguish recurring sound sequences that comprise words from the more accidental sound sequences that occur across word boundaries.” - Saffran, Aslin, & Newport (1996)

No regularity: *afraid + of* is an accidental sound sequence

who's afraid **of** the big bad wolf

word boundary

Transitional Probability

“Within a language, the transitional probability from one sound to the next will generally be highest when the two sounds follow one another in a word, whereas transitional probabilities spanning a word boundary will be relatively low.” - Saffran, Aslin, & Newport (1996)

Transitional Probability = Conditional Probability

$$\text{TrProb}(AB) = \text{Prob}(B | A)$$

Transitional probability of sequence AB is the conditional probability of B, given that A has been encountered.

$$\text{TrProb}(\text{"gob" "lin"}) = \text{Prob}(\text{"lin"} | \text{"gob"})$$

Transitional Probability

“Within a language, the transitional probability from one sound to the next will generally be highest when the two sounds follow one another in a word, whereas transitional probabilities spanning a word boundary will be relatively low.”
- Saffran, Aslin, & Newport (1996)

Transitional Probability = Conditional Probability

$\text{TrProb}(\text{"gob" | "lin"}) = \text{Prob}(\text{"lin" | "gob"})$

gob... ...ble, ...bler, ...bledygook, ...let, ...lin, ...stopper

(6 options)

$\text{Prob}(\text{"lin" | "gob"}) = 1/6$

Transitional Probability

“Within a language, the transitional probability from one sound to the next will generally be highest when the two sounds follow one another in a word, whereas transitional probabilities spanning a word boundary will be relatively low.”
- Saffran, Aslin, & Newport (1996)

$\text{Prob}(\text{"fraid" | "a"}) = \text{high}$

who's afraid of the big bad wolf

Transitional Probability

“Within a language, the transitional probability from one sound to the next will generally be highest when the two sounds follow one another in a word, whereas transitional probabilities spanning a word boundary will be relatively low.”
- Saffran, Aslin, & Newport (1996)

$\text{Prob}(\text{"of" | "fraid"}) = \text{lower}$

who's afraid of the big bad wolf

word boundary

Transitional Probability

“Within a language, the transitional probability from one sound to the next will generally be highest when the two sounds follow one another in a word, whereas transitional probabilities spanning a word boundary will be relatively low.”
- Saffran, Aslin, & Newport (1996)

$\text{Prob}(\text{"the" | "of"}) = \text{lower, but not as low as } \text{Prob}(\text{"of" | "afraid"})$

who's afraid of the big bad wolf

word boundary

Transitional Probability

“Within a language, the transitional probability from one sound to the next will generally be highest when the two sounds follow one another in a word, whereas transitional probabilities spanning a word boundary will be relatively low.”
- Saffran, Aslin, & Newport (1996)

$\text{Prob}(\text{"of"} \mid \text{"fraid"}) < \text{Prob}(\text{"fraid"} \mid \text{"a"})$
 $\text{Prob}(\text{"of"} \mid \text{"fraid"}) < \text{Prob}(\text{"the"} \mid \text{"of"})$

who's afraid of the big bad wolf

TrProb learner posits word boundary here,
at the minimum of the TrProbs

8-month old statistical learning

Saffran, Aslin, & Newport 1996

Familiarization-Preference Procedure (Jusczyk & Aslin 1995)

Habituation:

Infants exposed to auditory material that serves as potential learning experience

Test stimuli (tested immediately after familiarization):

(familiar) Items contained within auditory material

(novel) Items not contained within auditory material, but which are nonetheless highly similar to that material

8-month old statistical learning

Saffran, Aslin, & Newport 1996

Familiarization-Preference Procedure (Jusczyk & Aslin 1995)

Measure of infants' response:

Infants control duration of each test trial by their sustained visual fixation on a blinking light.

Idea: If infants have extracted information (based on transitional probabilities), then they will have different looking times for the different test stimuli.

Artificial Language

Saffran, Aslin, & Newport 1996

4 made-up words with 3 syllables each

Condition A:

tupiro, golabu, bidaku, padoti

Condition B:

dapiku, tilado, burobi, pagotu

Artificial Language

Saffran, Aslin, & Newport 1996

Infants were familiarized with a sequence of these words generated by speech synthesizer for 2 minutes. Speaker's voice was female and intonation was monotone. There were no acoustic indicators of word boundaries.

Sample speech:

tu pi ro go la bu bi da ku pa do ti go la bu tu pi ro pa do ti...

Artificial Language

Saffran, Aslin, & Newport 1996

The only cues to word boundaries were the transitional probabilities between syllables.

Within words, transitional probability of syllables = 1.0

Across word boundaries, transitional probability of syllables = 0.33

tu pi ro go la bu bi da ku pa do ti go la bu tu pi ro pa do ti...

Artificial Language

Saffran, Aslin, & Newport 1996

The only cues to word boundaries were the transitional probabilities between syllables.

Within words, transitional probability of syllables = 1.0

Across word boundaries, transitional probability of syllables = 0.33

TrProb("tu" "pi") = 1.0

tu pi ro go la bu bi da ku pa do ti go la bu tu pi ro pa do ti...

Artificial Language

Saffran, Aslin, & Newport 1996

The only cues to word boundaries were the transitional probabilities between syllables.

Within words, transitional probability of syllables = 1.0

Across word boundaries, transitional probability of syllables = 0.33

TrProb("tu" "pi") = 1.0

tu pi ro go la bu bi da ku pa do ti go la bu tu pi ro pa do ti...

Artificial Language

Saffran, Aslin, & Newport 1996

The only cues to word boundaries were the transitional probabilities between syllables.

Within words, transitional probability of syllables = 1.0

Across word boundaries, transitional probability of syllables = 0.33

TrProb("ro" "go") < 1.0 (0.3333...)

tu pi ro go la bu bi da ku pa do ti go la bu tu pi ro pa do ti...

Artificial Language

Saffran, Aslin, & Newport 1996

The only cues to word boundaries were the transitional probabilities between syllables.

Within words, transitional probability of syllables = 1.0

Across word boundaries, transitional probability of syllables = 0.33

TrProb("ro" "go") < 1.0 (0.3333...)

tu pi ro go la bu bi da ku pa do ti go la bu tu pi ro pa do ti...

word boundary

word boundary

Testing Infant Sensitivity

Saffran, Aslin, & Newport 1996

Expt 1, test trial:

Each infant presented with repetitions of 1 of 4 words

2 were "real" words

(ex: *tupiro*, *golabu*)

2 were "fake" words whose syllables were jumbled up

(ex: *ropitu*, *bulago*)

tu pi ro go la bu bi da ku pa do ti go la bu tu pi ro pa do ti...

Testing Infant Sensitivity

Saffran, Aslin, & Newport 1996

Expt 1, test trial:

Each infant presented with repetitions of 1 of 4 words

2 were "real" words

(ex: *tupiro*, *golabu*)

2 were "fake" words whose syllables were jumbled up

(ex: *ropitu*, *bulago*)

tu pi ro go la bu bi da ku pa do ti go la bu tu pi ro pa do ti...

Testing Infant Sensitivity

Saffran, Aslin, & Newport 1996

Expt 1, results:

Infants listened longer to novel items

(7.97 seconds for real words, 8.85 seconds for non-words)

Implication: Infants noticed the difference between real words and non-words from the artificial language after only 2 minutes of listening time!

Testing Infant Sensitivity

Saffran, Aslin, & Newport 1996

Expt 1, results:

Infants listened longer to novel items

(7.97 seconds for real words, 8.85 seconds for non-words)

Implication: Infants noticed the difference between real words and non-words from the artificial language after only 2 minutes of listening time!

But why?

Could be that they just noticed a familiar sequence of sounds, and didn't notice the different transitional probabilities.

Testing Infant Sensitivity

Saffran, Aslin, & Newport 1996

Expt 2, test trial:

Each infant presented with repetitions of 1 of 4 words

2 were "real" words

(ex: *tupiro*, *golabu*)

2 were "part" words whose syllables came from two different words in order

(ex: *pirogo*, *bubida*)

tu pi ro go la bu bi da ku pa do ti go la bu tu pi ro pa do ti...

Testing Infant Sensitivity

Saffran, Aslin, & Newport 1996

Expt 2, test trial:

Each infant presented with repetitions of 1 of 4 words

2 were "real" words

(ex: *tupiro*, *golabu*)

2 were "part" words whose syllables came from two different words in order

(ex: *pirogo*, *bubida*)

tu pi ro go la bu bi da ku pa do ti go la bu tu pi ro pa do ti...

Testing Infant Sensitivity

Saffran, Aslin, & Newport 1996

Expt 2, test trial:

Each infant presented with repetitions of 1 of 4 words

2 were "real" words

(ex: *tupiro*, *golabu*)

2 were "part" words whose syllables came from two different words in order

(ex: *pirogo*, *bubida*)

tu pi ro go la bu bi da ku pa do ti go la bu tu pi ro pa do ti...

Testing Infant Sensitivity

Saffran, Aslin, & Newport 1996

Expt 2, results:

Infants listened longer to novel items

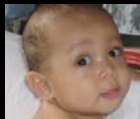
(6.77 seconds for real words, 7.60 seconds for part-words)

Implication: Infants noticed the difference between real words and part-words from the artificial language after only 2 minutes of listening time! They are sensitive to the transitional probability information.

Saffran, Aslin, & Newport (1996)

Experimental evidence suggests that 8 month old infants can track statistical information such as the transitional probability between syllables. This can help them solve the task of word segmentation.

Evidence comes from testing children in an artificial language paradigm, with very short exposure time.



Computational Modeling Data (Digital Children)



How good is transitional probability on real data?

Gambell & Yang (2006): Computational model goal

Real data, Psychologically plausible learning algorithm

Realistic data is important to use since the experimental study of Saffran, Aslin, & Newport (1996) used artificial language data

A psychologically plausible learning algorithm is important since we want to make sure whatever strategy the model uses is something a child could use, too. (Transitional probability would probably work, since Saffran, Aslin, & Newport (1996) showed that infants can track this kind of information in the artificial language.)

Survey of Infant Strategies

Possible strategy: learn from isolated words

Data: 9% of mother-to-child speech is isolated words

Problem: How does a child recognize an isolated word as such?
length won't work: "I-see" vs. "spaghetti"

Possible strategy: statistical properties like transitional probability between syllables

word boundaries postulated at local minima

pre tty bɑ by $p(\text{tty} \rightarrow \text{ba}) < p(\text{pre} \rightarrow \text{tty}), p(\text{ba} \rightarrow \text{by})$

Question: How well does this fare on real data sets (not artificial stimuli)?

Survey of Infant Strategies

Possible strategy: Metrical segmentation strategy

Children treat stressed syllable as beginning of word

- 90% of English content words are stress-initial

Problem: Stress systems differ from language to language

- the child would need to know that words are stress initial

...but to do that, the child needs words *first*

Possible strategy: phonotactic constraints (sequences of consonant clusters that go together, e.g. **str** vs. ***stl** in English); language-specific

- Infants seem to know these by 9 months

- posit boundary at improper sequence break: **stl** --> **st l** (first light)

Problem: May just be syllable boundary (restless)

Survey of Infant Strategies

Possible strategy: Memory

Use previous stored words (sound forms, not meanings) to recognize new words

- if child knows *new*, then can recognize *one* in *thatsanewone*

Problem: Needs to know words before can use this

A good point: "It seems...only language-independent strategies can set word segmentation in motion before the establishment and application of language-specific strategies"

How do we measure word segmentation performance?

Perfect word segmentation:
identify all the words in the speech stream (*recall*)
only identify syllables groups that are actually words (*precision*)

ðəbɪg bæ'd wə'lf
↓
ðə bɪg bæ'd wə'lf
the big bad wolf

How do we measure word segmentation performance?

Perfect word segmentation:
identify all the words in the speech stream (*recall*)
only identify syllables groups that are actually words (*precision*)

ðəbɪg bæ'd wə'lf
↓
ðə bɪg bæ'd wə'lf
the big bad wolf

Recall calculation:
Should have identified 4 words: the, big, bad, wolf
Identified 4 real words: the, big, bad, wolf
Recall Score: $4/4 = 1.0$

How do we measure word segmentation performance?

Perfect word segmentation:
identify all the words in the speech stream (*recall*)
only identify syllables groups that are actually words (*precision*)

ðəbɪg bæ'd wə'lf
↓
ðə bɪg bæ'd wə'lf
the big bad wolf

Precision calculation:
Identified 4 words: the, big, bad, wolf
Identified 4 real words: the, big, bad, wolf
Precision Score: $4/4 = 1.0$

How do we measure word segmentation performance?

Perfect word segmentation:
identify all the words in the speech stream (*recall*)
only identify syllables groups that are actually words (*precision*)

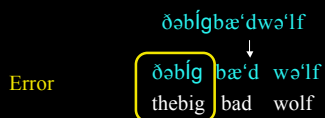
ðəbɪg bæ'd wə'lf
↓
ðəbɪg bæ'd wə'lf
thebig bad wolf

Error

How do we measure word segmentation performance?

Perfect word segmentation:

identify all the words in the speech stream (*recall*)
 only identify syllables groups that are actually words (*precision*)



Recall calculation:

Should have identified 4 words: the, big, bad, wolf

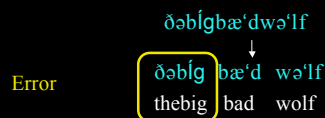
Identified 2 real words: big, bad

Recall Score: $2/4 = 0.5$

How do we measure word segmentation performance?

Perfect word segmentation:

identify all the words in the speech stream (*recall*)
 only identify syllables groups that are actually words (*precision*)



Precision calculation:

Identified 3 words: thebig, bad, wolf

Identified 2 real words: big, bad

Precision Score: $2/3 = 0.666\dots$

How do we measure word segmentation performance?

Perfect word segmentation:

identify all the words in the speech stream (*recall*)
 only identify syllables groups that are actually words (*precision*)

Want good scores on both of these measures

$$F = \frac{1}{\alpha^2 p^2 + (1 - \alpha)^2 r^2}$$

where p is precision, r is recall, and α is a factor that weighs the relative importance of p and r (and is often chosen to be 0.5 in practice).

Computational Model Goal

- psychologically plausible learning algorithm
- real data

Another good point: it's good if the information is in the data, but we also need to know how children could use it

On Psychological Plausibility

“On the one hand, previous computational models often *over-estimate the computational capacity of human learners*. For example, the algorithm in Brent & Cartwright (1996) produces a succession of lexicons, each of which is associated with an evaluation metric that is calculated over the entire learning corpus. A general optimization algorithm ensures that each iteration yields a better lexicon...unlikely that algorithms of such complexity are something a human learner is capable of using.” - Gambell & Yang (2006)

On Psychological Plausibility

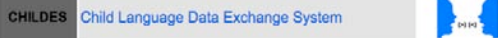
“On the other hand, previous computational models often *under-estimate* the human learner’s knowledge of linguistic representations. Most of these models are ‘synthetic’...the raw material for segmentation is a stream of segments...assumption probably makes the child’s job unnecessarily hard in light of the evidence that it is the syllable, rather than the segment, that makes up the primary units of perception” - Gambell & Yang (2006)

Where does the realistic data come from?

CHILDES

Child Language Data Exchange System
<http://childes.psy.cmu.edu/>

Large collection of child-directed speech data transcribed by researchers. Used to see what children’s input is actually like.



Where does the realistic data come from?

Gambell & Yang (2006)

Looked at Brown corpus files in CHILDES (226,178 words made up of 263,660 syllables).

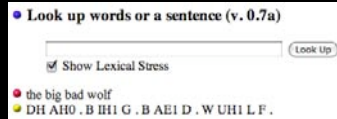
Converted the transcriptions to pronunciations using a pronunciation dictionary called the CMU Pronouncing Dictionary.

<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>



Where does the realistic data come from?

Converting transcriptions to pronunciations



Gambell and Yang (2006) tried to see if a model learning from transitional probabilities between syllables could correctly segment words from realistic data.

ðə bɪg bæ'd wɔ'lf

DH AH0 . B IH1 G . B AE1 D . W UH1 L F .

Segmenting Realistic Data

Gambell and Yang (2006) tried to see if a model learning from transitional probabilities between syllables could correctly segment words from realistic data.

ðə bɪg bæ'd wɔ'lf

DH AH0 . B IH1 G . B AE1 D . W UH1 L F .

Segmenting Realistic Data

Gambell and Yang (2006) tried to see if a model learning from transitional probabilities between syllables could correctly segment words from realistic data.

ðə bɪg bæ'd wɔ'lf

DH AH0 | B IH1 G . | B AE1 D . | W UH1 L F .

the big bad wolf

Modeling Statistical Learning With TrProb

"The model consists of two stages: training and testing. During the training stage, the learner gathers transitional probabilities over adjacent syllables in the learning data. The testing stage does not start until the entire learning data has been processed, and statistical learning is applied to the same data used in the training stage."

"There is a word boundary AB and CD if
 $TP(A \rightarrow B) > TP(B \rightarrow C) < TP(C \rightarrow D)$.

The conjectured word boundaries are then compared against the target segmentation."

Modeling Results for Transitional Probability

Precision: 41.6%

Recall: 23.3%



A learner relying only on transitional probability does not reliably segment words such as those in child-directed English.

About 60% of the words posited by the transitional probability learner are not actually words (41.6% precision) and almost 80% of the actual words are not extracted (23.3% recall).

(Even assuming perfect syllabification of the speech and neutralization of the effects of stress, and using the same data for training and testing.)

Why such poor performance?



"We were surprised by the low level of performance. Upon close examination of the learning data, however, it is not difficult to understand the reason. A necessary condition ... is that words must consist of multiple syllables. If the target sequence of segmentation contains only monosyllabic words, it is clear that statistical learning will fail. A sequence of monosyllabic words requires a word boundary after each syllable; a statistical learner, on the other hand, will only place a word boundary between two sequences of syllables for which the TPs within are higher than that in the middle... Saffran et al. (1996)... the pseudowords are uniformly three syllables long."- Gambell & Yang (2006)

Why such poor performance?



A brief demonstration



Why such poor performance?



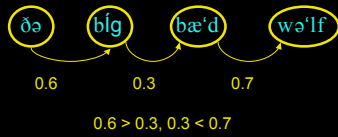
A brief demonstration



Why such poor performance?



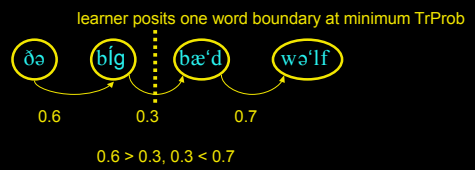
A brief demonstration



Why such poor performance?



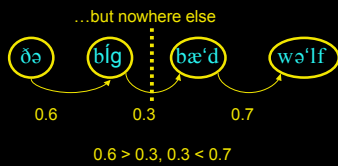
A brief demonstration



Why such poor performance?



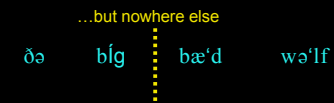
A brief demonstration



Why such poor performance?



A brief demonstration



Why such poor performance?



A brief demonstration

...but nowhere else

ðəbɪg



bæ'dwɔ'lf

Precision for this sequence: 0 words correct out of 2 posited
Recall: 0 words correct out of 4 that should have been posited

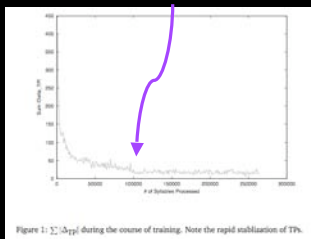
Why such poor performance?



"More specifically, a monosyllabic word is followed by another monosyllabic word 85% of the time. As long as this is the case, statistical learning cannot work." - Gambell & Yang (2006)

Would more data help? Probably not...

point of stabilization ~ 100,000 syllables
(children hear over 1,000,000 words in 6 months)



What about other models that have had success on data like this (Swingley 2005)?

"It is true that overall precision may be quite high for certain values of θ but it is worth noting that most of the three-syllable words determined by Swingley's criteria are wrong: the precision is consistently under 25-30%...regardless of the value of θ . Moreover, statistical criteria...produce very low recall...at best 22-27%." - Gambell & Yang (2006)

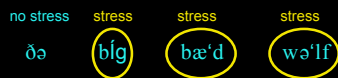
Additional Learning Bias

Gambell & Yang (2006) idea

Children are sensitive to the properties of their native language like stress patterns very early on. Maybe they can use those sensitivities to help them solve the word segmentation problem.

Unique Stress Constraint (USC)

A word can bear at most one **primary stress**.



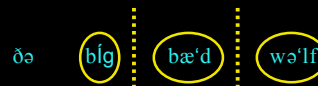
Additional Learning Bias

Gambell & Yang (2006) idea

Children are sensitive to the properties of their native language like stress patterns very early on. Maybe they can use those sensitivities to help them solve the word segmentation problem.

Unique Stress Constraint (USC)

A word can bear at most one **primary stress**.



Learner gains knowledge: These must be separate words

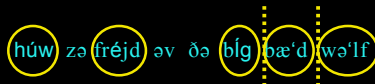
Additional Learning Bias

Gambell & Yang (2006) idea

Children are sensitive to the properties of their native language like stress patterns very early on. Maybe they can use those sensitivities to help them solve the word segmentation problem.

Unique Stress Constraint (USC)

A word can bear at most one **primary stress**.



Get these boundaries because stressed (strong) syllables are next to each other.

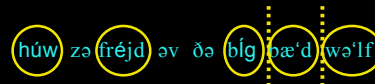
Additional Learning Bias

Gambell & Yang (2006) idea

Children are sensitive to the properties of their native language like stress patterns very early on. Maybe they can use those sensitivities to help them solve the word segmentation problem.

Unique Stress Constraint (USC)

A word can bear at most one **primary stress**.



Can use this in tandem with transitional probabilities when there are weak (unstressed) syllables between stressed syllables.

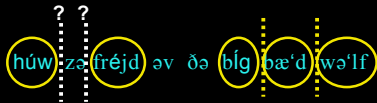
Additional Learning Bias

Gambell & Yang (2006) idea

Children are sensitive to the properties of their native language like stress patterns very early on. Maybe they can use those sensitivities to help them solve the word segmentation problem.

Unique Stress Constraint (USC)

A word can bear at most one **primary stress**.



There's a word boundary at one of these two: use minimum TrProb to figure out where.

USC + Transitional Probabilities

Precision: 73.5%

Recall: 71.2%



A learner relying only on transitional probability but who also has knowledge of the Unique Stress Constraint does a much better job at segmenting words such as those in child-directed English.

Only about 25% of the words posited by the transitional probability learner are not actually words (73.5% precision) and about 30% of the actual words are not extracted (71.2% recall).

"In fact, these figures are comparable to the highest performance in the literature." (Though see Goldwater et al. (2007)).

Another Strategy

Algebraic Learning (Gambell & Yang (2003))

Subtraction process of figuring out unknown words.

"Look, honey - it's a **big goblin**!"

bíggáblɪn



bíg = big (familiar word)

bíggáblɪn

bíg

gáblɪn = (new word)



Evidence of Algebraic Learning in Children

"Behave yourself!"

"I was have!"

(be-have = be + have)

"Was there an adult there?"

"No, there were two dults."

(a-dult = a + dult)

"Did she have the hiccups?"

"Yeah, she was hiccing-up."

(hicc-up = hicc + up)

Using Algebraic Learning + USC

StrongSyl WeakSyl1 WeakSyl2 StrongSyl
ma ny can come
"Many can come..."

Using Algebraic Learning + USC

Familiar word: "many"

StrongSyl WeakSyl1 WeakSyl2 StrongSyl
ma ny can come
"Many can come..."

Using Algebraic Learning + USC

Familiar word: "come"

StrongSyl WeakSyl1 WeakSyl2 StrongSyl
ma ny can come
"Many can come..."

Using Algebraic Learning + USC

This must be a word:
add it to memory

StrongSyl WeakSyl1 WeakSyl2 StrongSyl
ma ny can come
"Many can come..."

Algebraic Learner, More Generally

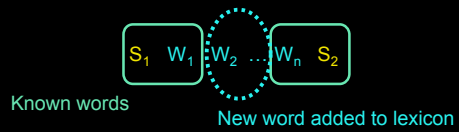
“However, USC may not resolve the word boundaries conclusively. This happens when the learner encounters $S_1W_1^nS_2$: the two S’s stand for strong syllables, and there are n syllables in between, where W_j^i stands for the substring that spans from the i th to the j th weak syllable.”

$S_1 W_1 W_2 \dots W_n S_2$

Algebraic Learner, More Generally

“However, USC may not resolve the word boundaries conclusively. This happens when the learner encounters $S_1W_1^nS_2$: the two S’s stand for strong syllables, and there are n syllables in between, where W_j^i stands for the substring that spans from the i th to the j th weak syllable.”

“If both $S_1W_1^{i-1}$ and $W_{j+1}^nS_2$ are, or are part of, known words on both sides of $S_1W_1^nS_2$, then W_j^i must be a word, and the learner adds W_j^i as a new word into the lexicon.”



Algebraic Learner, More Generally

“However, USC may not resolve the word boundaries conclusively. This happens when the learner encounters $S_1W_1^nS_2$: the two S’s stand for strong syllables, and there are n syllables in between, where W_j^i stands for the substring that spans from the i th to the j th weak syllable.”

“Otherwise...somewhat more complicated.”

“**Agnostic**: the learner ignores the string $S_1W_1^nS_2$ altogether and proceeds to segment the rest of utterance. No word is added.”

$S_1 W_1 W_2 \dots W_n S_2$

Ignore this entire syllable string

Algebraic Learner, More Generally

“However, USC may not resolve the word boundaries conclusively. This happens when the learner encounters $S_1W_1^nS_2$: the two S’s stand for strong syllables, and there are n syllables in between, where W_j^i stands for the substring that spans from the i th to the j th weak syllable.”

“Otherwise...somewhat more complicated.”

“**Random**: the learner picks a random position r ($1 \leq r \leq n$) and splits W_1^n into two substrings W_1^r and W_{r+1}^n ...no word is added to the lexicon.”

$S_1 W_1 W_2 \dots W_n S_2$

Guess $r = 2$, and split.

Algebraic Learning + USC

Agnostic
Precision: 85.9%
Recall: 89.9%



Random
Precision: 95.9%
Recall: 93.4%

"It may seem a bit surprising that the random algebraic learner yields the best segmentation results but this is not expected. The performance of the agnostic learner suffers from deliberately avoiding segmentation in a substring where word boundaries lie. The random learner, by contrast, always picks out *some* word boundary, which is very often correct. And this is purely due to the fact that words in child-directed English are generally short."

Gambell & Yang (2006) Conclusions

"The segmentation process can get off the ground only through language-independent means: experience-independent linguistic constraints such as the USC and experience-[in]dependent statistical learning are the only candidates among the proposed strategies."

"Statistical learning does not scale up to realistic settings."

"Simple principles on phonological structures such as the USC can constrain the applicability of statistical learning and improve its performance."

"Algebraic learning under USC, which has trivial computational cost and is in principle universally applicable, outperforms all other segmentation models."

Gambell & Yang (2006) Conclusions

"It is worth reiterating that our critical stance on statistical learning refers only to a specific kind of statistical learning that exploits local minima over adjacent linguistic units...we simply wish to reiterate the conclusion from decades of machine learning research that no learning, statistical or otherwise, is possible without the appropriate prior assumptions about the representation of the learning data and a constrained hypothesis space...present work, then, can be viewed as an attempt to articulate the specific linguistic constraints that might be built in for successful word segmentation to take place."

Extension: Other Languages

What about other languages besides English?

-English has predictable word order. Some languages don't. Would that destroy a transitional probability learner?

- English words are easily separable - but what about languages where the syntax is less separable from the morphology?

An example from *Chukchi*, a polysynthetic, incorporating, and agglutinating language:

```
Tameyrgaleivpaytarkan.  
Tameyrgaleivpaytarkan  
1.SG.SUBJ-great.head.hurt-PRES.1  
'I have a fierce headache.' (Borok 1961: 102)  
Tameyrgaleivpaytarkan has a 5:1 morpheme-to-word ratio with 3 incorporated lexical morphemes (moyrg 'great', leiv 'head', payt 'ache').
```