How useful are Transitional Probabilities in Adult-Directed Speech?

Kelly Enochson
George Mason University

Laboratory studies have shown that infants (Aslin, Saffran & Newport 1998) and adults (Saffran, Newport & Aslin 1996) are able to use transitional probabilities (TPs) between syllables to segment continuous speech into words. Syllables occurring within a word have high TPs compared to syllables across word boundaries, therefore learners could in principle segment speech by positing potential word boundaries at local TP minima. However, Yang (2004, 2006) showed that in practice the abundance of monosyllabic words in CDS makes identifying local minima insufficient for identifying word boundaries; TPs are only informative for words with two syllables or more. Adult-directed speech (ADS) likely has fewer monosyllabic words, and therefore is potentially more informative than CDS, in terms of TPs, for infant language learners. If this is the case, then ADS would provide a rich source of statistical information for L2 learners as well. The ability to track statistical information like TPs has been argued to be a domain-general cognitive mechanism (Kelly & Martin, 1994). As such it could be particularly useful to adult language learners, who potentially lack some of the domain-specific language acquisition faculties that infants possess (DeKeyser 2000).

The current study investigates whether ADS is indeed more informative in terms of TPs among syllables than CDS. Experiment 1 examines five corpora from the Michigan Corpus of Academic Spoken English (MICASE, Simpson, et al., 2002), using the algorithm developed in Yang (2004, 2006) for segmenting speech using TPs. Words are transcribed using the CMU pronouncing dictionary (Bartlett, et al., 2009). Yang gathered data from over 200,000 syllables, and found that transitional probabilities stabilized after about 100,000 syllables. For this reason, we limit our study to just over 100,000 syllables; our data include 137,201 syllables. Performance was evaluated using the F measure, which weighs precision (how many of the postulated words are actual words) and recall (how many of the actual words are postulated as words). Strong performance is reflected by an F measure closer to 1, while poor performance is reflected by an F measure closer to 0. Results of the first experiment yielded an F measure of .233, suggesting that ADS is *not* a better source of data for word segmentation using TPs. In fact, our result was lower than the F measure found for CDS, .299 (Yang 2004,2006).

To explore this result further, we address a possible difference between CDS and ADS, namely that that adult-directed speech contains longer words and a higher type/token ratio (Soderstrom 2007). This could mean that TPs do not stabilize around 100,000 syllables, but rather at a larger number. In Experiment 2, we doubled the number of syllables analyzed, including nine corpora from MICASE, consisting of 228,336 syllables. Results were almost exactly the same as in Experiment 1, with an F measure of .237.

These results suggest that adult-directed speech is not more informative than child-directed speech in terms of TPs among syllables. The explanation for this may be that, perhaps counter-intuitively, ADS does not have appreciably more multi-syllabic words than CDS. In Yang's CDS corpus, a one-syllable word is followed by another one-syllable word 85% of the time. In our data, this number is only somewhat lower: 77%. Our data consist of 61% monosyllabic words, which are uninformative in terms of TP. Additionally, the larger type vocabulary in ADS, particularly notable in this corpus of academic speech, may obfuscate TP information by lowering the "high" TPs of syllables that co-occur within a word. The relatively

low informativity of TPs in spontaneous natural speech—both CDS and ADS—suggests the need for additional word-segmentations mechanisms both for L1 and L2 learning.

References

Aslin, Richard N., Saffran, Jenny R., and Newport, Elissa L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9(4):321-324,1998.

Bartlett, S., Kondrak, G., and Cherry, C. (2009). On the syllabification of phonemes. NAACL-HTL.

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition* 22, 499-533.

Kelly, M.H. and Martin, S. (1994). Domain-general abilities applied to domain-specific tasks: Sensitivity to probabilities in perception, cognition, and language. *Lingua*, 92, 105-140.

Saffran, Jenny R., Newport, Elissa L., and Aslin, Richard N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35(4):606-621, 1996.

Simpson, R. C., S. L. Briggs, J. Ovens, and J. M. Swales. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review* 27: 501-532.

Swingley, Daniel. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50, 86-132.

Yang, Charles D. Universal grammar, statistics or both? (2004). *Trends in Cognitive Sciences*, 8(10):451-456, 2004.

Yang, Charles D. and Gambell, T. (2006) Word segmentation: Quick but not dirty. Manuscript, Yale University.