Psych 156A/ Ling 150: Acquisition of Language II

Lecture 7

Word Segmentation II

Announcements

Be working on HW2

Be working on word segmentation review questions

Midterm on Tuesday, 5/6/14

Midterm Review on Thursday, 5/1/14

Computational problem

Divide fluent speech into individual words

tuðəkǽsəlbijándðəgáblɪnsíɾi

tu ðə kǽsəl bijánd ðə gáblɪn síɾi

to the castle beyond the goblin city

How do we study this?

Experiments: Test infant abilities at different ages. See what they can and cannot do. But we have to guess (to some degree) how they manage to accomplish this.

Computational model: a program that simulates the mental processes occurring in a child's mind. This requires knowing what the input and output are, and then testing the strategies that can take the given input and transform it into the desired output.

Computational modeling
(Working with "digital children")

For example, in word segmentation, the input could be a sequence of syllables and the desired output is words (groups of syllables).

Input: "un der stand my po si tion"
Desired Output: "understand my position"

What is Computational Modeling?

In its essence, it's just a set of ***mathematical equations*** used to describe some process.

So, someone comes up to you and says "prove the Earth orbits the sun". Well, in order to prove that you need to collect some data, but you also need to know what you expect the data to look like.

The set of equations you came up with is a ***model*** of how the Earth orbits the sun.

Computational models are just like this, except they tend to be very complex and require computer simulations in order to answer.

Why use Computational Modeling?

For one, there are lots of practical applications:

- Weather forecasting
- Molecular protein folding simulations
- Netflix prediction

But models are also useful in psychology because another process they can model is ***learning***.

So if some theorist says "My experiment shows that children can use transitional probabilities (TPs) to segment words."

We can model that and see if it would be ***useful*** for children to use TPs in the real world.

Making Computational Modeling Useful

Just because our model works, doesn't make it useful:

If your phone can segment your fluent speech, does that mean it does it the same way an infant does? Probably not.

The closer we simulate how we think infants learn, the more useful our model is.

Simulating infants

- Use the kind of language input that infants receive
- Evaluate against how infants, not adults, perform
- Use learning strategies we know infants are capable of using

How does this relate to Word Segmentation?

We want to see if Transitional Probabilities are useful for actual infants

Gambell & Yang (2006): Computational model goals

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, and it's not clear how well the results they found will map to real 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.)

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*)

ðəbígbǽdwʌ́lf
↓
ðə bíg bǽd wʌ́lf
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*)

ðəbígbǽdwʌ́lf
↓
ðə bíg bǽd wʌ́lf
the big bad wolf

Recall calculation:
# of real words found / # of actual words
Identified 4 real words: the, big, bad, wolf
Should have identified 4 words: the, big, bad, wolf
Recall Score: 4 words found/4 should have found = 1.0

How do we measure word segmentation performance?

Ideally:
Compare our results versus what children actually segment

But…
Its really hard to test how a child segmented an utterance outside of a very controlled experiment.
Even if you could, testing an entire corpus (10,000s or 100,000s of utterances) might take years of data collection.

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*)

ðəbígbǽdwʌ́lf
↓
ðə bíg bǽd wʌ́lf
the big bad wolf

Precision calculation:
\# of real words found / # of words guessed
Identified 4 real words: the, big, bad, wolf
Identified 4 words total: the, big, bad, wolf
Precision Score: 4 real words found/4 words found= 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*)

ðəbígbǽdwʌ́lf
↓
Error ðəbíg bǽd wʌ́lf
thebig 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*)

ðəbígbǽdwʌ́lf
↓
Error ðəbíg bǽd wʌ́lf
thebig bad wolf

Recall calculation:
Identified 2 real words: bad, wolf
Should have identified 4 words: the, big, bad, wolf
Recall Score: 2 real words found/4 should have found = 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*)

ðəbígbǽdwʌ́lf
↓
Error ðəbíg bǽd wʌ́lf
thebig bad wolf

Precision calculation:
Identified 2 real words: bad, wolf
Identified 3 words total: thebig, bad, wolf
Precision Score: 2 real words/3 words identified = 0.666…

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 in order to be sure that word segmentation is really successful

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 word segmentation performance?

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

Perfect word 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 word segmentation performance?

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

Not-so-perfect word segmentation

Recall = 50% (0.50)
Precision = 67% (0.67)
F-score = 2*(0.50 * 0.67)/(0.50 + 0.67) = 0.57

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

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 (usually parents interacting with their children) transcribed by researchers. Used to see what children's input is actually like.

CHILDES Child Language Data Exchange System

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

The CMU Pronouncing Dictionary

Where does the realistic data come from?

Converting transcriptions to pronunciations

- **Look up words or a sentence (v. 0.7a)**

Look Up

☑ Show Lexical Stress

- the big bad wolf
- DH AH0 . B IH1 G . B AE1 D . W UH1 L F .

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

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

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	bǽd	wʌ́lf
DH AH0	B IH1 G	B AE1 D	W UH1 L F

"There is a word boundary AB and CD if

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

Transitional probability minimum

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.

Desired word segmentation

ðə	bíg	bǽd	wʌ́lf
DH AH0	B IH1 G	B AE1 D	W UH1 L F
the	big	bad	wolf

Modeling results for transitional probability

Precision: 41.6%

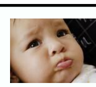
Recall: 23.3%

F-score: 29.9%

A learner relying only on transitional probability 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).

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 surrounding the sequences]..." - Gambell & Yang (2006)

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 surrounding the sequences]..." - Gambell & Yang (2006)

ðə bíg bǽd wʌ́lf

TrProb1 TrProb2 TrProb3

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 surrounding the sequences]…" - Gambell & Yang (2006)

ðə bíg bǽd wʌ́lf

0.6 0.3 0.7

Why such poor performance?

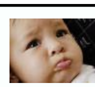
"We were surprised by the low level of performance. Upon close examination of the learning data, however, it is not difficult to understand the reason….a sequence of monosyllabic words requires a word boundary after each syllable; a [transitional probability] learner, on the other hand, will only place a word boundary between two sequences of syllables for which the [transitional probabilities] within [those sequences] are higher than [those surrounding the sequences]…" - Gambell & Yang (2006)

ðə bíg bǽd wʌ́lf

0.6 0.3 0.7

$0.6 > 0.3 < 0.7$

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 surrounding the sequences]…" - Gambell & Yang (2006)

learner posits one word boundary at minimum TrProb

ðə bíg | bǽd wʌ́lf

0.6 0.3 0.7

$0.6 > 0.3$, $0.3 < 0.7$

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 surrounding the sequences]…" - Gambell & Yang (2006)

…but nowhere else

ðə bíg | bǽd wʌ́lf

0.6 0.3 0.7

$0.6 > 0.3$, $0.3 < 0.7$

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 surrounding the sequences]…" - Gambell & Yang (2006)

…but nowhere else

ðə bíg | bǽd wʌ́lf

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 surrounding the sequences]…" - Gambell & Yang (2006)

…but nowhere else

ðəbíg | bǽdwʌ́lf

thebig | badwolf

Precision for this sequence: 0 words correct out of 2 found

Recall: 0 words correct out of 4 that should have been found

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, [a transitional probability learner] cannot work." - 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.

Hypothesis: Unique Stress Constraint (USC)

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

no stress	stress	stress	stress
ðə	bíg	bǽd	wʌ́lf
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 word segmentation problem.

Hypothesis: Unique Stress Constraint (USC)

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

ðə bíg bǽd wʌ́lf

the big bad wolf

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.

Hypothesis: Unique Stress Constraint (USC)

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

húz ə fréd əv ðə bíg bǽd wʌ́lf

who's a fraid of the big bad wolf

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.

Hypothesis: Unique Stress Constraint (USC)

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

húz ə fréd əv ðə bíg bǽd wʌ́lf

who's a fraid of the big bad wolf

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.

Hypothesis: Unique Stress Constraint (USC)

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

? ?

húz ə fréd əv ðə bíg bǽd wʌ́lf

who's a fraid of the big bad wolf

There's a word boundary at one of these two.

USC + Transitional Probabilities

Precision: 73.5%

Recall: 71.2%

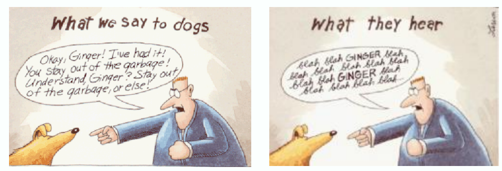
F-score: 72.3%

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 more formal version 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!"

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)

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

Using algebraic learning + USC

WeakSyl	StrongSyl	StrongSyl	StrongSyl
the	big	bad	wolf
ðə	bíg	bǽd	wʌ́lf

"the big bad wolf"

Using algebraic learning + USC

Familiar word: "the" (algebraic learning)

WeakSyl	StrongSyl	StrongSyl	StrongSyl
the	big	bad	wolf
ðə	bíg	bǽd	wʌ́lf

"the big bad wolf"

Using algebraic learning + USC

USC says these must be separate words

WeakSyl	StrongSyl	StrongSyl	StrongSyl
the	big	bad	wolf
ðə	bíg	bǽd	wʌ́lf

"the big bad wolf"

Using algebraic learning + USC

Correct segmentation!

WeakSyl	StrongSyl	StrongSyl	StrongSyl
the	big	bad	wolf
ðə	bíg	bǽd	wʌ́lf

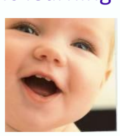
"the big bad wolf"

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.

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

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

Gambell & Yang 2006 critiques

Do children have access to the Unique Stress Constraint (USC)?

- Children definitely use transitional probabilities & algebraic learning – but how precise is their knowledge of lexical stress?

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.)

✗	✓
píma vs. latú	píma vs. pimá

Gambell & Yang 2006 critiques

Does dictionary stress really match actual stress patterns?

Gambell & Yang estimate: the *bíg bád wólf*
Typical speech: the big bad *wólf*

It's unclear how well this algorithm works with real stress patterns in fluent speech…

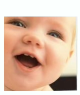
Actually, USC works very poorly unless you also add algebraic segmentation (Lignos 2011):

F-score = **31.2** (USC alone) = **92.9** (USC + Algebraic)

But Algebraic alone is almost as good = **90.4** (Algebraic)

More sophisticated learning strategies

What if children are capable of tracking more sophisticated distributional information (that is, they're not just restricted to transitional probability minima)? In that case, how well do they do on realistic data, if all they're using is statistical learning (no stress information)?



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)

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

Bayesian inference for word segmentation

What kind of hypotheses might a child have for word segmentation?

Observed data:
"to the ca stle be yond the go blin ci ty"

Hypothesis = sequence of vocabulary items producing this observable data

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 used twice*

Bayesian model

Learner expectations about word segmentation:
(1) Words tend to be shorter rather than longer
(2) Vocabulary tends to be small rather than large

Used by these research studies:
Goldwater, Griffiths, & Johnson 2009
Pearl, Goldwater, & Steyvers 2011
Phillips & Pearl 2012, Phillips & Pearl 2014

Bayesian model

Learner expectations about word segmentation:
(1) Words tend to be shorter rather than longer
(2) Vocabulary tends to be small rather than large

How would a Bayesian learner with these kind of expectations decide between the two hypotheses from before?

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

How long are words? Between 2 and 3 syllables, average = 2.2
How large is the vocabulary? 5 words

Bayesian model

Learner expectations about word segmentation:
(1) Words tend to be shorter rather than longer
(2) Vocabulary tends to be small rather than large

How would a Bayesian learner with these kind of expectations decide between the two hypotheses from before?

Hypothesis 2:
"to the castle beyond the goblin city"
Items: *to, the, castle, beyond, goblin, city*

How long are words? Between 1 and 2 syllables, average = 1.7
How large is the vocabulary? 6 words

Bayesian model

Comparing hypotheses - which is most likely?

Hypothesis 1: longer words, but fewer words
How long are words? Avg = 2.2 syllables
How large is the vocabulary? 5 words

Hypothesis 2: shorter words, but more words
How long are words? Avg = 1.7 syllables
How large is the vocabulary? 6 words

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: fewer words vs. shorter words).

Bayesian model

Comparing hypotheses - which is most likely?

Hypothesis 1: longer words, but fewer words
How long are words? Avg = 2.2 syllables
How large is the vocabulary? 5 words

Hypothesis 2: shorter words, but more words
How long are words? Avg = 1.7 syllables
How large is the vocabulary? 6 words

There will be some probability the Bayesian learner assigns to each hypothesis. The most probable hypothesis will be the one the learner chooses.

Bayesian model

Comparing hypotheses - which is most likely?

Hypothesis 1: longer words, but fewer words
How long are words? Avg = 2.2 syllables
How large is the vocabulary? 5 words

Probability: 0.33

Hypothesis 2: shorter words, but more words
How long are words? Avg = 1.7 syllables
How large is the vocabulary? 6 words

Probability: 0.67

There will be some probability the Bayesian learner assigns to each hypothesis. The most probable hypothesis will be the one the learner chooses.

Bayesian model

Comparing hypotheses - which is most likely?

Hypothesis 1: longer words, but fewer words
How long are words? Avg = 2.2 syllables
How large is the vocabulary? 5 words

Probability: 0.33

Hypothesis 2: shorter words, but more words
How long are words? Avg = 1.7 syllables
How large is the vocabulary? 6 words

Probability: 0.67

There will be some probability the Bayesian learner assigns to each hypothesis. The most probable hypothesis will be the one the learner chooses.

Bayesian Word Segmentation

- Go through each boundary, *b*, in the corpus
  - Do we prefer two small words or one large word
- Fewer word types
  - $P(w_i = w | w_1 \ldots w_{i-1}) = \frac{n_{i-1}(w) + \alpha P_0(w)}{i-1+\alpha}$
- Shorter word forms
  - $P_0(w = x_1 \ldots x_m) = \prod_{j=1}^{m} P(x_j)$
- Bigram
  - $P(w_i = w | w_{i-1} = w', w_1 \ldots w_{i-2}) = \frac{n_{i-1}(w', w) + \beta P_1(w)}{n_{i-1}(w') + \beta}$

Realistic Bayesian learners

Phillips and Pearl 2012 tested their Bayesian learners on realistic data: 28,391 utterances of child-directed speech from the Brent corpus in CHILDES. (Average utterance length: 3.4 words and 4.2 syllables)

Best performance by a Bayesian learner:

F-score: 86.3%

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

Statistical learning for word segmentation

Saffran et al. (1996) found that human infants are capable of tracking transitional probability between syllables and using that information to accomplish word segmentation in an artificial language.

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

More recent studies (Goldwater et al. 2009, Pearl et al. 2011, Phillips & Pearl 2012) 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 word segmentation - even before they use other strategies like lexical stress.

The power of computational modeling

Computational Modeling shows us what kinds of information are useful

But it also shows ***how*** using that information is useful

You might have all the pieces to a puzzle, but unless you know the rules for putting them together, you can't solve it

It also makes predictions!
e.g. The Bayesian models here predict:

- Infants should undersegment (i.e. "what's that" -> "what'sthat")
- Infants should oversegment morphology (i.e. "doing" -> "do ing")

Speaking of predictions

So, these models work on English, but what about other languages?

This is ***really*** important to the theory!

If Transitional Probabilities only work for English, then they can't be the basis of word segmentation.

But even though the experimental psychologists are ***very*** interested…

They can't test this themselves. It can only be done through computational modeling.

Cross-linguistic Word Segmentation

So, these models work on English, but what about other languages?

	English	German	Spanish	Italian	Farsi	Hungarian	Japanese
Random	38.2	34.2	28.9	22.9	21.4	25.7	23.8

Cross-linguistic Word Segmentation

So, these models work on English, but what about other languages?

	English	German	Spanish	Italian	Farsi	Hungarian	Japanese
Random	38.2	34.2	28.9	22.9	21.4	25.7	23.8
Algebraic (no stress)	87.8	82.4	58.3	40.0	35.1	49.8	30.1

Cross-linguistic Word Segmentation

So, these models work on English, but what about other languages?

	English	German	Spanish	Italian	Farsi	Hungarian	Japanese
Random	38.2	34.2	28.9	22.9	21.4	25.7	23.8
Algebraic (no stress)	87.8	82.4	58.3	40.0	35.1	49.8	30.1
Bayesian	77.1	73.1	64.8	71.3	69.6	66.2	66.5

To wrap up

Computational modeling is a tool that can aid psychologists by explicitly testing theories in ways that experiments can't.

Models also make predictions which can lead to ideas for new experiments!

But modeling has downsides too, by ***abstracting away*** from real children, we may lose something that's important to the learning process.

The closer we approximate the learning process (i.e. the more ***psychologically plausible*** our model), the better we can trust our model's conclusions.

Questions?

You should be able to do up through question 6 on HW2 and all of the word segmentation review questions.