Ling 151/Psych 156A: Acquisition of Language II

Lecture 9

Speech segmentation II

Announcements

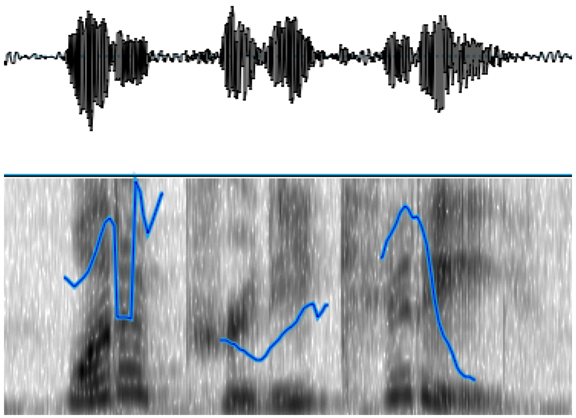
Be working on HW3 (due 1/31/18)

Be working on speech segmentation review questions

Midterm on 2/2/18 (review in class 1/31/18)

Acquisition task

Divide continuous (fluent) speech into individual units (typically words)



= wʌɾəpɹɪɾikɪɾi

wˈʌ ɾə pɹˈɪ ɾi kˈɪ ɾi

wʌɾ ə pɹɪɾi kɪɾi

what a pretty kitty!

Computational modeling

Simulates the mental processes occurring in a child's mind (usually implementing a mathematical description of those processes)

Computational modeling

Simulates the mental processes occurring in a child's mind (usually implementing a mathematical description of those processes)

Computational-level

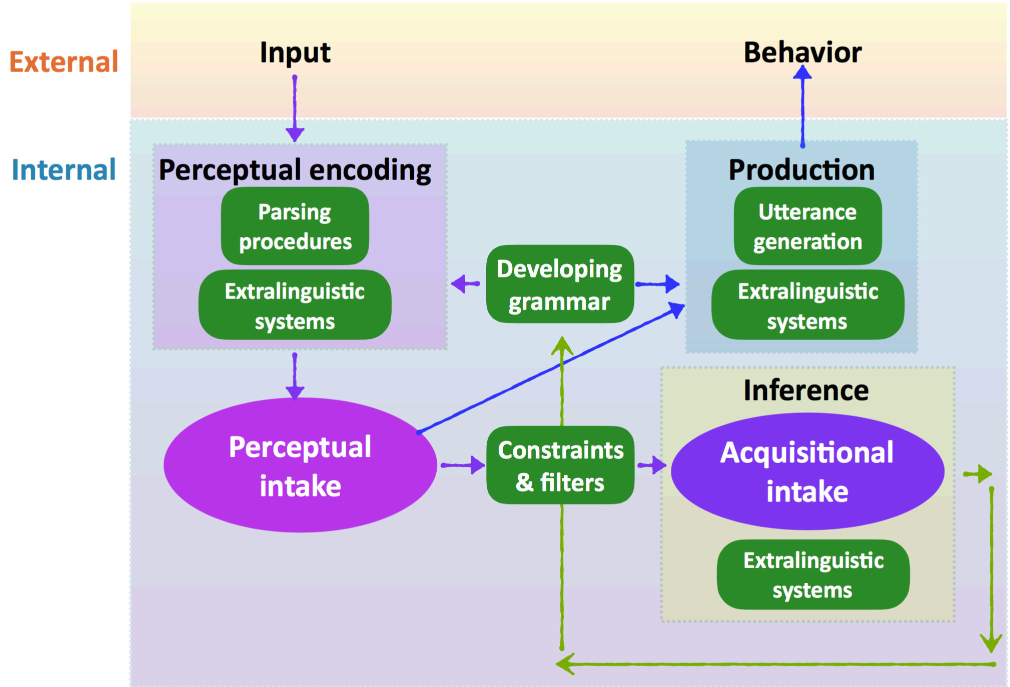
Algorithmic-level

Implementational-level

Important:
Empirically ground the model everywhere we can.

Computational modeling

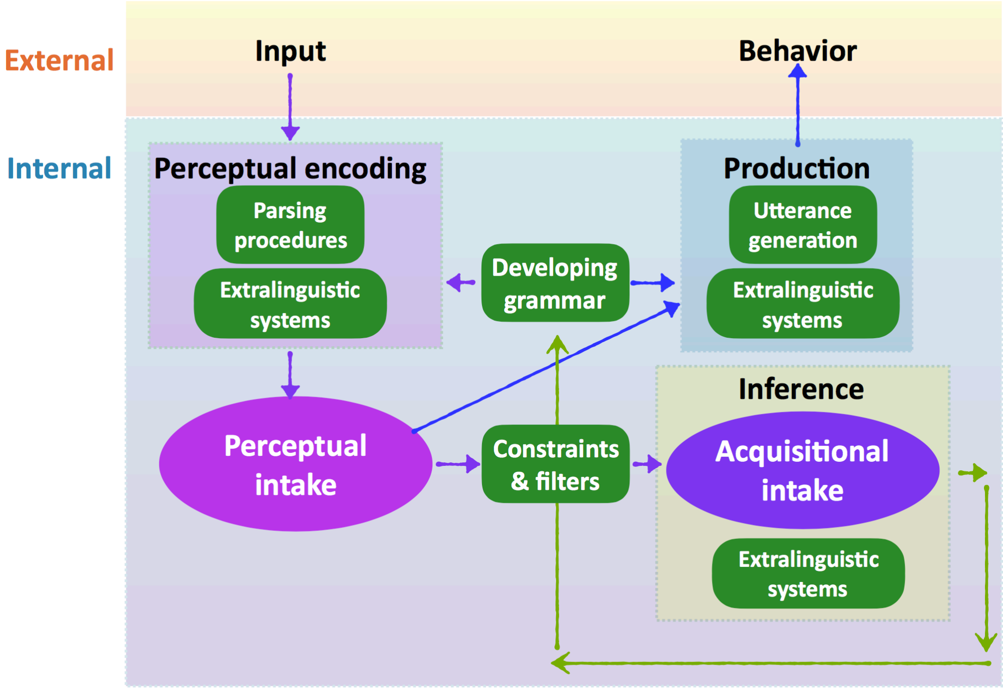
Important:
Empirically ground the model everywhere we can.



That way, when we get model results, we have some confidence that they're true about actual children.

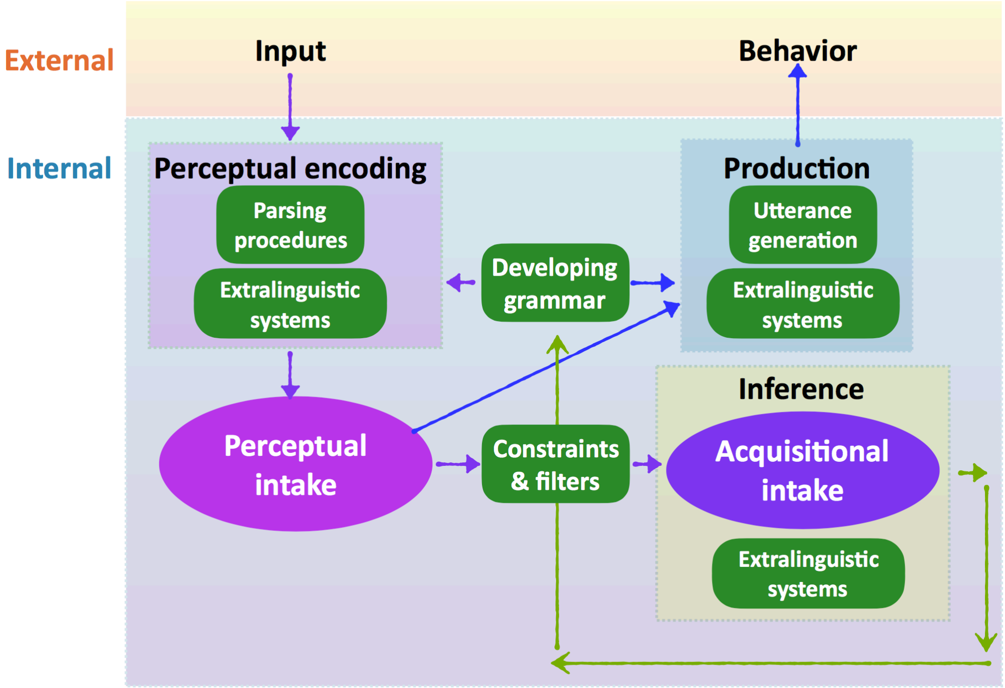
Computational modeling

Utility: Specify and evaluate theories of how acquisition works



Computational modeling

Utility: Specify and evaluate theories of how acquisition works

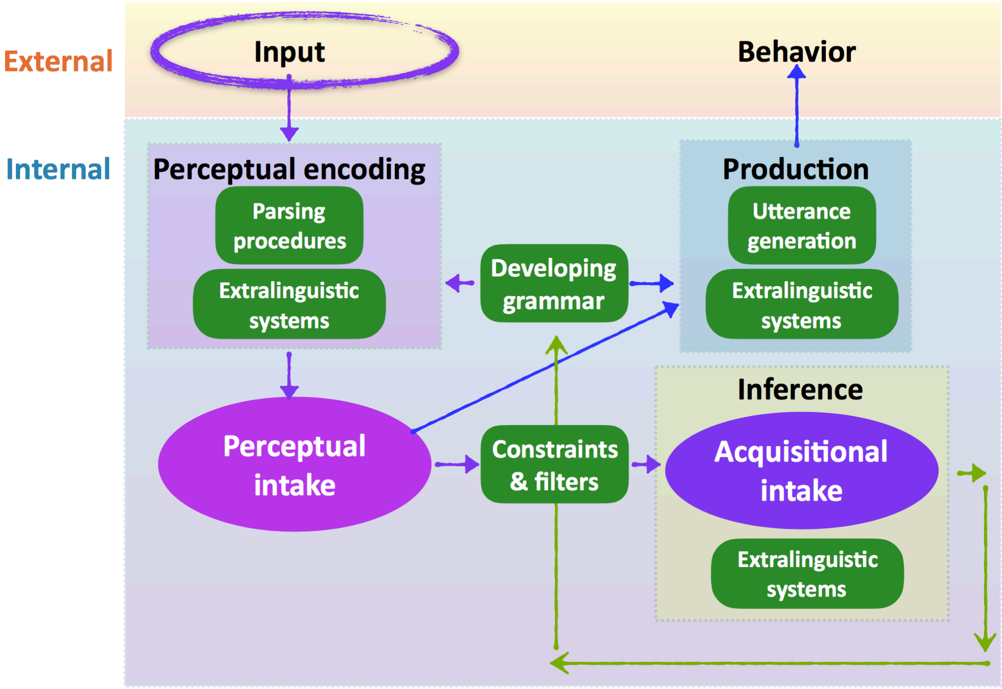


So…let's examine the statistical learning strategy for speech segmentation that relies on **transitional probabilities**.

How good is transitional probability on real data?

Gambell & Yang (2006): Computational model goal

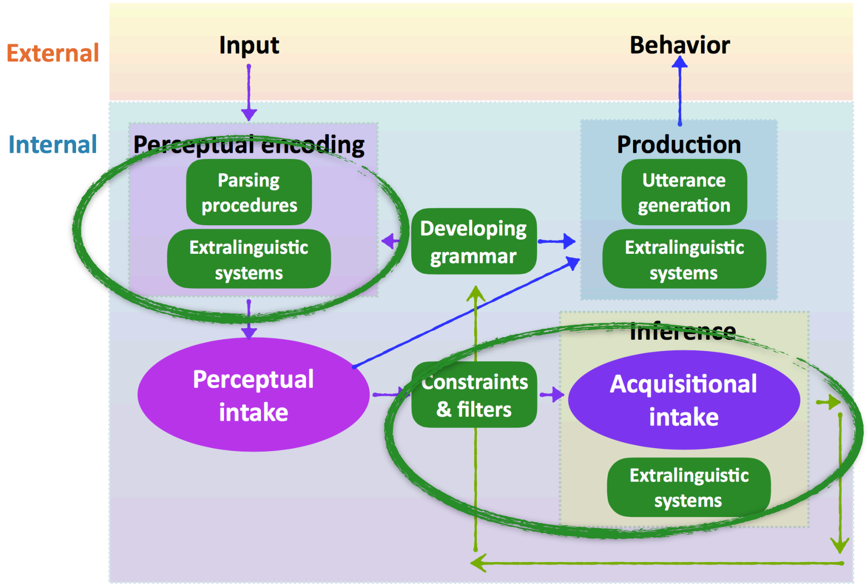
Realistic input is important to use since the experimental study of Saffran, Aslin, & Newport (1996) used artificial language data, and it's not clear how well the results they found will map to real language.



How good is transitional probability on real data?

Gambell & Yang (2006): Computational model goal

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. (Something based on transitional probability would probably work, since Saffran, Aslin, & Newport (1996) showed that infants can track this kind of information in the artificial language.)



How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

wʌˈɾəpɹɪˈɾikɪˈɾi

↓

wʌˈɾ ə pɹɪˈɾi kɪˈɾi

what a pretty kitty

How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

wʌˈɾəpɹɪˈɾikɪˈɾi

↓

wʌˈɾ ə pɹɪˈɾi kɪˈɾi

what a pretty kitty

Recall calculation:

\# of true words found / # of true words

Identified 4 true words: what, a, pretty, kitty

Should have identified 4 words: what, a, pretty, kitty

Recall Score: 4 true words found/4 should have found = **1.0**

How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

wʌˈɾəpɹɪˈɾikɪˈɾi

↓

wʌˈɾ ə pɹɪˈɾi kɪˈɾi

what a pretty kitty

Precision calculation:

\# of true words found / # of words guessed

Identified 4 true words: what, a, pretty, kitty

Identified 4 words total: what, a, pretty, kitty

Precision Score: 4 true words found/4 words found= **1.0**

How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

Undersegmentation

wʌˈɾəpɹɪˈɾikɪˈɾi

↓

wʌˈɾə	pɹɪˈɾi	kɪˈɾi
whata	pretty	kitty

How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

Undersegmentation

wʌˈɾəpɹɪˈɾikɪˈɾi

↓

wʌˈɾə pɹɪˈɾi kɪˈɾi

whata pretty kitty

Recall calculation:

Identified 2 true words: pretty, kitty

Should have identified 4 words: what, a, pretty, kitty

Recall Score: 2 true words found/4 should have found = **0.5**

How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

Undersegmentation

wʌˈɾəpɹɪˈɾikɪˈɾi

↓

wʌˈɾə pɹɪˈɾi kɪˈɾi

whata pretty kitty

Precision calculation:

Identified 2 true words: pretty, kitty

Identified 3 “words” total: whata, pretty, kitty

Precision Score: 2 true words/3 words identified = **0.666**…

How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

wʌˈɾəpɹɪˈɾikɪˈɾi

Undersegmentation **Oversegmentation**

wʌˈɾə pɹɪˈ ɾi kɪˈɾi

whata pre tty kitty

How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

wʌˈɾəpɹɪˈɾikɪˈɾi

Undersegmentation **Oversegmentation**

wʌˈɾə	pɹɪˈ	ɾi	kɪˈɾi
whata	pre	tty	kitty

Recall calculation:

Identified 1 true word: kitty

Should have identified 4 words: what, a, pretty, kitty

Recall Score: 1 true word found/4 should have found = **0.25**

How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

wʌˈɾəpɹɪˈɾikɪˈɾi

Undersegmentation **Oversegmentation**

wʌˈɾə	pɹɪˈ	ɾi	kɪˈɾi
whata	pre	tty	kitty

Precision calculation:

Identified 1 true word: kitty

Identified 4 “words” total: whata, pre, tty, kitty

Precision Score: 1 true word/4 words identified = **0.25**

How do we measure segmentation performance?

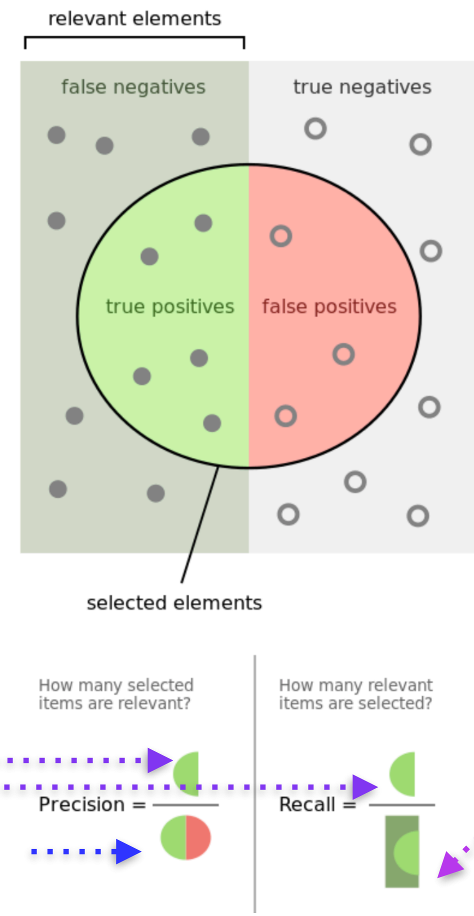
Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

What may be a helpful visualization if you're familiar with signal detection theory

https://en.wikipedia.org/wiki/Precision_and_recall#/media/File:Precisionrecall.svg



true words identified

all identified "words"

true words that should have been identified

How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

Want good enough scores on both of these measures in order to be sure that segmentation is really working

One score that combines precision and recall: F-score

- This is the harmonic mean of precision and recall

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

How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

Perfect segmentation

Recall = 100% (1.0)

Precision = 100% (1.0)

F-score = 2*(1.0 * 1.0)/(1.0 + 1.0) = 1.0

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

How do we measure segmentation performance?

Perfect adult-like segmentation:

identify all the words in the speech stream (*recall*)

only identify syllables groups that are actually words (*precision*)

Not-so-perfect segmentation

Recall = 50% (0.50)

Precision = 67% (0.67)

F-score = 2*(0.50 * 0.67)/(0.50 + 0.67) = 0.57

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

Back to modeling speech segmentation

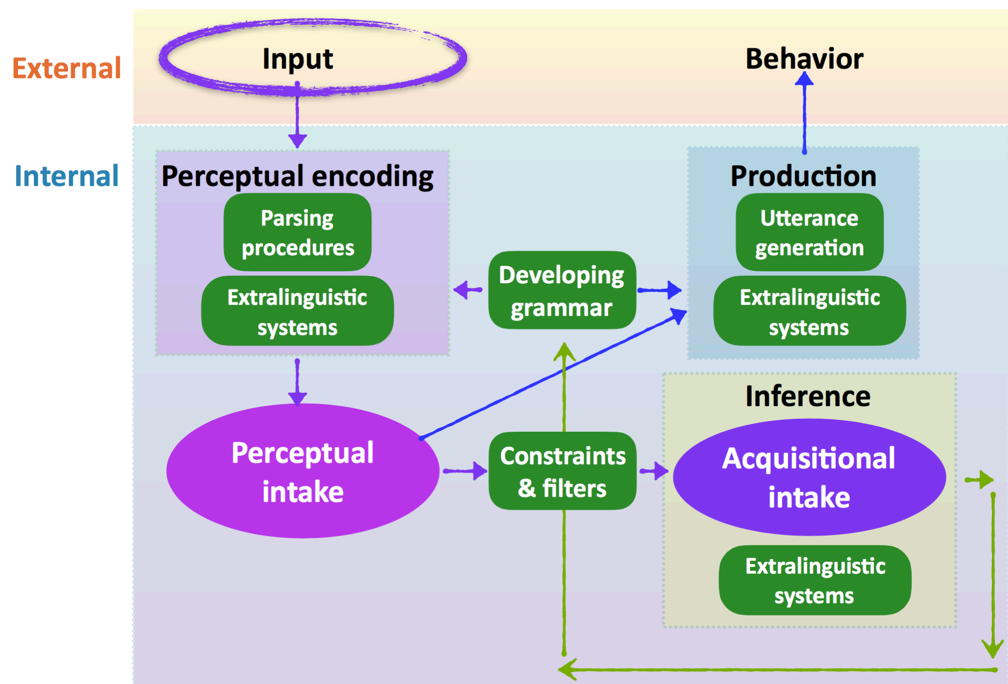
Where does the input data come from?

Gambell & Yang (2006)

http://childes.talkbank.org

CHILDES Child Language Data Exchange System

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



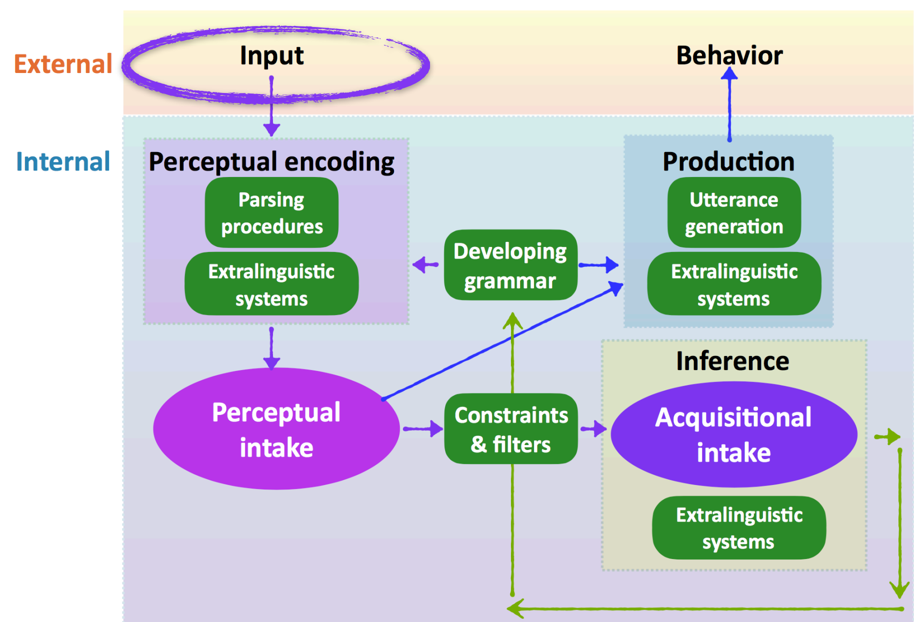
Where does the input data come from?

Gambell & Yang (2006)

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

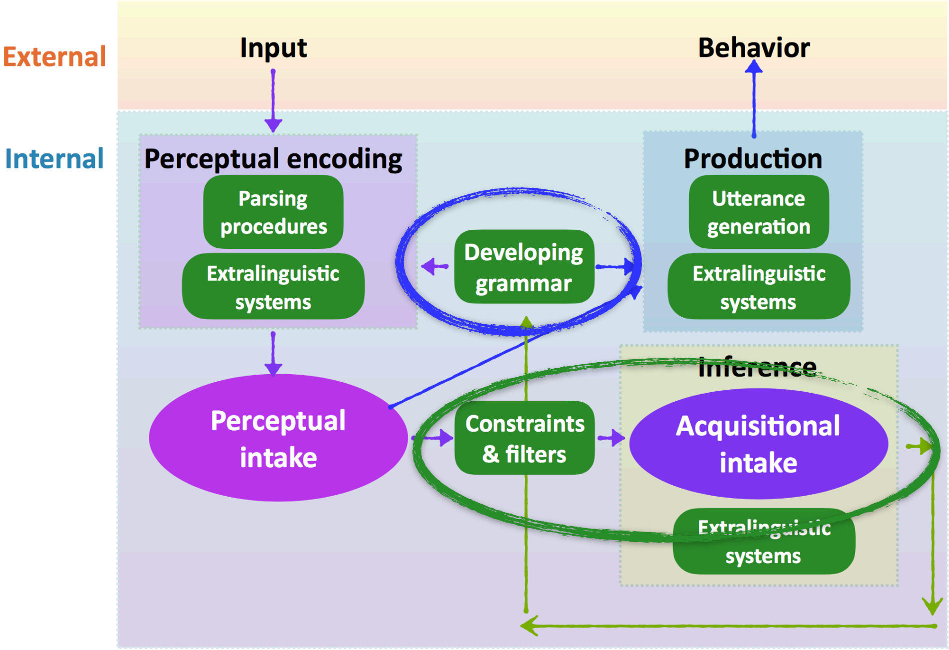
The CMU Pronouncing Dictionary

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



The modeled strategy

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



The modeled strategy

Gambell and Yang (2006)

Specific strategy implemented:
Place a boundary at a **transitional probability minimum**.

"There is a word boundary AB and CD if

TrProb(A --> B) > TrProb(B-->C) < TrProb(C --> D)."

The modeled strategy

Gambell and Yang (2006)

Specific strategy implemented:
Place a boundary at a **transitional probability minimum**.

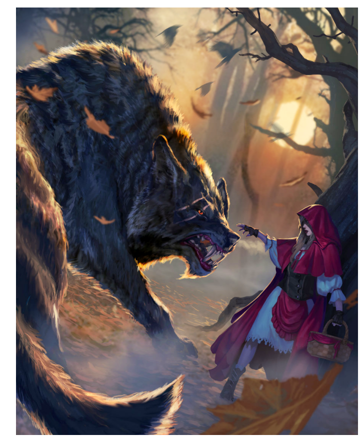
Desired segmentation

ðəbɪˈgbæˈdwʌˈlf

↓

ðə bɪˈg bæˈd wʌˈlf

the big bad wolf



Modeling results for transitional probability

Precision: 41.6%

Recall: 23.3%

F-score: 29.9%

A learner relying only on transitional probability this way 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 identified (23.3% recall).

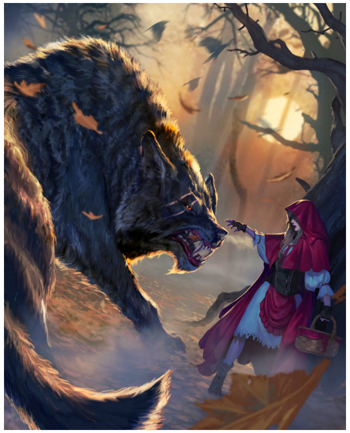
Why such poor performance?

"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 sequence of monosyllabic words requires a word boundary after each syllable; a [transitional probability] learner, on the other hand, will only place a word boundary between two sequences of syllables for which the [transitional probabilities] within [those sequences] are higher than [those of surrounding the sequences]..." - Gambell & Yang (2006)

Why such poor performance?

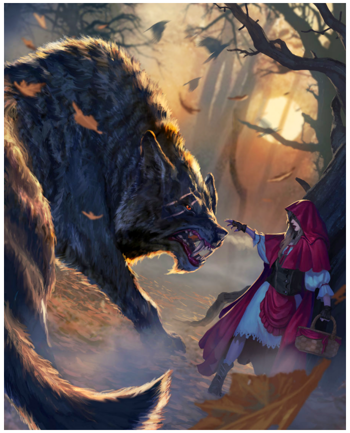
"a sequence of monosyllabic words requires a word boundary after each syllable" - Gambell & Yang (2006)

ðə | bɪˈg | bæˈd | wʌˈlf



Why such poor performance?

“a [transitional probability] learner, on the other hand, will only place a word boundary between two sequences of syllables for which the [transitional probabilities] within [those sequences] are higher than [those of surrounding the sequences]...” - Gambell & Yang (2006)

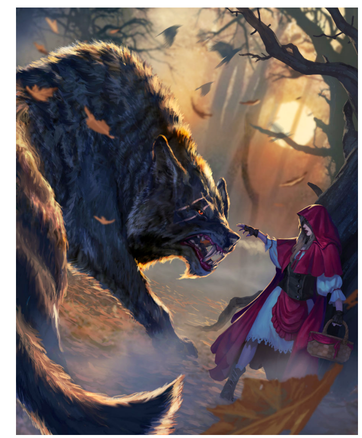


ðə bɪˈg bæˈd wʌˈlf

TrProb1 TrProb2 TrProb3

Why such poor performance?

"a [transitional probability] learner, on the other hand, will only place a word boundary between two sequences of syllables for which the [transitional probabilities] within [those sequences] are higher than [those of surrounding the sequences]…" - Gambell & Yang (2006)



ðə bɪˈg bæˈd wʌˈlf

0.6 **0.3** **0.7**

Best case scenario

0.6 > 0.3 < 0.7

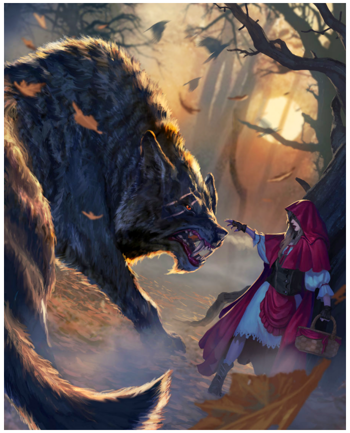
Why such poor performance?

"a [transitional probability] learner, on the other hand, will only place a word boundary between two sequences of syllables for which the [transitional probabilities] within [those sequences] are higher than [those of surrounding the sequences]..." - Gambell & Yang (2006)

learner posits one word boundary at minimum TrProb

ðə bɪˈg | bæˈd wʌˈlf

0.6 **0.3** **0.7**

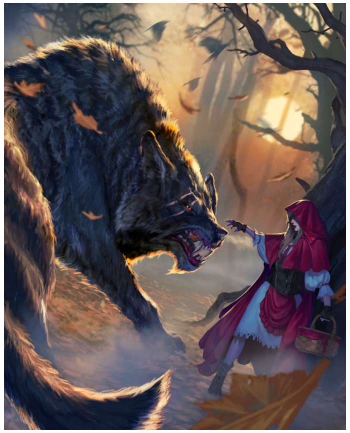


Best case scenario

0.6 > 0.3 < 0.7

Why such poor performance?

"a [transitional probability] learner, on the other hand, will only place a word boundary between two sequences of syllables for which the [transitional probabilities] within [those sequences] are higher than [those of surrounding the sequences]…" - Gambell & Yang (2006)



…and nowhere else

ðə bɪˈg | bæˈd wʌˈlf

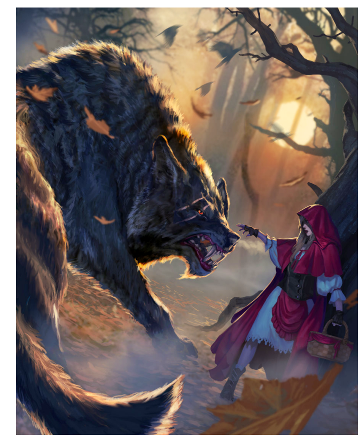
0.6 **0.3** **0.7**

Best case scenario

$0.6 > 0.3 < 0.7$

Why such poor performance?

"a [transitional probability] learner, on the other hand, will only place a word boundary between two sequences of syllables for which the [transitional probabilities] within [those sequences] are higher than [those of surrounding the sequences]..." - Gambell & Yang (2006)



ðəbɪˈg | bæˈdwʌˈlf

thebig | badwolf

Recall: 0 true words found out of 4 that should have been found = **0.0**

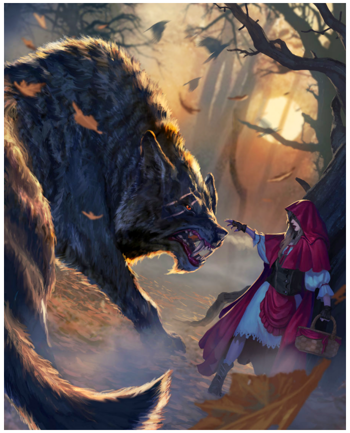
Precision: 0 true words found out of 2 "words" found = **0.0**

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 transitional probability learner] cannot work." - Gambell & Yang (2006)

ðəbɪˈg | bæˈdwʌˈlf

thebig | badwolf



Recall: 0 true words found out of 4 that should have been found = **0.0**

Precision: 0 true words found out of 2 "words" found = **0.0**

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 segmentation problem.

Hypothesis: Unique Stress Constraint (USC)

Children think a word can bear at most one primary stress.

no stress	stress	stress	stress
ðə	bɪˈg	bæˈd	wʌˈlf
the	big	bad	wolf

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 segmentation problem.

Hypothesis: Unique Stress Constraint (USC)

Children think a word can bear at most one primary stress.

no stress	stress	stress	stress
ðə	bɪˈg	bæˈd	wʌˈlf
the	big	bad	wolf

Learner gains knowledge: These must be separate words

Additional learning bias

Gambell & Yang (2006) idea

This Unique Stress Constraint (USC) knowledge could be used in combination with other cues like transitional probability.

huˈwz ə fɹeˈd əv ðə bɪˈg | bæˈd | wʌˈlf

who's a fraid of the big | bad | wolf

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

Additional learning bias

Gambell & Yang (2006) idea
This Unique Stress Constraint (USC) knowledge could be used in combination with other cues like transitional probability.

huˈwz ə fɹeˈd əv ðə bɪˈg bæˈd wʌˈlf
who's a fraid of the big bad wolf

There must be a boundary at one of these places because of the stressed syllables — the stressed syllables can't be in the same word.

Additional learning bias

Gambell & Yang (2006) idea

This Unique Stress Constraint (USC) knowledge could be used in combination with other cues like transitional probability.

huˈwz ə fɹeˈd əv ðə bɪˈg bæˈd wʌˈlf

who's a fraid of the big bad wolf

Maybe transitional probability can help decide and recover some of the boundaries correctly…

Additional learning bias

Gambell & Yang (2006) idea

This Unique Stress Constraint (USC) knowledge could be used in combination with other cues like transitional probability.

huˈwz ə fɹeˈd əv ðə bɪˈg bæˈd wʌˈlf

who's a fraid of the big bad wolf

TrProb1 TrProb2 TrProb3

Maybe transitional probability can help decide and recover some of the boundaries correctly…

Additional learning bias

Gambell & Yang (2006) idea

This Unique Stress Constraint (USC) knowledge could be used in combination with other cues like transitional probability.

huˈwz ə fɹeˈd əv ðə bɪˈg bæˈd wʌˈlf

who's a fraid of the big bad wolf

0.9 0.7 0.8

A minimum transitional probability learner would put a boundary here.

That's one more boundary that we needed!

USC + Transitional Probabilities

Precision: 73.5%

Recall: 71.2%

F-score: 72.3%

A learner relying 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).

Another strategy

Using words you recognize to help you figure out words you don't recognize (a implementation of the "familiar words" strategy)

Another strategy: Algebraic learning

Algebraic learning (Gambell & Yang 2003)

Subtraction process of figuring out unknown words.

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

bíggáblın

bíg= big (familiar word)

bíggáblın

bíg

gáblın = (new word)

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*

- 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*

Experimental evidence of algebraic learning

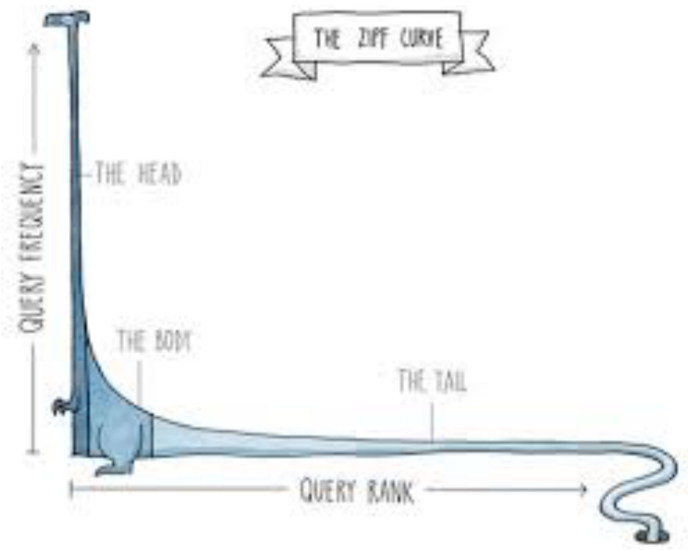
Experimental studies show young infants can use familiar words to segment novel words from their language

- Hallé, Durand, Bardies, & de Boysson 2008
  11-month-old French infants use French articles like *le, les,* and *la*

- Mersad & Nazzi 2012
  8-month-old French infants can use words like *mamã* to segment words in an artificial language

Computational support for algebraic learning

Kurumada, Meylan, & Frank (2013) discovered that the Zipfian nature of natural language data is much more beneficial to a segmentation strategy that looks for coherent chunks (like an algebraic learning strategy would).



Using algebraic learning + USC

no stress	stress	stress	stress
ðə	bɪˈg	bæˈd	wʌˈlf
the	big	bad	wolf

Using algebraic learning + USC

Familiar word: “the” (algebraic learning)

no stress	stress	stress	stress
ðə	bɪˈg	bæˈd	wʌˈlf
the	big	bad	wolf

Using algebraic learning + USC

USC: Only one stress per word - so two more boundaries go in to separate the stressed syllables

no stress	stress	stress	stress
ðə	bɪˈg	bæˈd	wʌˈlf
the	big	bad	wolf

Correct segmentation!

Algebraic learning + USC

Precision: 95.9%

Recall: 93.4%

F-score: 94.6%

A learner relying on algebraic learning and who also has knowledge of the Unique Stress Constraint does a really great job at segmenting words such as those in child-directed English - even better than one relying on the transitional probability between syllables.

Only about 5% of the words posited by the transitional probability learner are not actually words (95.9% precision) and about 7% of the actual words are not extracted (93.4% recall).

Gambell & Yang 2006 summary

Using a simple learning strategy involving transitional probabilities doesn't work so well on realistic data, even though experimental research suggests that infants are capable of tracking and learning from this information.

Gambell & Yang 2006 summary

Models of children that have additional knowledge about the stress patterns of words seem to have a much better chance of succeeding at segmentation if they learn via a simple transitional-probability-based strategy.

However, models of children that use algebraic learning (i.e., familiar words) and have additional knowledge about the stress patterns of words perform even better at segmentation than any of the models using a simple transitional probability strategy.

Gambell & Yang 2006 summary

Sandoval & Gómez 2016

	Age (in months)						
	5	6	7	8	9	10	11
Cue use							
statistical	✓[1]			✓[2,3]			✓[4]
lexical		✓[5]		✓[6,7]			✓[7]
metrical			✓[8]	✓[9]	✓[8]		
Cue weighting							
statistical *vs.*	✓[1] (metrical)		✓[12] (metrical)	✖[2] (metrical; coarticulation)	✖[12] (metrical)		✖[4] (metrical)
metrical *vs.*	✖[1] (statistical)		✖[12] (statistical)	✓[2] (statistical)	✓[12,13] (statistical; phonotactics)		✓[4] (statistical)

References: 1) Thiessen & Erikson, 2013; 2) Johnson & Jusczyk, 2001; 3) Saffran et al., 1996; 4) Johnson & Seidl, 2009; 5) Bortfeld et al., 2005; 6) Shi & Lepage, 2008; 7) Shi et al., 2006; 8) Curtin et al. 2005; 9) Jusczyk, Houston, et al., 1999; 10) Mattys & Jusczyk, 2001; 11) Jusczyk, Hohne et al., 1999; 12) Thiessen & Saffran, 2003; 13) Mattys et al. 1999.

Combining cues like familiar words (via algebraic learning) and metrical stress patterns seems like something that would work well for 8-month-olds.

Gambell & Yang 2006 summary

Sandoval & Gómez 2016

	Age (in months)						
	5	6	7	8	9	10	11
Cue use							
statistical	✓[1]			✓[2,3]			✓[4]
lexical		✓[5]		✓[6,7]			✓[7]
metrical			✓[8]	✓[9]	✓[8]		
Cue weighting							
statistical *vs.*	✓[1] (metrical)		✓[12] (metrical)	✗[2] (metrical; coarticulation)	✗[12] (metrical)		✗[4] (metrical)
metrical *vs.*	✗[1] (statistical)		✗[12] (statistical)	✓[2] (statistical)	✓[12,13] (statistical; phonotactics)		✓[4] (statistical)

References: 1) Thiessen & Erikson, 2013; 2) Johnson & Jusczyk, 2001; 3) Saffran et al., 1996; 4) Johnson & Seidl, 2009; 5) Bortfeld et al., 2005; 6) Shi & Lepage, 2008; 7) Shi et al., 2006; 8) Curtin et al. 2005; 9) Jusczyk, Houston, et al., 1999; 10) Mattys & Jusczyk, 2001; 11) Jusczyk, Hohne et al., 1999; 12) Thiessen & Saffran, 2003; 13) Mattys et al. 1999.

Börschinger & Johnson (2014) demonstrated how a very sophisticated statistical learner (a learner with some idea about how languages are organized) can quickly learn that the Unique Stress Constraint exists at the same time this learner is learning how to segment words out of fluent speech in English.

Gambell & Yang 2006 summary

Sandoval & Gómez 2016

	Age (in months) 5	6	7	8	9	10	11
Cue use							
statistical	✔[1]			✔[2,3]			✔[4]
lexical		✔[5]		✔[6,7]			✔[7]
metrical			✔[8] ✔[9]		✔[8]		
Cue weighting							
statistical *vs.*	✔[1] (metrical)		✔[12] (metrical)	✖[2] (metrical; coarticulation)	✖[12] (metrical)		✖[4] (metrical)
metrical *vs.*	✖[1] (statistical)		✖[12] (statistical)	✔[2] (statistical)	✔[12,13] (statistical; phonotactics)		✔[4] (statistical)

References: 1) Thiessen & Erikson, 2013; 2) Johnson & Jusczyk, 2001; 3) Saffran et al., 1996; 4) Johnson & Seidl, 2009; 5) Bortfeld et al., 2005; 6) Shi & Lepage, 2008; 7) Shi et al., 2006; 8) Curtin et al. 2005; 9) Jusczyk, Houston, et al., 1999; 10) Mattys & Jusczyk, 2001; 11) Jusczyk, Hohne et al., 1999; 12) Thiessen & Saffran, 2003; 13) Mattys et al. 1999.

But what about younger than this? Very young infants seem to rely on statistical cues alone to get started.

Gambell & Yang 2006 summary

Sandoval & Gómez 2016

	Age (in months) 5	6	7	8	9	10	11
Cue use							
statistical	✓[1]			✓[2,3]			✓[4]
lexical		✓[5]		✓[6,7]			✓[7]
metrical			✓[8] ✓[9]		✓[8]		
Cue weighting							
statistical *vs.*	✓[1] (metrical)		✓[12] (metrical)	✖[2] (metrical; coarticulation)	✖[12] (metrical)		✖[4] (metrical)
metrical *vs.*	✖[1] (statistical)		✖[12] (statistical)	✓[2] (statistical)	✓[12,13] (statistical; phonotactics)		✓[4] (statistical)

References: 1) Thiessen & Erikson, 2013; 2) Johnson & Jusczyk, 2001; 3) Saffran et al., 1996; 4) Johnson & Seidl, 2009; 5) Bortfeld et al., 2005; 6) Shi & Lepage, 2008; 7) Shi et al., 2006; 8) Curtin et al. 2005; 9) Jusczyk, Houston, et al., 1999; 10) Mattys & Jusczyk, 2001; 11) Jusczyk, Hohne et al., 1999; 12) Thiessen & Saffran, 2003; 13) Mattys et al. 1999.

✗ píma vs. latú

✓ píma vs. pimá

Skoruppa, Pons, Bosch, Christophe, Cabrol, & Peperkamp 2012:
6-month-old Spanish and French infants don't appear to even recognize the difference between words with initial vs. final lexical stress unless the word forms are identical. (No generalization of lexical stress patterns for words.)

Gambell & Yang 2006 summary

Sandoval & Gómez 2016

	Age (in months) 5	6	7	8	9	10	11
Cue use							
statistical	✔[1]			✔[2,3]			✔[4]
lexical		✔[5]		✔[6,7]			✔[7]
metrical			✔[8] ✔[9]		✔[8]		
Cue weighting							
statistical *vs.*	✔[1] (metrical)		✔[12] (metrical)	✖[2] (metrical; coarticulation)	✖[12] (metrical)		✖[4] (metrical)
metrical *vs.*	✖[1] (statistical)		✖[12] (statistical)	✔[2] (statistical)	✔[12,13] (statistical; phonotactics)		✔[4] (statistical)

References: 1) Thiessen & Erikson, 2013; 2) Johnson & Jusczyk, 2001; 3) Saffran et al., 1996; 4) Johnson & Seidl, 2009; 5) Bortfeld et al., 2005; 6) Shi & Lepage, 2008; 7) Shi et al., 2006; 8) Curtin et al. 2005; 9) Jusczyk, Houston, et al., 1999; 10) Mattys & Jusczyk, 2001; 11) Jusczyk, Hohne et al., 1999; 12) Thiessen & Saffran, 2003; 13) Mattys et al. 1999.

Is it possible that very young infants are using other (more sophisticated) statistical learning strategies?

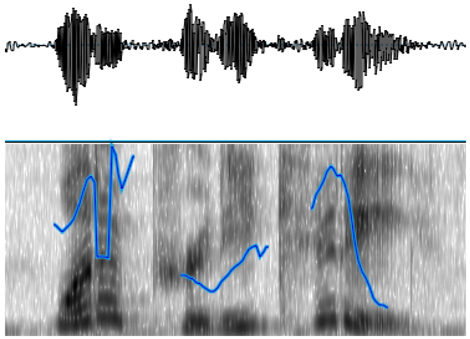
One idea: Bayesian inference

What if children can use Bayesian inference?
Human cognitive behavior is consistent with this kind of reasoning.
(Tenenbaum & Griffiths 2001, Griffiths & Tenenbaum 2005,
Xu & Tenenbaum 2007, Perfors et al. 2011, Pearl & Mis 2016)

Bayesian inference is a sophisticated kind of probabilistic reasoning that tries to find hypotheses that

(1) are consistent with the observed data
(2) conform to a child's prior expectations

One idea: Bayesian inference

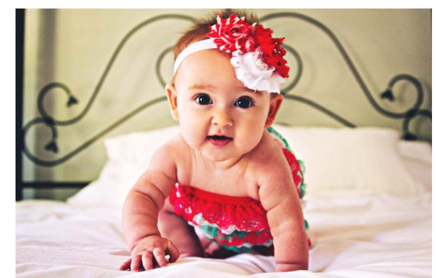


= wʌɾəpɹɪɾikɪɾi

wʌɾ ə pɹɪɾi kɪɾi

what a pretty kitty!

Investigating a Bayesian inference strategy for the very early stages of speech segmentation occurring around six months



Phillips & Pearl 2012, 2014a, 2014b, 2015a, 2015b, Pearl & Phillips in press

$$P(s|u) \propto P(s)P(u|s)$$

One idea: Bayesian inference

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

Strategy: Identify a proto-lexicon of words that best generates the observable fluent speech utterances

Mathematically encoded preferences:

wʌɾə
pɹɪɾi
kɪɾi

wʌ
ɾə
pɹɪɾikɪɾi

wʌɾə
pɹɪɾikɪɾi

Phillips & Pearl 2012, 2014a, 2014b, 2015a, 2015b, Pearl & Phillips in press

One idea: Bayesian inference

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

Strategy: Identify a proto-lexicon of words that best generates the observable fluent speech utterances

Mathematically encoded preferences:

(1) Prefer shorter words

wʌɾə
pɹɪɾi
kɪɾi

wʌ
ɾə
pɹɪɾikɪɾi

wʌɾə
pɹɪɾikɪɾi

Phillips & Pearl 2012, 2014a, 2014b, 2015a, 2015b, Pearl & Phillips in press

One idea: Bayesian inference

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

Strategy: Identify a proto-lexicon of words that best generates the observable fluent speech utterances

Mathematically encoded preferences:

(1) Prefer shorter words

(2) Prefer lexicons with fewer words

wʌɾə
pɹɪɾi
kɪɾi

wʌ
ɾə
pɹɪɾikɪɾi

wʌɾə
pɹɪɾikɪɾi

Phillips & Pearl 2012, 2014a, 2014b, 2015a, 2015b, Pearl & Phillips in press

One idea: Bayesian inference

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

posterior probability

Strategy: Identify a proto-lexicon of words that best generates the observable fluent speech utterances

Mathematically encoded preferences:

(1) Prefer shorter words

(2) Prefer lexicons with fewer words

wʌɾə
pɹɪɾi
kɪɾi

wʌ
ɾə
pɹɪɾikɪɾi

wʌɾə
pɹɪɾikɪɾi

Find the best segmentation

Phillips & Pearl 2012, 2014a, 2014b, 2015a, 2015b, Pearl & Phillips in press

One idea: Bayesian inference

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

prior probability

Strategy: Identify a proto-lexicon of words that best generates the observable fluent speech utterances

Mathematically encoded preferences:

(1) Prefer shorter words

(2) Prefer lexicons with fewer words

wʌɾə
pɹɪɾi
kɪɾi

wʌ
ɾə
pɹɪɾikɪɾi

wʌɾə
pɹɪɾikɪɾi

Find the best segmentation that balances these proto-lexicon preferences

Phillips & Pearl 2012, 2014a, 2014b, 2015a, 2015b, Pearl & Phillips in press

One idea: Bayesian inference

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

likelihood probability

Strategy: Identify a proto-lexicon of words that best generates the observable fluent speech utterances

Mathematically encoded preferences:

(1) Prefer shorter words

(2) Prefer lexicons with fewer words

wʌɾə
pɹɪɾi
kɪɾi

wʌ
ɾə
pɹɪɾikɪɾi

wʌɾə
pɹɪɾikɪɾi

Find the best segmentation that balances these proto-lexicon preferences

and can generate the observable fluent speech utterances

Phillips & Pearl 2012, 2014a, 2014b, 2015a, 2015b, Pearl & Phillips in press

Bayesian inference for speech segmentation

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

What kind of hypotheses might a child have for segmentation?

Bayesian inference for speech segmentation

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

Observed data: 

"tothecastlebeyondthegoblincity"

Hypothesis = sequence of lexical items producing this observable data

Bayesian inference for speech segmentation

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

Observed data:
"tothecastlebeyondthegoblincity"

Some sample hypotheses

Hypothesis 1:
"tothe castle beyond thegoblin city"
Items: *tothe, castle, beyond, thegoblin, city*

Hypothesis 2:
"to the castle beyond the goblin city"
Items: *to, the, castle, beyond, goblin, city*
Note: the *is observed twice in the utterance*

Bayesian inference for speech segmentation

"tothecastlebeyondthegoblincity"

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

Mathematically encoded preferences:

(1) Prefer shorter words

(2) Prefer lexicons with fewer words

Find the best segmentation that balances these proto-lexicon preferences and can generate the observable fluent speech utterances

Hypothesis 1:
"tothe castle beyond thegoblin city"
Items: *tothe, castle, beyond, thegoblin, city*

Hypothesis 2:
"to the castle beyond the goblin city"
Items: *to, the, castle, beyond, goblin, city*
Note: the *is observed twice in the utterance*

Bayesian inference for speech segmentation

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

“tothecastlebeyondthegoblincity”

Mathematically encoded preferences:

Find the best segmentation that balances these proto-lexicon preferences and can generate the observable fluent speech utterances

(2) Prefer lexicons with fewer words

Hypothesis 1:

“tothe castle beyond thegoblin city”

Items: *tothe, castle, beyond, thegoblin, city*

(1) Prefer shorter words

word length: 2.2 syl

Hypothesis 2:

“to the castle beyond the goblin city”

Items: *to, the, castle, beyond, goblin, city*

Note: the *is observed twice in the utterance*

✓

word length: 1.7 syl

Bayesian inference for speech segmentation

"tothecastlebeyondthegoblincity"

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

Mathematically encoded preferences:

(1) Prefer shorter words

Find the best segmentation that balances these proto-lexicon preferences and can generate the observable fluent speech utterances

(2) Prefer lexicons with fewer words

Hypothesis 1:
"tothe castle beyond thegoblin city"
Items: *tothe, castle, beyond, thegoblin, city*

 # words: 5 ✓

✓ **shorter words**

Hypothesis 2:
"to the castle beyond the goblin city"
Items: *to, the, castle, beyond, goblin, city*
Note: the *is observed twice in the utterance*

 # words: 6

Bayesian inference for speech segmentation

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

A Bayesian learner makes a decision based on how important each of its expectations is (in this case, it's a balance of the two constraints as determined by the mathematical implementation of the Bayesian strategy:
fewer words vs. shorter words).

"tothecastlebeyondthegoblincity"

Find the best segmentation that balances these proto-lexicon preferences and can generate the observable fluent speech utterances

Hypothesis 1	Hypothesis 2
✓	✓
fewer words	**shorter words**

Bayesian inference for speech segmentation

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

"tothecastlebeyondthegoblincity"

Find the best segmentation that balances these proto-lexicon preferences and can generate the observable fluent speech utterances

There will be some probability the Bayesian learner assigns to each hypothesis, based on this balance.

Hypothesis 1	Hypothesis 2
✓	✓
fewer words	**shorter words**
0.6	**0.4**

Bayesian inference for speech segmentation

Bayesian inference

$$P(s|u) \propto P(s)P(u|s)$$

“tothecastlebeyondthegoblincity”

Find the best segmentation that balances these proto-lexicon preferences and can generate the observable fluent speech utterances

The most probable hypothesis will be the one the learner chooses.

Hypothesis 1	Hypothesis 2
✓	
0.6	**0.4**

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

✓ **Is it useful?**

Computational-level modeled learners using this strategy segment fairly well, given realistic English child-directed speech data.

Best performance by a Bayesian learner on realistic English child-directed speech data had an **F-score of 86.3%**.

This is much better than what we found for a learner that hypothesizes a boundary at a transitional probability minimum (F-score = 29.9%). Statistical learning by itself isn't always so bad after all!

Phillips & Pearl 2012, 2014a, 2014b, 2015a, 2015b, Pearl & Phillips in press

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

Is it useful?

Is it useable?

Algorithmic-level modeled learners with cognitive constraints on their inference and memory can still use this strategy and segment English quite well.

Phillips & Pearl 2012, 2014a, 2014b, 2015a, 2015b, Pearl & Phillips in press

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

✓ **Is it useful?**

✓ **Is it useable?**

✓ **Does it work for different languages?**

It segments well for languages with different morphology and syllable properties: Spanish, Italian, German, Hungarian, Japanese, Farsi

Phillips & Pearl 2012, 2014a, 2014b, 2015a, 2015b, Pearl & Phillips in press

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

✓ **Is it useful?**

✓ **Is it useable?**

✓ **Does it work for different languages?**

Question: If we're modeling the speech segmentation occurring at 5-6 months, do we expect perfect adult-like segmentation?

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

✓ **Is it useful?**

✓ **Is it useable?**

✓ **Does it work for different languages?**

Hmmm…probably not, given the segmentation errors that persist even once children are able to speak.

"Two dults"
[Two adults]

"I am being have!"
[I am behaving!]
(in response to "Behave!")

"A B C D E F G, H I J K, elemenopi…"
[A B C D E F G, H I J K, L M N O P…

"Yeah, she was hiccing-up."
[hiccup = hicc + up]

"I don't want to go to your ami!"
[I don't want to go to Miami]

"Oh say can you see by the donzerly light?"
[Oh say can you see by the dawn's early light?]

Is it **useful**?

Is it **useable**?

Does it work for **different languages**?

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

Important point: What a six-month-old thinks are useful units to segment out of fluent speech may not match what we adults think of as words.

Example: "See the kitty playing with the string."

✓ **Is it useful?**

✓ **Is it useable?**

✓ **Does it work for different languages?**

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

Important point: What a six-month-old thinks are useful units to segment out of fluent speech may not match what we adults think of as words.

Example: "See the kitty playing with the string."

Useful unit smaller than a word:

-ing = ongoing action

Oversegmentation (split words up):

playing = play ing

✓ Is it useful?

✓ Is it useable?

✓ Does it work for different languages?

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

Important point: What a six-month-old thinks are useful units to segment out of fluent speech may not match what we adults think of as words.

Example: "See the kitty playing with the string."

Useful unit larger than a word:

thekitty = maps to specific concrete object

Undersegmentation (squish words together):

the kitty = thekitty

✓ Is it **useful**?

✓ Is it **useable**?

✓ **Does it work for different languages?**

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

Let's see this in action.

"See the kitty playing with the string."

Suppose we allow the following to count as useful units, even though they're technically missegmentations:

✓ *ing*, *thekitty*

✓ **Is it useful?**

✓ **Is it useable?**

✓ **Does it work for different languages?**

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

Let's see this in action.

Comparison "See thekitty play ing with the string." ✓ *ing*, *thekitty*

Suppose this is the segmentation the learner had:

See thekitty play ing withthe string

Recall:

5 "true words": see, thekitty, play, ing, string

7 words should have found: see, thekitty, play, ing, with, the, string

= 5/7 = **0.71**

✓ **Is it useful?**

✓ **Is it useable?**

✓ **Does it work for different languages?**

Bayesian inference for speech segmentation

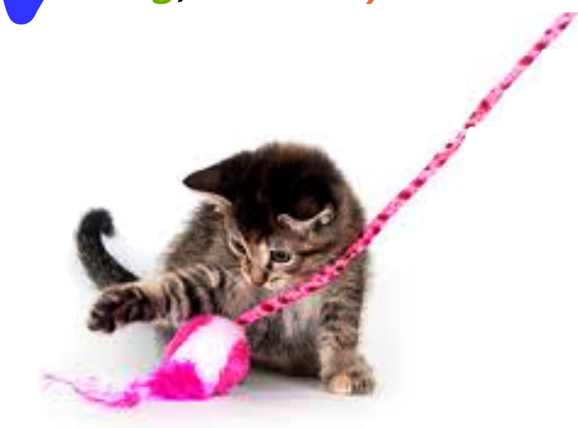
$$P(s|u) \propto P(s)P(u|s)$$

Let's see this in action.

Comparison "See thekitty play ing with the string." ✓ *ing*, *thekitty*

Suppose this is the segmentation the learner had:

See thekitty play ing withthe string



Precision:

5 "true words": see, thekitty, play, ing, string

6 "words" found: see, thekitty, play, ing, withthe, string

= 5/6 = **0.83**

✓ **Is it useful?**

✓ **Is it useable?**

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

When we count these “**useful units**” as reasonable segmentation output for a seven-month-old, Bayesian learners do really well cross-linguistically (Phillips & Pearl 2014b, Phillips & Pearl 2015, Pearl & Phillips in press): F-score: 77.4%. Again, this suggests Bayesian inference may work quite well as a statistical strategy in the absence of other cues.

Does it work for different languages?

✓

Is it **useful**?

Is it **useable**?

Does it work for **different languages**?

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

Also, the inferred units seem to be quite useful in practice — these units allow children to infer the correct stress-based cue for their language from the inferred proto-lexicons.

	Age (in months) 5	6	7	8	9	10	11
Cue use							
statistical	✓[1]			✓[2,3]			✓[4]
lexical		✓[5]		✓[6,7]			✓[7]
metrical			✓[8]	✓[9]	✓[8]		
Cue weighting							
statistical *vs.*	✓[1] (metrical)		✓[12] (metrical)	✖[2] (metrical; coarticulation)	✖[12] (metrical)		✖[4] (metrical)
metrical *vs.*	✖[1] (statistical)		✖[12] (statistical)	✓[2] (statistical)	✓[12,13] (statistical; phonotactics)		✓[4] (statistical)

References: 1) Thiessen & Erikson, 2013; 2) Johnson & Jusczyk, 2001; 3) Saffran et al., 1996; 4) Johnson & Seidl, 2009; 5) Bortfeld et al., 2005; 6) Shi & Lepage, 2008; 7) Shi et al., 2006; 8) Curtin et al. 2005; 9) Jusczyk, Houston, et al., 1999; 10) Mattys & Jusczyk, 2001; 11) Jusczyk, Hohne et al., 1999; 12) Thiessen & Saffran, 2003; 13) Mattys et al. 1999.

Is it useful?

Is it useable?

Does it work for different languages?

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

Also, the inferred units seem to be quite useful in practice — these units also allow more successful word-meaning mapping.

thekitty

Bayesian inference for speech segmentation

$$P(s|u) \propto P(s)P(u|s)$$

✓ **Is it useful?**

✓ **Is it useable?**

✓ **Does it work for different languages?**

This kind of Bayesian inference seems to be a good proposal for a very early speech segmentation strategy that depends on statistical cues.

Statistical learning for segmentation

Gambell & Yang (2006) found that the statistical learning strategy of positing word boundaries at transitional probability minima failed on realistic child-directed speech data.

More recent studies found that more sophisticated statistical learning -- Bayesian inference -- did much better on realistic child-directed speech data, suggesting that children may be able to use statistical learning to help them with segmentation - even before they use other strategies like lexical stress.

Statistical learning for segmentation

Notably, Bayesian inference learning strategies can work for learning to segment a variety of languages, especially if we recognize that an infant's segmentation may not perfectly match an adult's segmentation.

Questions?

You should be able to do all the questions on HW3 and all of the speech segmentation review questions.