A New Corpus of Colloquial Korean and its Applications

Kevin Tang1, Brent de Chene2

1University College London, Department of Linguistics & 2Waseda University, School of Education; kevin.tang.10@ucl.ac.uk, 2dechene@waseda.jp

Introduction

Validity Speech produced outside the phonetics laboratory provides ecological validation for experimental findings. Case studies We show how a newly constructed subtitle corpus of Korean can model variation in spontaneous speech with case studies involving (a) noun inflection and (b) vowel epenthesis in stop-final English loanwords.

Motivation

Existing Corpora of Korean Spoken The ETIH (2006) database contains 30 hours of road speech (24,300 sentences) of a single speaker. × Too small – unreliable estimation of low frequency words (at least 6 million words required, Brysbaert and New, 2009). Written The 21st Century Sejong Corpus (www.sejong.or.kr) (95.5 mil. of which 5.2 million words spoken) and the Trends 21 corpus (Hungru Kim et al., 2011) (480 mil. of newspaper texts, not openly available) × Formal and edited (normalised)

SUBTLEX: Constructing Corpora from Subtitles

Essentially transcribed spoken speech and of 50–500 mil. words × Wide range of genres, tenors, persons, speech acts in dialogues × Outperform written-corpora in terms of % variance explained of behavioural task measures, e.g. English (Brysbaert and New, 2009), Polish (Man德拉 et al., 2014), Dutch (Kuileveld, Brysbaert, and New, 2010). Brazilian Portuguese (Tang, 2012). × No phonetic recordings, translated mainly from English TV/films

Method

Mined 98,393 Korean subtitle files from the web. Cleanned irrelevant information – subtitle line number, time indications, e-mail addresses and websites. Filtered Non Korean files. De-duplicated as popular films get uploaded more often. Enriched with HaeNammu morphological analyser. ≈ 90 million eojeols (orthographic words). 3.6 million word types

Conclusions


Illustrated that the methodological innovation of using speech-like text corpora, such as SUBTLEX, can shed light on cognitive questions about the spoken language and is complementary to experimental and theoretical constructs in linguistics.

Acknowledgements

We would like to thank Andrew Nevis for suggesting the project, Paweł Mandera for developing the software with compiling the corpus and especially Jieun Bark for her extensive assistance with the Korean data.

Ongoing Regularisation in Noun Inflection

Background

• A variety of obstruents and clusters occur stem-finally in Korean nouns and verbs. • Before V-initial suffixes, these are reyalphabetised into onsets and surface clusters. • Before C-initial suffixes, however, they are confined to codas and subject to neutralization and cluster reduction. • They consequently alternate with /p t k/, the only permissible coda obstruents. In verbs, these alternations are stable, indicating that they involve no lexical irregularity: this implies in turn that the compulsive reduplication of verb stem bases are basic. For noun stems, in contrast, there is reason to believe that: • neutralised preconsonantal alternants are the default representations; • alternants other than the default are irregular, except that there is a rule taking stem-final /t s/ to before a vowel (Ko, 1989). The evidence for this analysis is the ongoing elimination of irregular alternants and the productivity of the t-s rule.

What can SUBTLEX tell us about these changes?

• It provides evidence that (as argued for changes affecting coro- nal obstruents by Kang (2003a)) they are analogically rather than phonologically motivated (for claims to the contrary, see Kang, 2003b). (95.5 mil. of which 5.2 million words spoken) and the Trends 21 corpus (Hungru Kim et al., 2011) (480 mil. of newspaper texts, not openly available) × Formal and edited (normalised)

Vowel Epenthesis after Postvocalic Word-Final Stops

Why “gag” /gæg/ /goʊ/ /kæg/ /kæ.g/? • Place of Final Stop • Voicing of Final Stop • Tenesness of Final Vowel • Word length • Stress of Final Syllable

Previous work

Kang, 2003a – Evaluated predictors in isolation from each other. Rhee and Choi, 2001 – Analysed the relative contribution of pre- dictors using a simple main-effects logistic model.

Reanalysis with SUBTLEX

Since Words are better modelled as a random effect (as opposed as a fixed effect) (Clark, 1973), we reanalysed previous findings and ex- plored the complex interactions involved in vowel epenthesis with ad- ditional predictors (Final Vowels, Words and Source Language Freq.). We analysed the epenthesis variations of ≈ 450 English loanwords estimated using SUBTLEX–KR (instead of South Korean words, 1998) with: • Mixed-effects Logistic Models predicting the binary epenthesis text/corpus and Words as a random effect. • Conditional Inference Trees predicting the level of epenthesis of a given word type using a log-logistic metric Log(Freq of Vowel Epens/Freq of Non Vowel Epens).

We examined the complex interactions (Single Tree) and the conditional importance of the predictors (Random Forest).

This evidence is that (a) innovative stems (loanwords) are invari- ant exemplars of default patterns rather than showing the variation according to lexical statistics that ProbMatch (Zuraw, 2000) predict- els; (b) established stems show variation in principle if only if they are irregular, rather than displaying either the uniform invari- ance (Zuraw, 2000) or the uniform variability (see Jun, 2010:146 on Korean *-stems) that a ProbMatch theory could postulate.

Findings

Words as a random effect The random-incorrect of Words captured a large portion of the variance – $R_{random}^2 = 0.447$, $R_{fixed}^2 = 0.496$ (Nakagawa and Schielzer, 2013).

Unimportant All mixed-effects Regression Estimates (from Mixed models) (2), Conditional Variable Importance (from Random Forest) (3) and the single Conditional Inference Tree (Fig. 4) – suggested that Stress of Final Syllable and Monoling- ularity are relatively weak predictors.

Interactions Final Vowels is most important according to both the Forest (Fig. 3) and the Tree (Fig. 4). Teneness is insufficient to capture all the predictive power of Final Vowels (3 times lower, Fig. 3). Final Vowels suggested three levels of vowel quality, [æ i ɛ] and [ɛ i ɔ] and [o i ʌ] and [ʌ i ɔ]. New predictor Random Forest highlighted the importance of Source Language Frequency – more important than most of the predictors (ranked 2nd, Fig. 3). Interestingly, the direction of its effect appeared to be dependent on Final Vowels (Fig. 4).

Variability The NAKI, browsed survey (국립국어원, 1997), which was based on newspapers and magazines, severely underesti- mated the amount of variability in vowel epenthesis, presum- ably due to editing. It estimated only 6% of words with variable epenthesis, while SUBTLEX estimated 41%.

References


