Adaptation of maternal speech in statistical word segmentation of Korean

Mun, Seongmin[1] & Ko, Eon − Suk[2]
Department of English Language and Literature, Chosun university[1,2]

Keywords: word segmentation, child language acquisition, statistical learning mechanisms, Korean

For the basic task in first language acquisition, children must learn to segment words (i.e., word boundaries) from the speech of their caregiver (e.g., Bernard et al., 2020, Stärk et al., 2020). Furthermore, previous studies report that it is much easier to segment words by using child-directed speech (CDS) than adult-directed speech (ADS) (Fernald, 2000). In this regard, statistical approaches including transitional probability (TP; Saffran et al., 1996) are drawing attention to the understanding of statistical mechanisms of word segmentation in child language acquisition (Aslin et al., 1998).

At the beginning of our study, we pose three questions: (i) what the optimal algorithm is likely at work for segmentation in Korean, a language typologically different from the major Indo-European languages that have been investigated for this task, (ii) are distributional cues enhanced in some way in CDS compared to ADS? (iii) do the differences by age can affect the model performance? To answer these questions, we report a statistical simulation based on the model performance by using different algorithms and measures of TP with the manipulation of utterance types (i.e., CDS vs ADS) and different age groups ranging from zero to two.

For this purpose, as a linguistic resource, we used the Ko corpus in the CHILDES dataset (Ko et al., 2020; 149,395 syllable tokens for CDS and 24,746 syllable tokens for ADS) for CDS and ADS, with Call Friend Korean corpus for ADS (122,444 syllable tokens for ADS). In addition, to simulate a child's linguistic environment, we changed the written form of the corpus to a spoken form by using the phonological rules of Korean. For model training, we devised the syllable-based TP models by employing two algorithms (i.e., absolute, and relative) and two measures (i.e., Forward TP, and Backward TP). We then employed the k-fold cross-validation technique to obtain a normalized result from each model (Stone, 1974). We set the value of k as 10 and repeated the cross-validation 10 times, with each sub-sample used exactly once as the test set for the model training. Model performance was measured by comparing the word boundaries in the original input sentence with the word boundaries via each model.

We note three major findings of this current study. First, as shown in Figure 1, we found that the model by using FTP with a relative algorithm showed better performance than the other models. Second, we found that our model showed better performance when it trained by using CDS than ADS. This trend aligns with the advantages of CDS for word segmentation (Fernald, 2000). Third, as shown in Figure 2, when we tried to see the details by different age groups, we found that the CDS for the age 1 group showed better performance than the other age groups. Our TP model successfully demonstrates the ability to formulate statistical learning mechanisms of word segmentation for Korean. The success of our statistical learner adds to the cross-linguistic evidence for the effectiveness of statistical approaches in modelling child language acquisition.

References

Aslin, R.N., Saffran, J.R., & Newport, E.L. (1998). Computation of conditional probability statistics by 8-month-old infants. Psychological Science, 9 (4), 321–324. doi:10.1111/1467-9280.00063.

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G. R., Larsen, E., Johnson, M., Fibla, L., Dupoux, E., Daland, R., Cao, X. N., & Cristia, A. (2020). WordSeg: Standardizing unsupervised word form segmentation from text. Behavior Research Methods, 52(1), 264– 278.

Fernald, A. (2000). Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition. Phonetica,57(2-4), 242–254.

Ko, E.-S., Jo, J., On, K.-W. & Zhang, B.-T. (2020). Introducing the Ko Corpus of Korean mother-child interaction, Frontiers in Psychology. doi: 10.3389/fpsyg.2020.602623.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. Science, 274(5294), 1926-1928.

Stärk K, Kidd E, Frost RLA. (2020). Word segmentation cues in German child-directed speech: A corpus analysis. Language and Speech. doi: 10.1177/0023830920979016.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, 36(2):111–147.