



# How useful are transitional probabilities in adult-directed speech?

Kelly Enochson  
George Mason University

## Summary

**Background:** Infants are able to compute transitional probability (TP) and use this information to segment continuous speech into words (Saffran, Newport & Aslin 1996; Aslin, Saffran & Newport 1998).

**Problem:** Yang (2004, 2006) used an algorithm that segments speech based on TP to examine its usefulness on a corpus of child-directed speech (CDS). It was unsuccessful because of the abundance of monosyllabic words.

**Research Question:** Is the transitional probability information available in adult-directed speech more informative than in child-directed speech?

## Experiment 1

**Transitional Probability:**

$$TP(A \rightarrow B) = \frac{\Pr(AB)}{\Pr(A)}$$

**Method**

- Data come from Michigan Corpus of Academic Spoken English (MICASE) (Simpson, R. C., S. L. Briggs, J. Ovens, and J. M. Swales. 2002)
- Data transcribed using CMU Pronouncing Dictionary (Bartlett, Susan., Kondrak, Grzegorz., and Cherry, Colin. 2009)
- Maximize Onset
- Separate learning stage and testing stage
- TPs are computed over all the data
- Word boundaries are postulated at the points of *local minima* - where the transitional probability is lower than its neighbors.
- Yang used 226,178 words and 263,660 syllables (p.11)
- TPs stabilize after about 100,000 syllables (p.14)

**Experiment 1**

- 5 study groups
- Words = 113,607
- Syllables = 137,201

## Experiment 1 Results

**Performance measures**

*Precision* = how many of the postulated words are actual words

*Recall* = how many of the actual words are postulated as words

*F measure* ( $\alpha = .5$ )

$$F_{\alpha} = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$$

**Results**

*Yang's results*

Precision: 41.6%

Recall: 23.3%

**F measure: .299**

*Experiment 1 results*

Precision: 37.0%

Recall: 17.0%

**F measure: .233**

## Experiment 2

**Possible explanations for poor performance**

- ADS also contains many monosyllabic words
- TPs only work for 2- and 3-syllable words. Longer words also fail using local minima.
- CDS contains low type/token ratio: reduced number of word types, simplifying vocabulary (Soderstrom 2007)
  - Larger vocabulary of ADS potentially obfuscates statistical information.
  - ADS potentially requires larger input to achieve stable TPs.

## Experiment 2 Results

**Experiment 2**

7 study groups,  
2 advising sessions  
Words = 190,909  
Syllables = 228,336

**Results**

Precision: 37.6%  
Recall: 17.3%  
**F measure: .237**

**Contact Information**

kenochso@gmu.edu  
mason.gmu.edu/~kenochso

## Discussion

**Word Length:**

- These corpora are comprised of data that are 61% monosyllabic words.
- A monosyllabic word is followed by another monosyllabic word 77% of the time
- c.f. Yang's CDS corpus, where a monosyllabic word is followed by another monosyllabic word 85% of the time
- The corpora consist of 1.9% (Exp 1) and 1.7% (Exp 2) words that are more than 3 syllables long

**Stress Information**

Yang found that identifying words using only information about primary stress was more effective than TPs or a method that combines TPs and stress information.

*Yang's results - stress*

Precision: 81.5%

Recall: 90.1%

**F measure: .857**

*Experiment 2 results - stress*

Precision: 68.9%

Recall: 46.4%

**F measure: .554**

\*\* My data did not include utterance boundaries as a delimiter. That limits what the program can get "for free" from my corpus.

**Research question:** Is adult-directed speech more informative for language learners than child-directed speech, in terms of transitional probability?

**Answer:** No.

- F measure increases with an increase in corpus size, but only slightly.
- One-syllable and more-than-three-syllable words do not work with TP - local minima.
- Larger vocabulary in ADS potentially makes TP less informative
- ADS is not markedly different from CDS in terms of TP - local minima

## References

- Yang, Charles D. (2004) Universal grammar, statistics or both? *Trends in Cognitive Sciences*, 8(10):451-456.
- Yang, Charles D. and Gambell, T. (2006) Word segmentation: Quick but not dirty. Manuscript, Yale University.