A New Corpus of Colloquial Korean and its Applications

Kevin Tang, Brent de Chene
University College London, Waseda University
kevin.tang.10@ucl.ac.uk, dechene@waseda.jp

Motivation: Speech produced outside the phonetics laboratory provides ecological validation for experimental findings. Using a newly constructed subtitle corpus, we present here a corpus-based approach to modeling variations in spontaneous speech. While everyday spontaneous speech is inexpensive to collect from field work and digital media, the transcription time and cost remain obstacles to creating a large corpus, capable of capturing a wide range of sensitive linguistic variations. The existing corpora of Korean are too small to be representative of the language. For instance, the spoken-transcripts in the 21st Century Sejong Corpora (Kim and Kang, 1998) provide 5.2 million words and the ETRI database (2006) consists of 24,300 sentences read by a single speaker. The lack of a suitable corpus motivated us to construct our own corpus of conversational Korean.

SUBTLEX, a method of using film subtitles to construct corpora, was developed by New et al. (2007) with French, and has since been applied to many other languages (see, e.g., Brysbaert and New, 2009; Keuleers, Brysbaert, and New, 2010). Crucially these film subtitle frequencies have been proven to be excellent predictors of behavioural task measures such as reaction times in lexical decision tasks. This is primarily due to their spoken register, as they are essentially transcribed spoken speech, as well as their large corpus size. Following this method, we set out to compile a subtitle corpus of Korean, SUBTLEX-KR.

Method: A Korean subtitle website was datamined, and the files preprocessed to remove irrelevant information. Non-Korean files were filtered by applying a language detection model (Shuyo, 2010). Subtitle files were screened for duplicates using the Kullback–Leibler divergence (Kullback and Leibler, 1951). The resulting corpus contained 90 million word tokens and 3.6 million word types, where "word" = *eojol,* a unit between spaces in Korean writing.

Applications: SUBTLEX-KR is useful in investigating a wide range of Korean phonological phenomena that involve variation, both when that variation represents regularization in progress and when it represents alternative resolutions of conflicting forces. Here we will briefly treat three such phenomena. The first is variation in the realization before vowel-initial clitics of noun stems that end historically in marked obstruents (/p^h, t^h, c^h, c, k^h, k'/) or clusters (/ps, ks, lk/, etc.). For many such stems, the historically expected prevocalic allomorph varies with an allomorph that coincides with the prepausal and preconsonantal form (for stems ending in coronals, with a form resulting from a productive rule that takes *t* to *s* prevocalically at the end of a noun stem). After noting evidence that, as widely assumed, underlying forms of noun stems have been reanalyzed as coinciding with prepausal or isolation forms, we will introduce two possible interpretations of this variation. On one of them, variation is due to stochastic rules that mirror lexical statistics (Jun 2010); on the other, variation is the result of ongoing elimination of irregular allomorphs, with concomitant expansion of the range of the corresponding default forms. These two interpretations make different predictions about which stems should vary and which should be stable. While existing corpora fail to supply evidence relevant to the choice between them, we will show that SUBTLEX-KR provides clear evidence in favor of the interpretation on which variation results from ongoing regularization. The second phenomenon we will investigate is variation involving the residual vowel harmony alternation *a*～*ə* in verbal inflectional suffixes. We will compare results from SUBTLEX-KR with earlier results from other corpora (Hong 2008, Kang 2012) and present an interpretation of this case as well in terms of ongoing regularization. Finally, we will consider variation in the epenthesis of /ɨ/ after stop-final English loanwords and show that SUBTLEX-KR allows us to test the conclusions of Kang (2003) with regard to this phenomenon.

References

Brysbaert, M., New, B. 2009. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. In: *Behavior Research Methods* 41.4, 977–990. / Electronics and Telecommunications Research Institute (ETRI). 2006. Database of conversational sentences for speech synthesis. http://slrdb.etri.re.kr/DBSearch/Voice.asp. / Hong, S-H. 2008. Variation and exceptions in the vowel harmony of Korean suffixes. *Journal of Studies in Language* 24:405-428. / Jun, J. 2010. Stem-final obstruent variation in Korean. *Journal of East Asian Linguistics* 19:137-179. / Kang, H. 2012. *Diachrony in synchrony: Korean vowel harmony in Verbal Conjugation.* Dissertation, Stony Brook University. / Kang, Y-J. 2003. Perceptual similarity in loanword adaptation: English postvocalic word-final stops in Korean. *Phonology* 20:219–273. / Keuleers, E., Brysbaert, M., New, B. 2010. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. In: *Behavior Research Methods* 42.3, 643–650. / Kim, H-G., Kang, B-M. 1998. 21st Century Sejong Project-Compiling Korean Corpora. In: *Development* 1999, 2000. / Kullback, S., Leibler, R.A. 1951. On information and sufficiency. In: *The Annals of Mathematical Statistics* 22.1, 79–86. / New, B. et al. 2007. The use of film subtitles to estimate word frequencies. In: *Applied Psycholinguistics* 28.4, 661–677. / Shuyo, N. 2010. Language Detection Library for Java. url: http://code.google.com/p/language-detection/.