THE EFFECT OF ORTHOGRAPHIC KNOWLEDGE ON THE SEGMENTATION OF SPEECH

Bruce L. Derwing and Terrance M. Nearey

University of Alberta, Canada

Yeo Bom Yoon

Seoul National University of Education, Korea

ABSTRACT

As literate English speakers, we are accustomed to the idea that words are made up of individual consonant and vowel sounds, called segments, that these sounds group into larger units called syllables, and that syllables naturally break down into intermediate units that include the rhyme (rime). But are elements like the segment, syllable, and rime universal, i.e., appropriate for the description of all languages? There is experimental evidence that speakers of some languages (e.g., Mandarin or Taiwanese Chinese) may not segment words into units smaller than the whole syllable, while in other languages (e.g., Korean and Japanese) units called the body or the mora may supplant the rime. However, the native speakers tested so far in all four of the language groups mentioned (English, Chinese, Japanese, and Korean) were all relatively well educated, literate, and often even bilingual. Thus they were all exposed to the writing systems of their own and/or their second language, which might have predisposed them to perform the way they did. Since knowledge of the writing systems has not been controlled in previous studies, in the present research we will test speakers of these languages who have not been subjected to the influence of L1 spelling. Such speakers include preliterate children, adult nonliterates, and bilinguals with "split literacy" (e.g., second generation immigrants in Canada and the US, who have learned to speak Japanese, Korean, or a Chinese language natively, but who are literate only in English).

1. PRIOR RESEARCH

The research outlined here is part of a larger cross-linguistic investigation of phonological units in languages of diverse types. Previous research has focused on the status of the syllable (e.g., CVC) and a variety of its hypothesized subsyllabic constituents, including the segment (C or V), the rime (VC), the body (CV)[1], and the mora (a timing unit that can have several phonetic manifestations, including CV). A variety of diverse experimental tasks have been employed in this effort, including (1) word blending (what are the preferred break points when one syllable is blended with another?); (2) global sound similarity judgments (SSJs; what units figure in predicting these?); and (3) concept formation (what is the relative difficulty of identifying a target set defined in terms of one type of unit vs. another?) In a comparison of English with Korean CVC syllables, for example, it was found that English speakers preferred onset plus rime word blends, with break points before the vowel (e.g., SIEVE + FUZZ ¶ SUZZ), while Korean speakers preferred body plus coda blends, breaking after the vowel (e.g., KANG + SEM ¶ ÆKAM)[1,2]. Moreover, in a linear regression analysis of SSJ ratings for CVC-CVC pairs, it was also found that both shared individual segments (Cs or V) and a shared rime (CV) unit made significant independent contributions to mean similarity scores from English speakers, while it was the body (CV) unit that complemented the segments in the results of a comparable task by Korean speakers [3,4]. Furthermore, in a concept formation study (not done in English), Korean speakers found that a target set of words all sharing the common body element /ka/ was easier to identify than a target set of words all sharing the common rime element /ak/ [5]. Taken together, these results indicate that English and Korean CVC syllables are segmented differently, with constituent break points as indicated by the hyphens: into C-VC in English vs. CV-C in Korean.[2]

2. THE INFLUENCE OF ORTHOGRAPHY

It is encouraging that all three of the experimental tasks described above have led to consistent results in the English vs. Korean case.[3] This is because the three tasks are different in the kind of responses they call for and/or the levels of metalinguistic awareness that they presumably invoke. In word blending, for example, subjects produce novel blends (or, in a forced-choice version of the task, choose between one blend and another), a task which would seem to overtly focuses their attention directly on intrasyllabic break points, and presumably to the units that are defined by these breaks. In the concept formation task, too, subjects are directed to discover the phonological units that all members of the target set share, though guided only by feedback as to which stimuli belong to the target set and which do not. In comparing syllable-pairs for similarity in sound, however, subjects are not required to attend to constituent elements at all, but merely to make a global intuitive assessment of overall similarity. The fact that all three tasks lead to the same basic results shows, at the very least, that the findings are not the result of a strategy that is linked to any specific task.

Despite this consistency across the three experimental tasks indicated, we know that at least one confounding factor remains that has not been controlled in any of the studies reported. Specifically, all of the subjects tested in the research described have been literate adults, and most highly educated university students, as well, and hence well versed in the orthographic norms of the languages in which they were tested.

On the one hand, since standard English spelling contains largely units (letters and digraphs) that represent individual phonemic segments—an exception is the letter X, which can

represent the sequence /ks/, as in the word MIX, or even the syllable /‰ks/, as in X-RAY—it does not systematically represent syllabic constituents such as the onset or the rime (though there may exist collocational patterns that might lead to the discovery of such units by literate speakers [6]). On the other hand, the standard Korean orthography contains not only symbols for individual segments but also bundles these elements into syllable-like packages, by stacking the letters in vertical arrays, as with TNß (for the word /sun/ 'pure') and tKß (for the word /san/ 'mountain'). In the first example (where the vowel letter is written with a horizontal orientation), the three letters are simply written one below the next, introducing no particular associative connections between them. However, in the second example (which is typical for CVC strings in which the vowel letter is written with a vertical orientation), notice that the first two letters (representing CV) are written on the same (top) line, with the letter for the coda consonant written below it. This orthographic convention thus suggests that vowels are more closely associated with preceding consonants than with following ones, introducing a potential bias in favor of a body or CV constituent. To insure that this bias was not responsible for the results obtained in our earlier experiments with literate, adult subjects, therefore, we are expanding our tests to include subjects who do not know how to read or write Korean and who would not be subject to the bias that this orthographic convention introduces.[4]

3. TESTING PRELITERATE CHILDREN

While our long-term plans include the testing of three different types of nonliterate Korean native speakers (see abstract), the first group that we have chosen to work with is young, preliterate children. Since pilot testing has indicated that some of the tests that we used with literate adults (such as concept formation) are not well suited for testing children, we have been exploring some new experimental vehicles through which both literate and nonliterate children (or adults) can be tested and compared. Two such new tests have looked promising in pilot testing with English-speaking children, and these are the ones that we are also adapting to the testing of Korean-speaking children.

The first and most promising of these new techniques is a List Recall task. In this task, children are presented with a mixed series of two types of lists of monosyllabic CVC nonsense words, each representing the names of some pictured made-up animals. In one list type, all of the names rhyme, i.e., they all end with the same VC sequence (e.g., /-ip/, as in TEEP, HEEP, MEEP, NEEP); members of each list of the other type all share a common body or CV sequence (e.g., /t‰-/, as in TEP, TETCH, TEM, TENG). Nonsense words are used in order to avoid familiarity and frequency effects. Pilot work suggests that lists that share a viable subsyllabic unit for the language involved (i.e., rimes for onset-rime languages and bodies for body-coda languages) will be recalled better than the opposite lists. (Complete data for both English and Korean speakers will be presented at the conference.)

A second technique that has proven effective in pilot work with children is a Unit Reduplication task. In this task, children are presented with a series of monosyllabic CVC words and, on the basis of modeling with puppets, are asked to both repeat the word unchanged (puppet #1) and then to repeat it with its key subsyllabic element copied (puppet #2), either at the beginning (for the body subtask; e.g., /sup/ ¶ /susup/) or at the end (for the rime subtask; e.g., /sup/ ¶ /supup/. Pilot work suggests that the rime subtask is easier for speakers of onset-rime languages and the body subtask for those of body-coda languages. (Again, complete data for both English and Korean speakers will be presented at the conference.)

Finally, a Reading Test has also been introduced, in order to separate subjects into groups of Readers vs. Nonreaders. In this test, subjects are asked to identify a series of pictures (e.g., of a cake) and then to select the correct spelling of the word from a choice of four alternatives (e.g., HOT, RAKE, CAVE, CAKE). Notice that these alternative spellings have, respectively, none, two, three, and all letters in common with the correct standard spelling of the word. The alternatives are, of course, presented in a different order for each word. Spellers are distinguished from Nonspellers by comparing their number correct scores (on 20 items) with the expectation due to chance (25%).

4. SUMMARY AND CONCLUSIONS

The most interesting comparisons for our purposes will be those contrasting Readers vs. Nonreaders on the List Recall and Unit Reduplication tasks described above. (This work is still in progress but the full results will be presented at the conference.)

5. REFERENCES

1. Derwing, B.L., Yoon, Y.B., and Cho, S.W. "The Organization of the Korean Syllable: Experimental Evidence," *Japanese /Korean Linguistics*, Vol.2, 1993, pp223-238.

2. Wiebe, G.E., and Derwing, B.L. "A Forced-Choice Word-Blending Task for Testing Intrasyllabic Breakpoints in English, Korean and Taiwanese," *21st LACUS Forum*, 1994, pp142-151.

3. Derwing, B.L., and Nearey, T.M. "Sound Similarity and Segment Prominence: A Cross-linguistic Study," *Proc. ICSLP*, 1994, pp351-354.

4. Yoon, Y.B., and Derwing, B.L., "The Sound Similarity Judgement of Korean CVCs by Korean and English Speakers," *Proc. Canadian Linguistic Asso*., 1994, pp657-665.

5. Yoon, Y.B., *Experimental studies of the syllable and the segment in Korean*, PhD dissertation, University of Alberta, 1994.

6. Treiman, R. "Distributional Constraints and Syllable Structure in English," *J. of Phonetics* 16:221-229, 1988.

7. Vennemann, T., *Preference laws for syllable structure and the explanation of sound change*, Mouton de Gruyter, Berlin, 1988.

8. Wang, H.S., and Derwing, B.L. "Is Taiwanese a 'Body' Language?" *Proc. Canadian Linguistic Asso.*, 1993, pp679-694.

9. Derwing, B.L., and Wang, H.S. "Concept Formation as a Tool for the Investigation of Phonological Units in Taiwanese," *Proc. 13th Int. Congress of Phonetic Sciences*, Vol. 3, 1995, pp362-365.

10. Derwing, B.L., and Wiebe, G.E.. "Syllable, Mora or Segment? Evidence from Global Sound Similarity Judgements in Japanese," *Proc. Canadian Linguistic Asso.*, 1994, pp155-163.

11. Otake, T., Hatano, G., Cutler, A., and Mehler, J. "Mora or Syllable? Speech Segmentation in Japanese," *J. Memory and Language* 32: 258-278, 1993.

6. NOTES

1. The use of the term 'body' for a CV unit was first proposed by Vennemann [7] and has since become standard practice in experimental phonology, though the term is evidently used in a quite different sense in the visual word recognition literature.

2. Comparable studies also support an onset-rime analysis for Taiwanese Chinese [8,9] and the mora as a significant unit in Japanese [10] (see also [11], which involves a different approach). Interestingly, the segment has emerged as a significant unit in all of the languages so far tested [3].

3. As indicated in Derwing and Wang [8], there are some inconsistencies in the research on Taiwanese which have yet to be completely resolved.

4. Comparable writing conventions also contaminate the Japanese and Taiwanese research. Specifically, the Japanese writing system is unabashedly mora-based, thus casting into doubt the literate adult research in that language, and the Taiwanese situation is complicated by an onset-rime based training orthography that is used in the early school years in Taiwan (see [9] for further details).

7. ACKNOWLEDGEMENTS

Thanks to Marni Manegre for her assistance with the test design, stimulus selection, and data collection, during the pilot phase of this investigation. This work was supported by an SSHRC research grant awarded to the first two authors.