Word Segmentation: Quick but not Dirty

Timothy Gambell and
Charles Yang

What's the Problem?

- One task that language learners have to master is to figure out where word boundaries are in speech.
  - In normal conversation, we don't pause between words, it's one continuous flow of speech. The infant has to somehow figure out the words of its language- Infants as young as 7.5 months have already begun to do this. We're going to explore how they could do this- not just how an ideal learner could, in principle segment words, but we're going to think about how actual kids can do this.

Outline of the paper

- Strategies for Word Segmentation
  - Previous suggestions of word segmentation strategies
- Statistical Learning is Ineffective
  - Where we show simple statistical learning doesn't fare as well as it should.
- Segmentation under Linguistic Constraints
  - Where we see an alternative to pure statistical learning.

Previously Proposed Solutions

- Children could recognize isolated words
  - The suggestion is that children learn individual words when they hear them in isolation. They can then use these words to "bootstrap" their knowledge.
  - About 9% of utterances directed at children by their mothers are single word utterances.
  - But, how does the child recognize the single word utterances
    - by length seems unlikely, "I see" is shorter than "spaghetti"
    - Gambell and Yang want to know the mechanism by which children pick out single word utterances- they say none have been proposed in the literature, but their suggestion might help with this question

Previously Proposed Solutions

- Transitional Probabilities $\mathrm{TP}(A \rightarrow B) = \frac{\Pr(AB)}{\Pr(A)}$
  - The thought is that different syllables occur next to each other more frequently when they make part of a word than when they do not, and children can get information about word boundaries this way.
  - This method does not rely on previous knowledge of the particular language the child is learning- it's supposed to work well regardless of the language.
  - This seems to work well in the lab, but has yet to be tested in "the wild" (i.e. on actual language, rather than artificial language)- we're going to look at this later.

Transitional Probabilities

Saffran Aslin and Newport showed that children as young as 7 months can distinguish between words and non-words, and even words and part-words in an artificial language, after only 2 minutes of exposure

This may be a domain general learning device, infants have been shown to be able to pick out allowable sequences among non-linguistic sounds (tones) and even shapes.

Monkeys can even distinguish between words and non-words, using the same artificial language given to the infants in Saffran et al's study!

Previously Proposed Solutions

- Metrical Segmentation Strategy
  - In English, most (around 90%) of the words start out with a stressed syllable, the thought is that children could use this information to help with word segmentation.
  - 7.5 month old English speaking kids do better at recognizing words with strong/weak stress pattern than weak/strong stress pattern.
  - But, you need to know something about your language before you can implement this strategy.
  - Also, what about languages without such a predictable stress pattern?

Previously Proposed Solutions

- Phonotactic Constraints
  - Some strings of sounds could, in principle, be English words, and others could not. *Blanze* and *slan* could be words in English, but *kzit* and *vtalp* could not be English words. There are certain consonant combinations that are not allowable in English. The thought is that when the learner comes in contact with these consonant combinations, they can figure out a word boundary.
  - 9 month old infants have been shown to be sensitive to this constraint.
  - But, the unallowable consonant combination might just signify a boundary between syllables, and not a word boundary- "mb" in "embed"
  - This method also assumes some familiarity with the language.
  - This method might be most helpful in segmenting syllables.

Previously Proposed Solutions

- Allophonic and articulatory cues
  - Certain articulations of sounds can differ depending on whether or not the sound comes at the beginning of the word. In English the allophone /t/ serves as a good example, It is aspirated at the beginning of a word, and uaspirated at the end. For example: “tab” and “cat”
  - The thought is that children could use this sort of information to mark word boundaries.
  - However, this relies on much knowledge of the language, this is clearly not a way to “get off the ground” learning language.
  - 9 month olds can’t use this strategy, but 10.5 month olds can! (nitrates vs. night rates)

Previously Proposed Solutions

- Memory
  - Kids may learn sound patterns before they learn the meanings of words.
  - They might be able to learn new words by extracting sound patterns from phrases with some familiar sound patterns.

Previously Proposed Solutions

- These possible solutions are by no means mutually exclusive, learners could employ all or any of these strategies to help them learn language. But, there’s a problem, most of these possible solutions presupposes some familiarity with their language. One of the big puzzles is to figure out how we “get off the ground” with language acquisition- how is it that children get to know enough about their language to start using these tricks?
- We also would like to know how the learning strategies interact with one another, and how they work across languages

Modeling Word Segmentation: Preliminaries

- This paper aims to explore the psychologically plausible algorithms that children may actually use for word segmentation, it does not aim at just trying to examine the information available to the child.
- Both methods are important, but previous research has focused more on the possible available data, and not on psychologically plausible learning mechanisms
- Just knowing that certain statistical regularities exist in a corpora tells us nothing about whether that information can be extracted with psychologically plausible means, this paper wants to focus on just the methods kids could actually use.

Preliminaries

- Computational models of word segmentation, traditionally, have made assumptions about the learning process that are not well suited for understanding child language learning
  - Previous computational models often overestimate the computational capacity of human learners
  - And, they underestimate human's knowledge of linguistic representations- it seems that we get that syllables are the primary source that words are built up out of rather early- we don't seem to have to start with segments and move up to syllables, as some models assume.

Precision vs. Recall

- Recall
  - How many correct answers the learner gives out of how many total correct answers there are.

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

- Precision
  - How many correct answers the learner gives, out of how many answers they give

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

- F-measure

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

Previous Word Segmentation Results

- The highest performance was Brent (1999)
  - Precision and recall around 70%-80%
- Other models typically do worse
  - Precision and recall around 40%-50%

The Input

- For input for their model, they took a phonetic transcription of child directed speech from the CHILDES corpus, labeled for different sorts of stresses, with spaces between words, and punctuation removed.
  - "cat" becomes "K AE1 T"
  - "catapult" becomes "K AE1 T AH0 P AH0 L T"

Statistical Learning is Ineffective

- So, we’ve reviewed possible word segmentation strategies, we like them, in that they all seem psychologically plausible. It is also likely that they don’t all come on-line at once, but gradually. The transitional probabilities strategy looks most promising for getting us off the ground- that is, this strategy does not presuppose some knowledge about the learner’s language. So, we’ll look at a statistical learning method.

The Statistical Learning Model

- Training Stage
  - The learner gathers statistical information about transitional probabilities over all syllables in the data.
- Testing Stage
  - The model is tested on the same data it was trained on- unfair advantage to the model? Psychological plausibility?
  - A word boundary is posited at the point of local minima.

  So, if, for the syllable concatenation ABCD,

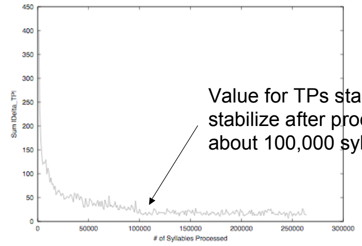
  TP(A→B) > TP(B→C) < TP(C→D)

  then a word boundary is posited between AB and CD

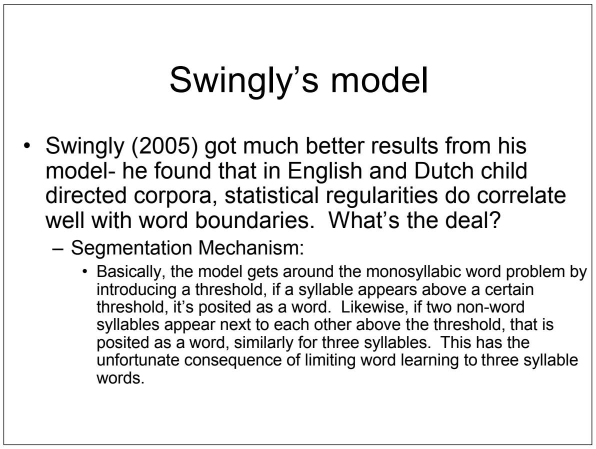
Results from this Model

- Precision = 41.6%
- Recall = 23.3%
  - Why isn’t this good enough?
    - I’m guessing that it’s because more than half the time, the learner guesses an incorrect word- how can the learner tell when they get a wrong word?
- A possible reason why these results are so poor: We need words with multiple syllables for this strategy to be effective, but most words in child directed speech are monosyllabic
  - The learning data consisted of 226,178 words, and 263,660 syllables. A monosyllabic word is followed by another monosyllabic word 85% of the time.

- Further, even giving the learner more training data is unlikely to help, we find that after the model has processed about 1000,000 sylllables, the total change in the values for the TPs is considerably reduced, and doesn’t change much.

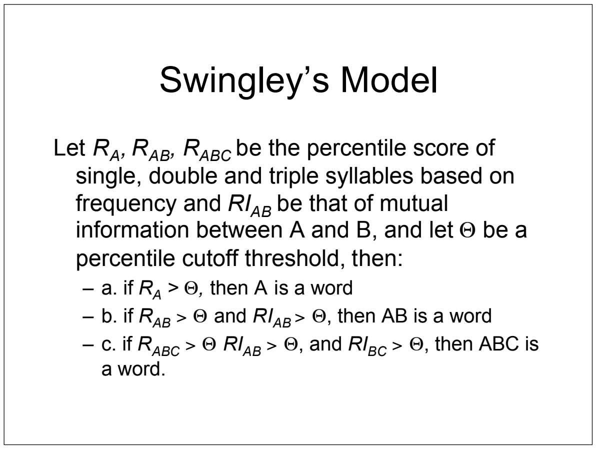


Value for TPs starts to stabilize after processing about 100,000 syllables



Swingly's model

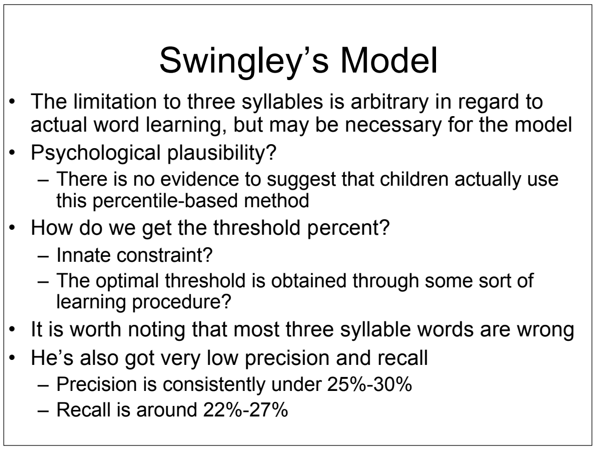
- Swingly (2005) got much better results from his model- he found that in English and Dutch child directed corpora, statistical regularities do correlate well with word boundaries. What's the deal?
  - Segmentation Mechanism:
    - Basically, the model gets around the monosyllabic word problem by introducing a threshold, if a syllable appears above a certain threshold, it's posited as a word. Likewise, if two non-word syllables appear next to each other above the threshold, that is posited as a word, similarly for three syllables. This has the unfortunate consequence of limiting word learning to three syllable words.



Swingley's Model

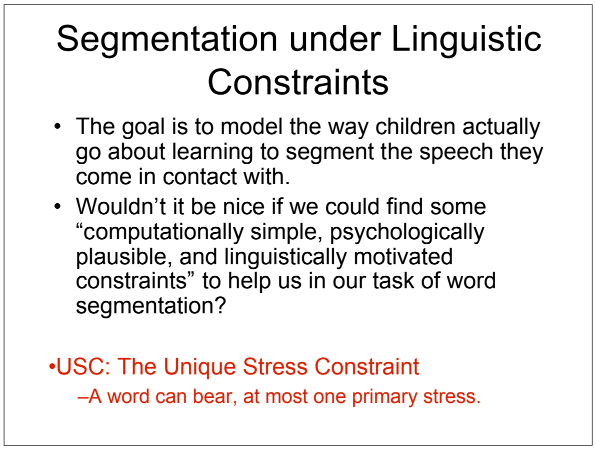
Let R_A, R_{AB}, R_{ABC} be the percentile score of single, double and triple syllables based on frequency and RI_{AB} be that of mutual information between A and B, and let Θ be a percentile cutoff threshold, then:

- a. if $R_A > \Theta$, then A is a word
- b. if $R_{AB} > \Theta$ and $RI_{AB} > \Theta$, then AB is a word
- c. if $R_{ABC} > \Theta$ $RI_{AB} > \Theta$, and $RI_{BC} > \Theta$, then ABC is a word.



Swingley's Model

- The limitation to three syllables is arbitrary in regard to actual word learning, but may be necessary for the model
- Psychological plausibility?
  - There is no evidence to suggest that children actually use this percentile-based method
- How do we get the threshold percent?
  - Innate constraint?
  - The optimal threshold is obtained through some sort of learning procedure?
- It is worth noting that most three syllable words are wrong
- He's also got very low precision and recall
  - Precision is consistently under 25%-30%
  - Recall is around 22%-27%

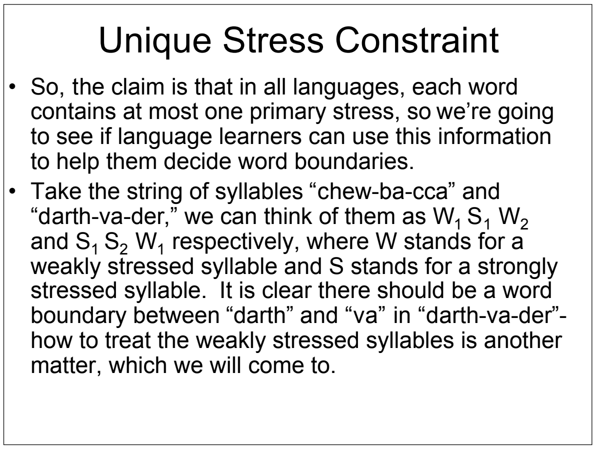


Segmentation under Linguistic Constraints

- The goal is to model the way children actually go about learning to segment the speech they come in contact with.
- Wouldn't it be nice if we could find some "computationally simple, psychologically plausible, and linguistically motivated constraints" to help us in our task of word segmentation?

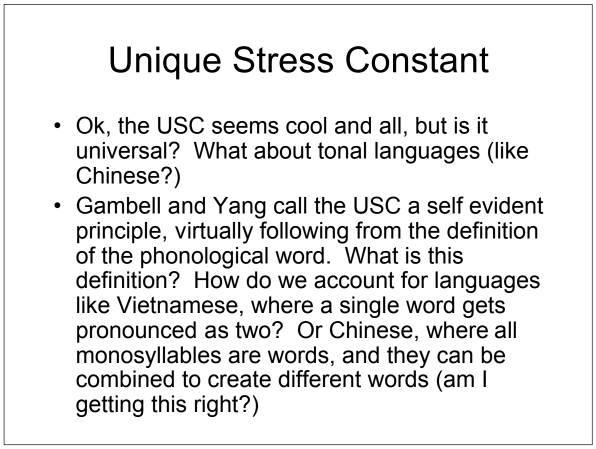
•USC: The Unique Stress Constraint

–A word can bear, at most one primary stress.



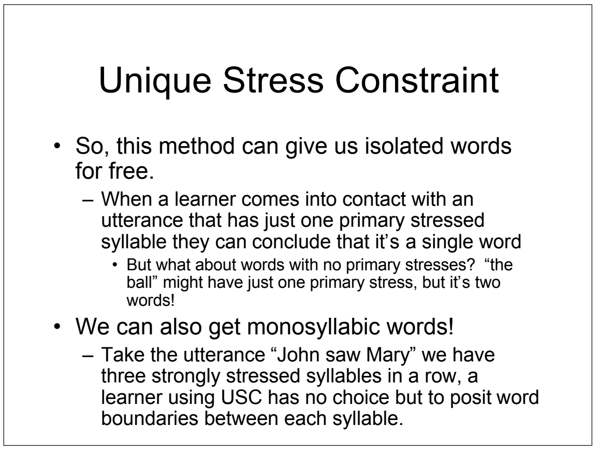
Unique Stress Constraint

- So, the claim is that in all languages, each word contains at most one primary stress, so we're going to see if language learners can use this information to help them decide word boundaries.
- Take the string of syllables "chew-ba-cca" and "darth-va-der," we can think of them as W_1 S_1 W_2 and S_1 S_2 W_1 respectively, where W stands for a weakly stressed syllable and S stands for a strongly stressed syllable. It is clear there should be a word boundary between "darth" and "va" in "darth-va-der"- how to treat the weakly stressed syllables is another matter, which we will come to.



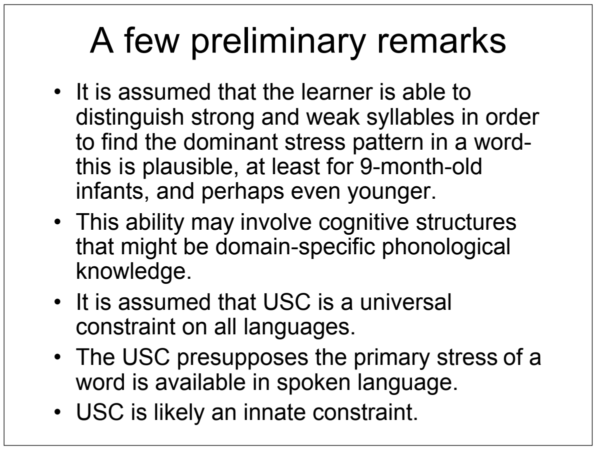
Unique Stress Constant

- Ok, the USC seems cool and all, but is it universal? What about tonal languages (like Chinese?)
- Gambell and Yang call the USC a self evident principle, virtually following from the definition of the phonological word. What is this definition? How do we account for languages like Vietnamese, where a single word gets pronounced as two? Or Chinese, where all monosyllables are words, and they can be combined to create different words (am I getting this right?)



Unique Stress Constraint

- So, this method can give us isolated words for free.
  - When a learner comes into contact with an utterance that has just one primary stressed syllable they can conclude that it's a single word
    - But what about words with no primary stresses? "the ball" might have just one primary stress, but it's two words!
- We can also get monosyllabic words!
  - Take the utterance "John saw Mary" we have three strongly stressed syllables in a row, a learner using USC has no choice but to posit word boundaries between each syllable.



A few preliminary remarks

- It is assumed that the learner is able to distinguish strong and weak syllables in order to find the dominant stress pattern in a word- this is plausible, at least for 9-month-old infants, and perhaps even younger.
- This ability may involve cognitive structures that might be domain-specific phonological knowledge.
- It is assumed that USC is a universal constraint on all languages.
- The USC presupposes the primary stress of a word is available in spoken language.
- USC is likely an innate constraint.

The Models

- Statistical Learning with USC
- Algebraic Learners
  - Agnostic Learner
  - Random Learner

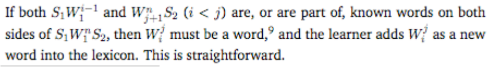
Statistical Learner with USC

- Training stage just like SL
- Testing stage:
  - Learner scans sequence of input syllables, and
    - If two strong syllables are adjacent, a word boundary is posited between them
    - If there are more than one weak syllables between two strong ones, then a word boundary is posited where the pairwise TP is lowest.
  - Results on this model: Precision=73.5%, and Recall=71.2%- comparable to the best results in the literature.

Algebraic Learning

- Motivation: Calculating transitional probabilities is HARD!
  - With each new utterance of a syllable, the learner has to readjust potentially thousands of transitional probabilities.
  - We want to see if a simpler learning mechanism could get us the same results without having to use such a computationally expensive method.

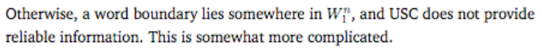
Algebraic Learning

- We consider two algebraic learners, an Agnostic learner and a Random learner. They both agree on this case:
  - 

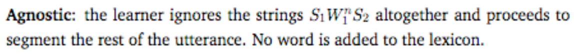
    If both $S_1W_1^{i-1}$ and $W_{j+1}^{n}S_2$ $(i < j)$ are, or are part of, known words on both sides of $S_1W_1^{n}S_2$, then W_i^j must be a word,[9] and the learner adds W_i^j as a new word into the lexicon. This is straightforward.

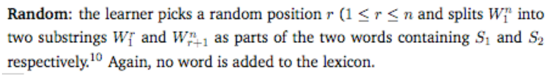
However, the case gets more complicated if we have to segment between unfamiliar words with weak syllables separating strong ones.

Algebraic Learning

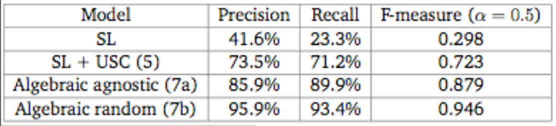

Otherwise, a word boundary lies somewhere in W_1^n, and USC does not provide reliable information. This is somewhat more complicated.

In this case, there are different ways to proceed, we will look at two.


Agnostic: the learner ignores the strings $S_1 W_1^n S_2$ altogether and proceeds to segment the rest of the utterance. No word is added to the lexicon.


Random: the learner picks a random position r ($1 \le r \le n$ and splits W_1^n into two substrings W_1^r and W_{r+1}^n as parts of the two words containing S_1 and S_2 respectively.[10] Again, no word is added to the lexicon.

Results



Model	Precision	Recall	F-measure ($\alpha = 0.5$)
SL	41.6%	23.3%	0.298
SL + USC (5)	73.5%	71.2%	0.723
Algebraic agnostic (7a)	85.9%	89.9%	0.879
Algebraic random (7b)	95.9%	93.4%	0.946

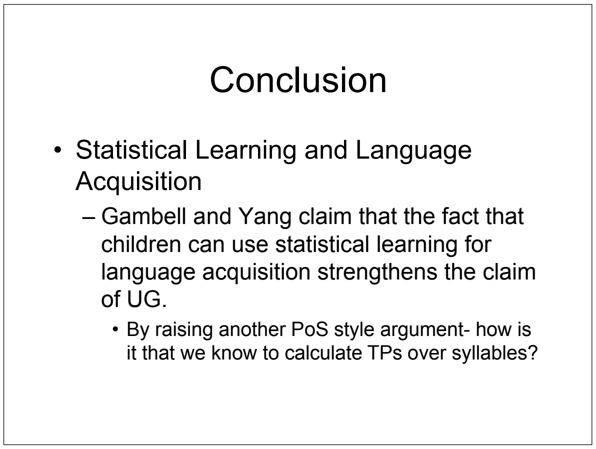
- So, the random algebraic learner does best overall. This, they claim is due to the learner guessing at words, rather than making no predictions at all, generally the words in the corpora are short, so the random learner gets a lot of correct guesses.
- They suggest that in actual word segmentation children might rely on language specific metrical segmentation

Conclusions

- The segmentation process can get off the ground only through the use of language independent means: experience-independent linguistic constraints such as USC and experience-dependent statistical learning are the only candidates among the proposed strategies for word segmentation. More of an assumption than a conclusion?
- Statistical learning does not scale up to realistic settings of language acquisition.
- Simple principles on phonological structures such as USC can constrain the applicability of statistical learning and improve its performance, though the computational cost of statistical learning may still be prohibitive.
- Algebraic learning under USC, which has trivial computational cost and is in principle universally applicable, outperforms all other segmentation models.

Conclusion

- Direction for Future Work
  - How do children learn to pick out statistical tendencies in speech?
    - This work helped clarify some logical issues, but we'd like to know how kids actually do this
  - How good are infants at identifying syllables in speech?
  - Algebraic learners might learn too quick
    - The algebraic models very rapidly learned to segment words, if we're interested in what kids actually do, we might want to take their gradual learning into account.
  - Different sorts of languages might make USC a bit more complicated.



Conclusion

- Statistical Learning and Language Acquisition
  - Gambell and Yang claim that the fact that children can use statistical learning for language acquisition strengthens the claim of UG.
    - By raising another PoS style argument- how is it that we know to calculate TPs over syllables?