Does Korean defeat phonotactic word segmentation?

Robert Daland
Department of Linguistics
University of California, Los Angeles
3125 Campbell Hall, Box 951543
Los Angeles, CA 90095-1543, USA
r.daland@gmail.com

Kie Zuraw
Department of Linguistics
University of California, Los Angeles
3125 Campbell Hall, Box 951543
Los Angeles, CA 90095-1543, USA
kie@ucla.edu

Abstract

Computational models of infant word segmentation have not been tested on a wide range of languages. This paper applies a phonotactic segmentation model to Korean. In contrast to the undersegmentation pattern previously found in English and Russian, the model exhibited more oversegmentation errors and more errors overall. Despite the high error rate, analysis suggested that lexical acquisition might not be problematic, provided that infants attend only to frequently segmented items.

1 Introduction

The process by which infants learn to parse the acoustic signal into word-sized units—word segmentation—is an active area of research in developmental psychology (Polka and Sundara 2012; Saffran et al. 1996) and cognitive modeling (Daland and Pierrehumbert 2011 [DP11], Goldwater et al. 2009 [GGJ09]). Word segmentation is a classic bootstrapping problem: to learn words, infants must segment the input, because around 90% of the novel word types they hear are never uttered in isolation (Aslin et al. 1996; van de Weijer 1998). However, in order to segment infants must know some words, or generalizations about the properties of words. How can infants form generalizations about words before learning words themselves?

1.1 DiBS

Two approaches in the literature might be termed *lexical* and *phonotactic*. Under the lexical approach, exemplified by GGJ09, infants are assumed to exploit the Zipfian distribution of language, identifying frequently recurring and mutually predictive sequences as words. In the phonotactic approach, infants are assumed to leverage universal and/or language-specific knowledge about the phonological *content* of sequences to infer the optimal segmentation. The present study focuses on the phonotactic approach outlined in DP11, termed DiBS. For other examples of approaches that use phonotactics, see Fleck 2008, Blanchard et al. 2010.

A (Di)phone-(B)ased (S)egmentation model consists of an inventory of segment-segment sequences, with an estimated probability that a word boundary falls between the two segments. For example, when [pd] occurs in English, the probability of an intervening word boundary is very high: $Pr(\# \mid [pd]) \approx 1$. These probabilities are the parameters of the model to be learned. In the supervised setting (*baseline* model), these parameters may be estimated directly from data in which the word boundaries are labeled: $Pr(\# \mid pd) = Fr(\# \wedge pd) / (Fr(\# \wedge pd) + Fr(\neg\# \wedge pd))$ where $Fr(\# \wedge pd)$ is the number of [pd] sequences separated by a word boundary, and $Fr(\neg\# \wedge pd)$ the number of [pd]'s not separated by a word boundary. For assessment purposes, these probabilities are converted to hard decisions.

DP11 describe an unsupervised learning algorithm for DiBS that exploits a positional independence assumption, treating phrase edges as a proxy for word edges (*phrasal* model). This learning model's performance on English is on par with state-of-the-art lexical models (GGJ09), reflecting the high positional informativeness of diphones in English. We apply the baseline and phrasal models to Korean.

1.2 Linguistic properties of Korean

Korean is unrelated to languages previously modeled (English, Dutch, French, Spanish, Ara-

bic, Greek, Russian), and it is an interesting test case for both phonotactic and lexical approaches.

Korean syntax and morphology (Sohn 1999) present a particular challenge for unsupervised learning. Most noun phrases are marked with a limited set of case suffixes, and clauses generally end in a verb, inflected with suffixes ending in a limited set of sounds ([a,ʌ,i,jo]). Thus, the phrase-final distribution may not reflect the overall word-final distribution—problematic for some phonotactic approaches. Similarly, the high frequency and positional predictability of affixes could lead a lexical model to treat them as words. A range of phonological processes apply in Korean, even across word boundaries (Sohn 1999), yielding extensive allomorphy. Phonotactic models may be robust to this kind of variation, but it is challenging for current lexical models (see DP11).

Korean consonantal phonology gives diphones several informative properties, including:

- Various consonant clusters (obstruent-lenis, lenis-nasal, *et al.*) are possible only if they span a word boundary
- Various consonants cannot precede a word boundary
- [ŋ] cannot follow a word boundary

Conversely, unlike in previously studied languages, vowel-vowel sequences are common word-internally. This is likely to be problematic for phonotactic models, but not for lexical ones.

2 Methods

We obtained a *phonetic corpus* representing Korean speech by applying a grapheme-to-phonetic converter to a text corpus. First, we conducted an analysis of this phonetic corpus, with results in Table 1. Next, for comparability with previous studies, two 750,000-word samples (representing approximately one month of child input each) were randomly drawn from the phonetic corpus—the *training* and *test* corpora. The phrasal and baseline DiBS models described above were trained and tested on these corpora; results are reported in Table 2. Finally, we inspected one 'day' worth of segmentations, and offer a qualitative assessment of errors.

2.1 Corpus and phonetic conversion

The Korean Advanced Institute of Science and Technology Raw Corpus, available from the Semantic Web Research Center, semantic-web.kaist.ac.kr/home/index.php/KAIST_Corpus contains approximately 70,000,000 words from speeches, novels, newspapers, and more. The corpus was preprocessed to supply phrase breaks at punctuation marks and strip XML.

The grapheme-to-phonetic conversion system of Kim et al. (2002) was generously shared by its creators. It includes morphosyntactic processing, phrase-break detection, and a dictionary of phonetic exceptions. It applies regular and lexically-conditioned phonological rules, but not optional rules. Kim et al. reported per-grapheme accuracy of 99.7% in one corpus and 99.98% in another.

An example of original text and the phonetic conversion is given below, with phonological changes in bold:

orthographic: 경기도$_1$ 여주에서$_2$ 출생$_3$. 중앙대$_4$문예창작학과를$_5$ 졸업했다$_6$.

phonetic: ㄱㅕ ㅇㄱㅣ ㄷㅗ$_1$ ㅕㅈㅜ ㅔㅅㅓ$_2$ ㅊㅜㄹ**ㅆ**ㅐ ㅇ$_3$ # ㅈㅜ ㅇㅏ ㅇㄷㅐ$_4$ ㅁㅜㄴ**ㅔ**ㅊㅏ ㅇㅈㅏ**ㅋ**ㅏㄱ**ㄲ**ㅏ르ㅡ르$_5$ ㅈㅗ르ㅓ **ㅛ**ㅐ**ㄷㅍ**ㅏ$_6$

IPA: k jʌ ŋ k i t o$_1$ jʌ tɕ u e s ʌ$_2$ tɕ^h u l s* ɛ ŋ$_3$ # tɕ u ŋ a ŋ t ɛ$_4$ m u n e tɕ^h a ŋ tɕ a k^h a k k* wa l ɨ l$_5$ tɕ o l ʌ p^h ɛ t t* a$_6$

(the * diacritic indicates tense consonants)

gloss: Born$_3$ in Yeoju$_2$, Gyeonggi-do$_1$. Graduated$_6$ from Jungang University$_4$ Department of Creative Writing$_5$.

We relied on spaces in the corpus to indicate word boundaries, although, as in all languages, there can be inconsistencies in written Korean.

2.2 Error analysis

An under-researched issue is the nature of the errors that segmentation algorithms make. For a given input word in the test corpus, we defined the *output projection* as the minimal sequence of segmented words containing the entire input word. For example, if *the#kitty* were segmented as *thekitty*, then *thekitty* would be the output projection for both *the* and *kitty*. Similarly, for a posited word in the segmentation/output of the test corpus, we defined the *input projection*. For example, if *the#kitty* were segmented as *theki#tty*, then the *the#kitty* would be the input projection of both *theki* and *tty*. For each word, we examined the input-output relationship. Several questions were of interest. Are highly frequent items segmented frequently enough that the child is likely to be able to learn them? Is it

the case that all or most items which are segmented frequently are themselves words? Are there predicted errors which seem especially serious or difficult to overcome?

3 Results and discussion

The 1350 distinct diphones found in the phonetic corpus were grouped into phonological classes. Table 1 indicates the probabilities (percentage) that a word boundary falls inside the diphone; when the class contains 3 or more diphones, the median and range are shown. Because of various phonological processes, some sequences cannot exist (blank cells), some can occur only word-internally (marked *int*), and some can occur only across word boundaries (marked *span*). For example, the velar nasal [ŋ] cannot begin a word, so diphones of the form *Xŋ* must be word-internal. Conversely, a lenis-/h/ sequence indicates a word boundary, because within a word a lenis stop merges with following /h/ to become an aspirated stop. If all diphones in a cell have a spanning rate above 90%, the cell says *span**, and if below 10%, *int**. This means that all the diphones in that class are highly informative; other classes contain a mix of more and less informative diphones.

The performance of the DiBS models is shown in Table 2. An undersegmentation error is a true word boundary which the segmentation algorithm fails to find (*miss*), while an oversegmentation error is a falsely posited boundary (*false alarm*). The under- and over-segmentation error rates are defined as the number of such errors per word (percent). We also report the precision, recall, and F scores for boundary detection, word token segmentation, and type segmentation (for details see DP11, GGJ09).

model	**baseline**	**phrasal**
under (errs per wd)	43.4	72.5
over (errs per wd)	17.7	22.0
prec (bdry/tok/type)	68/36/34	28/11/12
recall (bdry/tok/type)	46/27/29	11/6/8
F (bdry/tok/type)	55/31/31	15/8/9

Table 2: Results of DiBS models

On the basis of the fact that the oversegmentation error rate in English and Russian was consistently below 10% (<1 error/10 wds), DP11 conjectured that phonotactic segmenters will, cross-linguistically, avoid significant oversegmentation. The results in Table 2 provide a counterexample: oversegmentation is distinctly higher than in English and Russian. Indeed, Korean is a more challenging language for purely phonotactic segmentation.

3.1 Phonotactic cues to word segmentation

Because phonological processes are more likely to apply word-internally, word-internal sequences are more predictable (Aslin et al. 1996; DP11; GGJ09; Saffran et al. 1996; van de Weijer 1998). The phonology of Korean is a potentially

seg. 2 seg. 1	lenis stop	lenis non-stop	tense	asp.	h	n	m	ŋ	liquid	vowel	diphth.
lenis stop	span	100 4-100	int*	27 5-53	span	100 98-100	span		100 10-100	int*	7 0-100
lenis non-stop										int	int
tense										int	int
aspirated										int	int
h										int	int
n	65 29-66	46, 57	38 18-82	45 32-67	35	32	61		span*	12 1-37	53 20-99
m	19 14-21	18, 18	14 4-57	14 12-26	14	int*	21		span	int*	12 1-92
ŋ	12 11-13	10, 12	9 6-55	11 10-15	int*	int*	10		span	6 0-64	18 4-86
liquid	55 43-63	84, 88	71 6-90	53 17-68	42	90	53		int*	3 0-14	39 7-95
vowel	16 6-87	32 12-82	36 4-97	18 3-88	38 9-84	5 1-31	13 2-70	int	int*	44 1-90	51 3-100
diphthong	10 0-79	12 0-55	21 0-100	11 0-87	16 0-88	3 0-15	19 0-74	int	int*	26 0-100	31 0-100

Table 1: Diphone behavior

rich source of information for word segmentation: obstruent-initial diphones are generally informative as to the presence/absence of word boundaries. However, as we suspected, vowel-vowel sequences are problematic, since they occur freely both within words and across word boundaries. Korean differs from English in that most English diphones occur nearly exclusively within words, or nearly exclusively across word boundaries (DP11), while in Korean most sonorant-obstruent sequences occur both within and across words.

3.2 Errors and word-learning

It seems reasonable to assume that word-learning is best facilitated by seeing multiple occurrences of a word. A segmentation that is produced only once might be ignored; thus we defined an input or output projection as frequent if it occurred more than once in the test sample.

A word learner relying on a phonotactic model could expect to successfully identify many frequent words. For 73 of the 100 most frequent input words, the only frequent output projection in the baseline model was the input word itself, meaning that the word was segmented correctly in most contexts. For 20 there was no frequent output projection, meaning that the word was not segmented consistently across contexts, which we assume is noise to the learner. In the phrasal model, for 16 items the most frequent output projection was the input word itself and for 64 there was no frequent output projection.

Conversely, of the 100 most frequent potential words identified by the baseline model, in 26 cases the most frequent input projection was the output word itself: a real word was correctly identified. In 26 cases there was no frequent input projection, and in 48 another input projection was at least as frequent as the output word. One such example is [mjʌn] 'cotton', frequently segmented out when it was a bound morpheme ('if' or 'how many'). The most frequently segmented item was [ke], which can be a freestanding word ('there/thing'), but was often segmented out from words suffixed with [-ke] '-ly/to' and [-eke] 'to'.

What do these results mean for a child using a phonotactic strategy? First, many of the types segmented in a day would be experienced only once (and presumably ignored). Second, infants would not go far astray if they learned frequently-segmented items as words.

3.3 Phrase edges and independence

We suspected the reason that the phrasal DiBS model performed so much worse than baseline was its assumption that phrase-edge distributions approximate word-edge distributions. Phrase beginnings were a good proxy for word beginnings, but there were mismatches phrase-finally. For example, [a] is much more frequent phrase-finally than word-finally (because of common verb suffixes ending in [a]), while [n] is much more frequent word-finally (because of non-sentence-final suffixes ending in [n]). The positional independence assumption is too strong.

4 Conclusion

This paper extends previous studies by applying a computational learning model of phonotactic word segmentation to Korean. Various properties of Korean led us to believe it would challenge both unsupervised phonotactic and lexical approaches.

Phonological and morphological analysis of errors yielded novel insights. For example, the generally greater error rate in Korean is partly caused by a high tolerance for vowel-vowel sequences within words. Interactions between morphology and word order result in violations of a key positional independence assumption.

Phonotactic segmentation was distinctly worse than in previous languages (English, Russian), particularly for oversegmentation errors. This implies the segmentation of simplistic diphone models is not cross-linguistically stable, a finding that aligns with other cross-linguistic comparisons of segmentation algorithms. In general, distinctly worse performance is found for languages other than English (Sesotho: Blanchard et al. 2010; Arabic and Spanish: Fleck 2008). These facts suggest that the successful segmentation model must incorporate richer phonotactics, or integrate some lexical processing. On the bright side, we found that frequently segmented items were mostly words, so a high segmentation error rate does not necessarily translate to a high error rate for word-learning.

References

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.). *Signal to syntax.* Mahwah, NJ: LEA, pp. 117–134.

Blanchard, D., Heinz, J., & Golinkoff, R. (2010). Modeling the contribution of phonotactic cues to

the problem of word segmentation. *Journal of Child Language* 37(3), 487-511.

Daland, R. & Pierrehumbert, J.B. (2011). Learnability of diphone-based segmentation. *Cognitive Science* 35(1), 119-155.

Fleck, M. (2008). Lexicalized phonotactic word segmentation. *Proceedings of ACL-08: HLT*, 130-138.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112(1), 21-54.

Kim, B., Lee, G., & Lee, J.-H. (2002). Morpheme-based grapheme to phoneme conversion using phonetic patterns and morphophonemic connectivity information. *ACM Trans. Asian Lang. Inf. Process.* 1(1), 65-82.

Polka, L. & Sundara, M. (2012). Word segmentation in monolingual infants acquiring Canadian-English and Canadian-French: Native language, cross-language and cross-dialect comparisons. *Infancy* 17(2), 198-232.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 275(5294), 1926-1928.

Sohn, H.-M. (1999). *The Korean Language*. Cambridge: Cambridge University Press.

van de Weijer, J. (1998). Language input for word discovery. *MPI series in psycholinguistics (No. 9).*