

Psych 215L:  
Language Acquisition

Lecture 6  
Word Segmentation

Computational Problem

Divide spoken speech into words

tuðəkæ̀səlbi\_jándðəgáblɪn\_síri

Computational Problem

Divide spoken speech into words

tuðəkæ̀səlbi\_jándðəgáblɪn\_síri



tu ðə kæ̀səl bi\_jánd ðə gáblɪn síri  
to the castle beyond the goblin city

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)

to the castle beyond the goblin city

### 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: *ca + stle* is a common sound sequence

to the castle beyond the goblin city

### 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: *stle + be* is an accidental sound sequence

to the castle beyond the goblin city

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.

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

Read as “the probability of ‘lin’, given that ‘gob’ has just been encountered”

### 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

$$\text{TrProb}(\text{"go" "blin"}) = \text{Prob}(\text{"blin" | "go"})$$

Example of how to calculate TrProb:

go...  
 ...bble, ...bbler, ...bbledygook, ...blet, ...blin  
 (5 options for what could follow "go")

$$\text{TrProb}(\text{"go" "blin"}) = \text{Prob}(\text{"blin" | "go"}) = 1/5$$

### 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)

Idea:  $\text{Prob}(\text{"stle" | "ca"}) = \text{high}$   
 Why? "ca" is often followed by "stle"

to the **castle** beyond the goblin city

### 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)

Idea:  $\text{Prob}(\text{"be" | "stle"}) = \text{lower}$   
 Why? "stle" is not usually followed by "be"

to the castle **be** yond the goblin city  
 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)

$\text{Prob}(\text{"yond" | "be"}) = \text{higher}$   
 Why? "be" is often followed by "yond", among other options

to the castle **beyond** the goblin city

### 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)

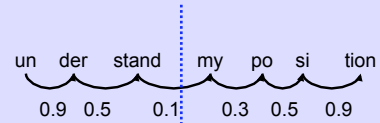
Prob("be" | "stle") < Prob("stle" | "ca")  
 Prob("be" | "stle") < Prob("yond" | "be")

to the castle beyond the goblin city

TrProb learner posits word boundary here, at the **minimum of the transitional probabilities**

Important: doesn't matter what the probability actually is, so long as it's a minimum when compared to the probabilities surrounding it

### Transitional Probability Example



**0.5 < 0.1 0.1 < 0.3**

0.1 = Transitional probability minimum, compared with surrounding transitional probabilities (0.5, 0.3)

Word boundary is here

### 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 (2 minutes).



### Computational Modeling Data (Digital Children)



### How good is transitional probability on real data?

Gambell & Yang (2006): Computational model goal

Realistic data + Psychologically plausible learning algorithm

Realistic data is important to use since the experimental study of Saffran, Aslin, & Newport (1996) used artificial language data (though see Gómez & Gerken (2000) for the value of artificial language studies and Finn & Hudson Kam (2008) and Onnis et al. (2005) for other issues with them); Johnson & Tyler (2010) show that transitional probability tracking abilities can be disrupted by having words of varying lengths - even in an artificial language.

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

Also, experimental evidence from Lew-Williams et al. (2011) suggests that hearing words in isolation is very helpful for segmentation for 6- to 8-month-old infants.

Potential problem: How does a child recognize that a word is isolated? Might just get word chunks instead.

length won't work: "I-see" vs. "spaghetti"

### Survey of Infant Strategies

Possible strategy: statistical properties like transitional probability between syllables

word boundaries postulated at local minima

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

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*

(though see Swingley (2005) for how a child sensitive to mutual information could extract syllable sequences that give clues to the overall stress system of a language before knowing too many words)

### Survey of Infant Strategies

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)

(However, see Blanchard et al. (2010) for using these kind of phonotactic constraints successfully to accomplish word segmentation on realistic child-directed speech data)

### 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*

Mersad & Nazzi 2012: Infants can use familiar words to segment artificial languages with words of different lengths.

Problem: Need to know some words before can use this

Gambell & Yang say: "It seems...*only language-independent strategies can set word segmentation in motion* before the establishment and application of language-specific strategies" – though see Blanchard et al. (2010) for a model that learns language-dependent phonotactic knowledge at the same time word segmentation is occurring.

### How do we measure word segmentation performance?

Perfect word segmentation:

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

ðəbɪgbædwɔːlf



ðə bɪg bæd wɔːlf  
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 syllable groups that are actually words (*precision*)

ðəbɪgbædwɔːlf



ðə bɪg bæd wɔːlf  
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 syllable groups that are actually words (*precision*)

ðəbɪgbædwɔːlf  
 ↓  
 ðə bɪg bæd wɔːlf  
 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 syllable groups that are actually words (*precision*)

ðəbɪgbædwɔːlf  
 ↓  
 ðəbɪg bæd wɔːlf  
 thebig bad wolf

Error

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 syllable groups that are actually words (*precision*)

ðəbɪgbædwɔːlf  
 ↓  
 ðəbɪg bæd wɔːlf  
 thebig bad wolf

Error

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 syllable groups that are actually words (*precision*)

ðəbɪgbædwɔːlf  
 ↓  
 ðəbɪg bæd wɔːlf  
 thebig bad wolf

Error

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?

Note: Recall and precision can be calculated over word tokens (as done here), but also over word boundaries and lexicon items.

ðəbɪg bæd wɔlf

thebig bad wolf

Word boundaries:  
 # identified correctly = 2  
 # identified = 2  
 # should have identified = 3

Boundary precision =  $2/2 = 1.00$   
 Boundary recall =  $2/3 = 0.666$

Note: Boundary precision > boundary recall  
 = indication of undersegmentation

### How do we measure word segmentation performance?

Note: Recall and precision can be calculated over word tokens (as done here), but also over word boundaries and lexicon items.

ðəbɪg bæd wɔlf

thebig bad wolf

Lexicon items:  
 # identified correctly = 2 (bad, wolf)  
 # identified = 3 (thebig, bad, wolf)  
 # should have identified = 3 (the, bad, wolf)

Lexicon item precision =  $2/3 = 0.666$   
 Lexicon item recall =  $2/3 = 0.666$

### How do we measure word segmentation performance?

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

Want good scores on both of these measures  
 (F-score is harmonic mean)

$$F = \frac{1}{\frac{1}{p} + \frac{1}{r}}$$

where p is precision, r is recall, and n 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

- real data
- psychologically plausible learning algorithm

A related point for computational vs. algorithmic-level models: it's good if the information is in the data (which is the way optimal or ideal learner models operate), but we also need to know how children could use those data.



### 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)

Goldwater, Griffiths, & Johnson (2009) and Johnson & Goldwater (2009) explore an ideal Bayesian learner for word segmentation - though the same issue of adapting the probabilistic learner to human limitations arises.

### 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).

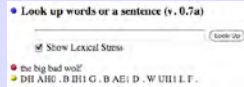
Converted the transcriptions to pronunciations using a pronunciation dictionary called the CMU Pronouncing Dictionary.

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



### Where does the realistic data come from?

Converting transcriptions to pronunciations



Given "the big bad wolf", the pronouncing dictionary produces output like this (which maps to the phonemic representation):

the big bad wolf  
 DH AH0 . B IH1 G . B AE1 D . W UH1 L F .  
 ðə bíg bæd wɔlf

### 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 bíg bæd wɔlf  
 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 bíg bæd wɔlf  
 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?

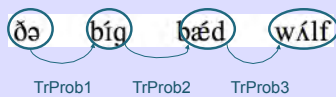


"We were surprised by the low level of performance. Upon close examination of the learning data, however, it is not difficult to understand the reason. A necessary condition ... is that words must consist of multiple syllables. If the target sequence of segmentation contains only monosyllabic words, it is clear that [this kind of] 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?



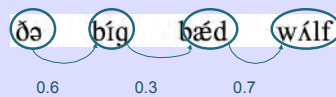
A brief demonstration




### Why such poor performance?

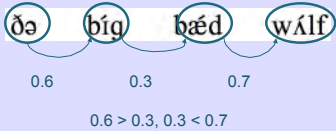


A brief demonstration




Why such poor performance? 

A brief demonstration



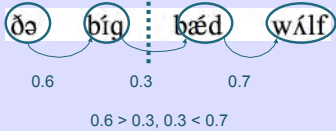
0.6      0.3      0.7

$0.6 > 0.3, 0.3 < 0.7$

Why such poor performance? 


A brief demonstration

learner posits one word boundary at minimum TrProb



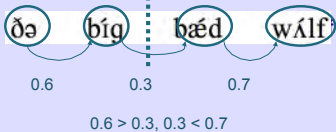
0.6      0.3      0.7

$0.6 > 0.3, 0.3 < 0.7$

Why such poor performance? 

A brief demonstration

...but nowhere else



0.6      0.3      0.7

$0.6 > 0.3, 0.3 < 0.7$

Why such poor performance? 

A brief demonstration

...but nowhere else



0.6      0.3      0.7

$0.6 > 0.3, 0.3 < 0.7$

Why such poor performance?



A brief demonstration

...but nowhere else

ðæbíg      bædwɔlf  
 thebig      badwolf

Note: undersegmentation

Word token & lexicon item precision: 0/2 = 0  
 Word token & lexicon item recall: 0/4 = 0

Boundary precision: 1/1 = 1  
 Boundary recall: 1/3 = .33

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, [this kind of] 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)

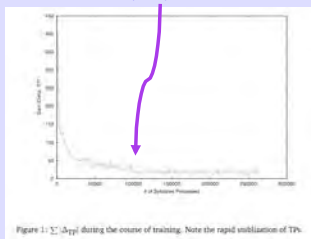


Figure 1:  $\sum |\Delta_{t+1}^i|$  during the course of training. Note the rapid stabilization of TEs.

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  $\theta$  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  $\theta$ . Moreover, statistical criteria...produce very low recall...at best 22-27%." - Gambell & Yang (2006)

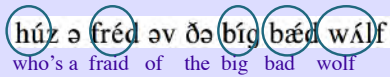
### Additional Learning Bias

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.



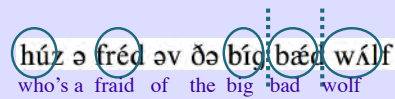
### Additional Learning Bias

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

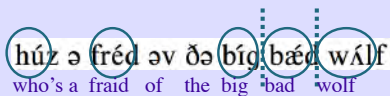
### Additional Learning Bias

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.

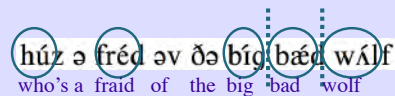
### Additional Learning Bias

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.

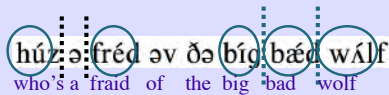
### Additional Learning Bias

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. (2009), Johnson & Goldwater (2009), and Blanchard et al (2010).)

### Another Strategy

Algebraic Learning (Gambell & Yang (2003))

Subtraction process of figuring out unknown words.

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



bɪg = big (familiar word)

bɪggáblɪn  
bɪg

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."

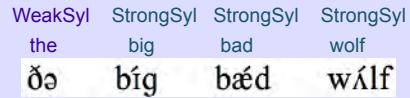
(hicc-up = hicc + up)

### Experimental Evidence of Algebraic Learning

Experimental studies show young infants can use familiar words to segment novel words from their language

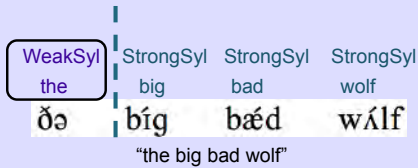
- Bortfeld, Morgan, Golinkoff, & Rathbun 2005:  
6-month-old English infants use their own name or *Mommy/Mama*
- Hallé, Durand, Bardies, & de Boysson 2008  
11-month-old French infants use French articles like *le, les, and la*
- Shi, Werker, & Cutler 2006  
11-month-old English infants use English articles like *her, its, and the*
- Shi, Cutler, Werker, & Cruickshank 2006  
11-month-old English infants (but not 8-month-old English infants) use the English article *the*

### Using Algebraic Learning + USC



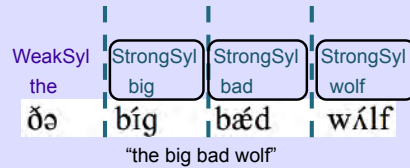
### Using Algebraic Learning + USC

Familiar word: "the" (algebraic learning)



### Using Algebraic Learning + USC

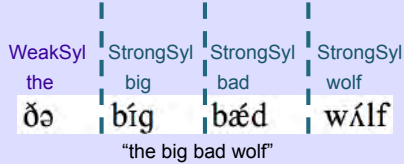
USC says these must be separate words





### Using Algebraic Learning + USC

Correct segmentation!



### Algebraic Learner, More Generally

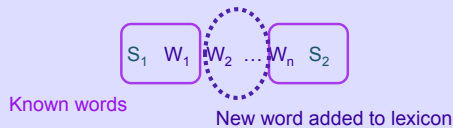
"However, USC may not resolve the word boundaries conclusively. This happens when the learner encounters  $S_1W_i^nS_2$ : the two S's stand for strong syllables, and there are  $n$  syllables in between, where  $W_j$  stands for the substring that spans from the  $i$ th to the  $j$ th weak syllable."

$S_1 W_1 W_2 \dots W_n S_2$

### Algebraic Learner, More Generally

"However, USC may not resolve the word boundaries conclusively. This happens when the learner encounters  $S_1W_i^nS_2$ : the two S's stand for strong syllables, and there are  $n$  syllables in between, where  $W_j$  stands for the substring that spans from the  $i$ th to the  $j$ th weak syllable."

"If both  $S_1W_{i-1}$  and  $W_{j+1}^nS_2$  are, or are part of, known words on both sides of  $S_1W_i^nS_2$ , then  $W_j$  must be a word, and the learner adds  $W_j$  as a new word into the lexicon."

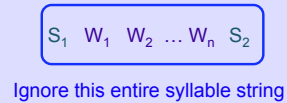


### Algebraic Learner, More Generally

"However, USC may not resolve the word boundaries conclusively. This happens when the learner encounters  $S_1W_i^nS_2$ : the two S's stand for strong syllables, and there are  $n$  syllables in between, where  $W_j$  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  $S_1W_i^nS_2$  altogether and proceeds to segment the rest of utterance. No word is added."

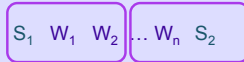


### Algebraic Learner, More Generally

"However, USC may not resolve the word boundaries conclusively. This happens when the learner encounters  $S_1 W_1^n S_2$ : the two S's stand for strong syllables, and there are  $n$  syllables in between, where  $W_i$  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."



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 unexpected. 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: Still true today?

"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."

**Willits, Seidenberg, & Saffran (2009): performance is greatly affected by what units are statistically tracked**

## 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? What about other statistical learners (see Blanchard et al. (2010))?

- English words are easily separable - but what about languages where the syntax is less separable from the morphology? (Specifically, what happens when "words" are longer?)

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

Tameygalivtpeytarkin.  
 1st.SG.SUBJ-great-head-hurt-PRES.1  
 1 have a fierce headache. (Skorik 1981: 102)  
 Tameygalivtpeytarkin has a 5:1 morpheme-to-word ratio with 3 incorporated lexical morphemes (mays 'great', javi 'head', payi 'ache').

## Another important metric for model: Matching developmental trajectories

Lignos 2011

"While many computational models have been created to explore how children might learn to segment words, the focus has largely been on achieving higher levels of performance and exploring cues suggested by artificial learning experiments. We propose a broader focus that includes designing models that display properties of infants' performance as they begin to segment words."

Some empirically-based properties

- (1) Unit of representation = syllable
- (2) Undersegmenting function word (e.g., *it's-a*) happens earlier
- (3) Oversegmenting function words that begin other words (e.g., *behave*) should happen later
- (4) Ends of utterances should be more salient/useful for identifying new words

## Another important metric for model: Matching developmental trajectories

Lignos 2011: A model that has these properties

Similar to Gambell & Yang's USC + Algebraic, except...

- (1) Inserts boundaries left-to-right (from beginning to end of utterance)
- (2) Considers multiple segmentations when there is uncertainty
- (3) Can use stress information (specifically, the USC constraint)

"Rather than focusing on cues in artificial learning experiments which may or may not generalize to the natural development of word segmentation in children, we have shown how a simple algorithm for segmentation mimics many of the patterns seen in infants' developing competence. We believe this work opens the door to a promising line of research that will make a stronger effort to see simulations of language acquisition as not just an unsupervised learning task but rather a modeling task that must take into account a broad variety of phenomena."

## Additional Material

### 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 monotone speech: 

[http://whyfiles.org/058language/images/baby\\_stream.aiff](http://whyfiles.org/058language/images/baby_stream.aiff)

*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 = TrProb("go" "la"), TrProb("pa" "do")

*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

$\text{TrProb}(\text{"ro" "go"}, \text{TrProb}(\text{"ro" "pa"}) = 0.3333... <$   
 $1.0 = \text{TrPrb}(\text{"pi" ro}), \text{TrProb}(\text{"go" "la"}), \text{TrProb}(\text{"pa" "do"})$

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 (non-words)

(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 (non-words)  
(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 ("tupiro" is familiar while "ropitu" never appeared), and didn't notice the differences in 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 (part-words)

(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.