A New Corpus of Colloquial Korean and its Applications

Kevin Tang[1], Brent de Chene[2]
[1]University College London, Department of Linguistics & [2]Waseda University, School of Education; [1]kevin.tang.10@ucl.ac.uk, [2]dechene@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 ETRI (2006) database contains 30 hours of read speech (24,300 sentences) of a single speaker.

× Too small – unreliable estimation of low frequency words (at least 16 million words required, Brysbaert and New, 2009)

Written The 21st Century Sejong Corpora (`www.sejong.or.kr`) (95.5 mil, of which 5.2 million words spoken) and the Trends 21 corpus (Hung-Gyu Kim et al., 2011) (400 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, tenses, 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 (Mandera et al., 2014), Dutch (Keuleers, 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.

Cleaned 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 HanNanum morphological analyzer.

⟹ 90 million eojeols (orthographic words), 3.6 million word types.

Conclusions

Developed a new corpus of colloquial Korean which will remain accessible as a potentially valuable resource for a number of researchers in adjacent fields.
⟹ `http://tang-kevin.github.io/Tools.html` ⟸

Documented variation that makes little or no appearance in edited texts/corpora and elucidated its characteristics.

Confirmed previous findings by Kang, 2003a and de Chene, 2014.

Demonstrated how using a combination of statistical models can reveal hidden patterns in the data (Tagliamonte and Baayen, 2012).

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 Nevins** for suggesting the project, **Paweł Mandera** for his help 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 resyllabified into onsets and surface unmodified.
- 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 contrastive prevocalic alternants of verb stems are basic. For noun stems, in contrast, there is reason to believe that:

- neutralized preconsonantal alternants are the default representations;
- alternants other than the default are irregular, except that
- there is a rule taking stem-final *t* to *s* before a vowel (Ko, 1989).

The evidence for this analysis is the ongoing elimination of irregular alternants and the productivity of the *t*-to-*s* rule.

What can SUBTLEX tell us about these changes?

- It provides evidence that (as argued for changes affecting coronal obstruents by Kang (2003b)) they are analogically rather than phonologically motivated (for claims to the contrary, see 국립국어원, 2004:7 and, for coronals, Hyunsoon Kim, 2001).

This evidence is that for stems with irregular allomorphs, regularisation rate in the corpus is inversely rather than directly proportional to corpus frequency (for the relationship between frequency and (a) sound change (b) analogy see e.g. Hooper, 1976). (See Fig. 1)

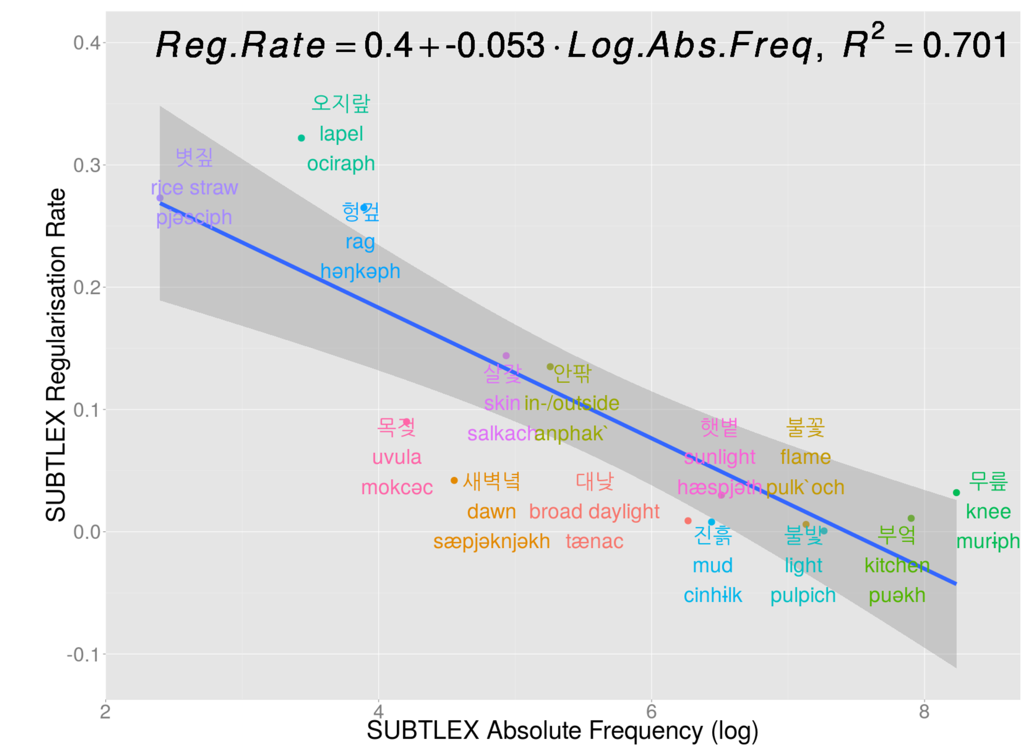
- It provides evidence that they represent the consequences of a "Probability Maximization" rather than a "Probability Matching" response to the problem posed by the alternation of basic X with multiple Y_i (see de Chene, 2014).

ProbMatch: $X \sim Y_i$ is analyzed by postulating multiple stochastic rules $R_i : X \to Y_i$ whose strength is proportional to their lexical frequency.

ProbMax: $X \sim Y_i$ is analyzed by postulating (at most) a single rule $R : X \to Y_{max}$, where Y_{max} is the Y_i with highest lexical frequency; other Y_i are irregular and subject to elimination over time. (If $Y_{max} = X$, no rule is postulated, and the elimination of irregular Y_i results in leveling.)

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

Figure 1: Correlation between Regularisation Rate and Corpus Frequency



Vowel Epenthesis after Postvocalic Word-Final Stops

Why "gag" /gæg/ ⟶ /kæ.ki/?

- Place of Final Stop
- Voicing of Final Stop
- Tenseness of Final Vowel
- Monosyllabicity
- Stress of Final Syllable
- Source Language Freq. – 1^{st} Principal Component of the frequency norms in SUBTLEX–US & UK
- Final Vowels and Words

Previous work

Kang, 2003a – Evaluated predictors in isolation from each other.

Rhee and Choi, 2001 – Analysed the relative contribution of predictors 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 explored the complex interactions involved in vowel epenthesis with additional predictors (Final Vowels, Words and Source Language Freq.).

We analysed the epenthesis variations of ≈ 450 English loanwords estimated using SUBTLEX–KR (instead of 국립국어원, 1990) with:

Mixed-effects Logistic Models predicting the binary epenthesis output of a given word token with *Words* as a random effect.

Conditional Inference Trees predicting the level of epenthesis of a given word type using a log-ratio metric $Log(Freq\,of\,Vowel\,Epen/Freq\,of\,Non\,Vowel\,Epen)$.

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

Figure 2: Regression Estimates from Mixed-Effects Models

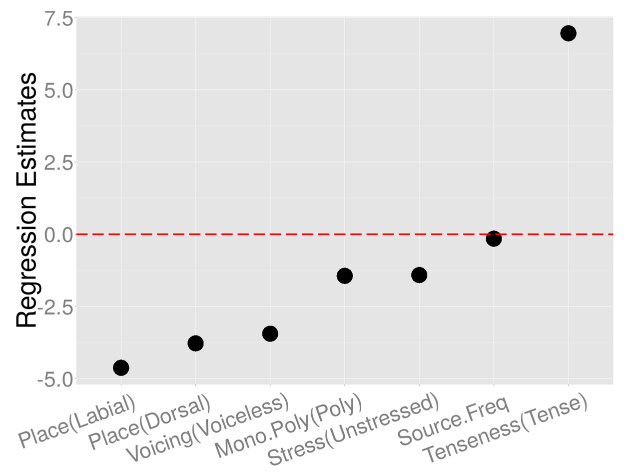


Figure 3: Conditional Variable Importance from Random Forest

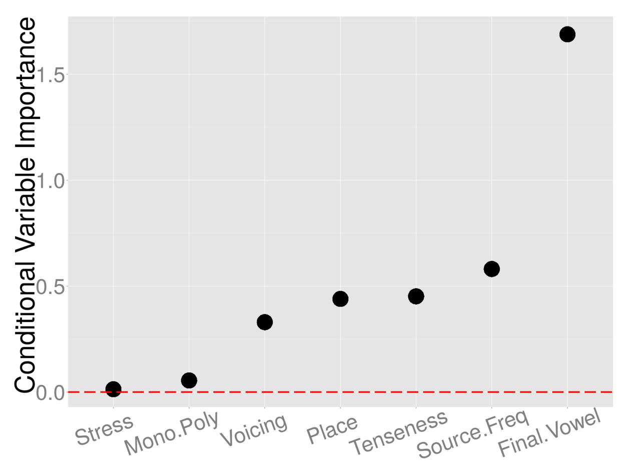
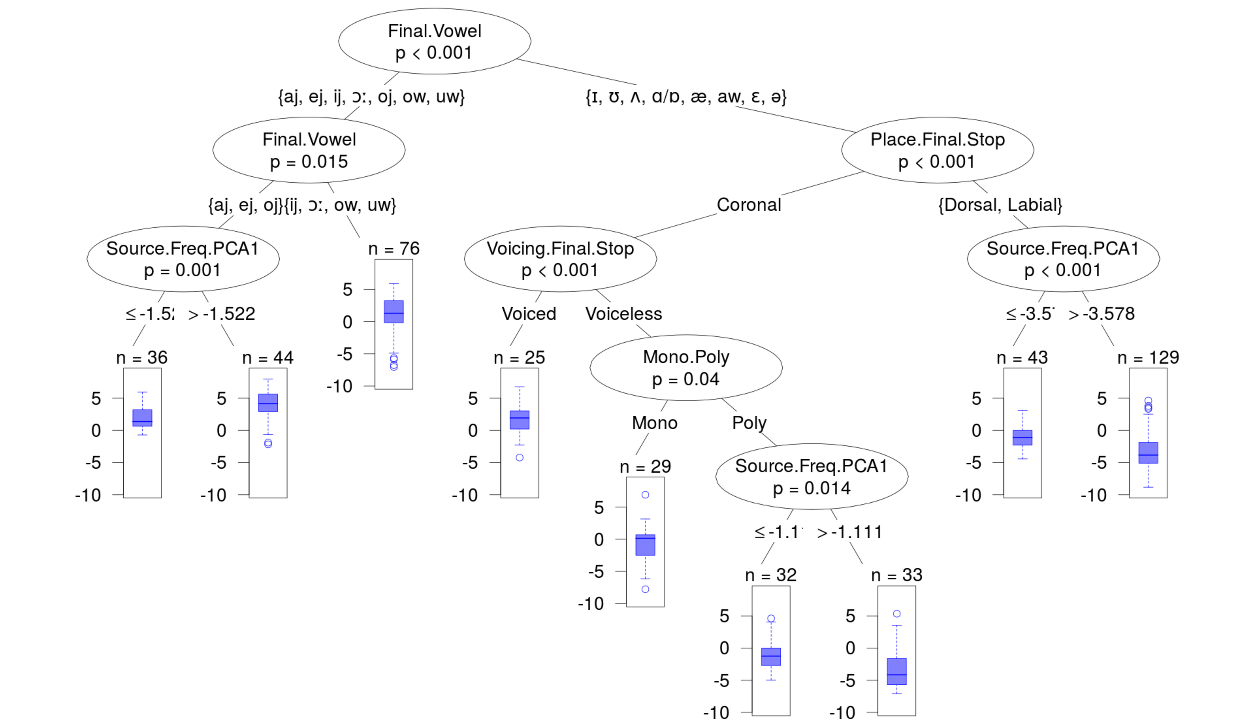


Figure 4: Conditional Inference Tree (boxplots represent log-ratios of epenthesis)



Findings

Words as a random effect The random-intercept of *Words* captured a large portion of the variance – $R^2_{Random} = 0.447$, $R^2_{Fixed} = 0.496$ (Nakagawa and Schielzeth, 2013).

Unimportance All models – Regression Estimates (from Mixed models) (Fig. 2), Conditional Variable Importance (from Random Forest) (Fig. 3) and the single Conditional Inference Tree (Fig. 4) – suggested that *Stress of Final Syllable* and *Monosyllabicity* are relatively weak predictors.

Interactions *Final Vowels* is most important according to both the Forest (Fig. 3) and the Tree (Fig. 4). *Tenseness* is insufficient to capture all the predictive power of *Final Vowels* (3 times lower, Fig. 3). *Final Vowels* suggested three levels of vowel quality, {aj, ej, oj}, {ij, ɔː, ow, uw} and {ɪ, ʊ, ɑ/ɒ, æ, ɛ, ə}.

New predictor Random Forest highlighted the importance of *Source Language Frequency* – more important than most of the predictors (ranked 2^{nd}, Fig. 3). Interestingly, the direction of its effect appeared to be dependent on *Final Vowels* (Fig. 4). Such a pattern is difficult to discover with a linear model.

Variability The NAKL loanword survey (국립국어원, 1990), which was based on newspapers and magazines, severely underestimated the amount of variability in vowel epenthesis, presumably due to editing. It estimated only 6% of the words with variable epenthesis, while SUBTLEX estimated 41%.

References

Brysbaert, M. and B. New (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, pp. 977–990.

Clark, Herbert H (1973). "The language-as-fixed-effect fallacy: A critique of language statistics in psychological research". In: *Journal of verbal learning and verbal behavior* 12.4, pp. 335–359.

de Chene, Brent (2014). "Probability Matching versus Probability Maximization in Morphophonology: The Case of Korean Noun Inflection". In: *Theoretical and applied linguistics at Kobe Shoin* 17, pp. 1–13.

ETRI (2006). *Database of conversational sentences for speech synthesis (Electronics and Telecommunications Research Institute)*. `http://slrdb.etri.re.kr/DBSearch/Voice.asp`.

Hooper, Joan B (1976). "Word frequency in lexical diffusion and the source of morphophonological change". In: *Current progress in historical linguistics*, pp. 96–105.

Jun, Jongho (2010). "Stem-final obstruent variation in Korean". In: *Journal of East Asian Linguistics* 19.2, pp. 137–179.

Kang, Yoonjung (2003a). "Perceptual similarity in loanword adaptation: English postvocalic word-final stops in Korean". In: *Phonology* 20.02, pp. 219–273.

— (2003b). "Sound changes affecting noun-final coronal obstruents in Korean". In: *Japanese/Korean Linguistics* 12, pp. 117–127.

Keuleers, E., M. Brysbaert, and B. New (2010). "SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles". In: *Behavior Research Methods* 42.3, pp. 643–650.

Kim, Hung-Gyu et al. (2011). "Trends 21 Corpus: A Large Annotated Korean Newspaper Corpus for Linguistic and Cultural Studies". Digital Humanities 2011. URL: `http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-257.xml;query=;brand=default`.

Kim, Hyunsoon (2001). "A phonetically based account of phonological stop assibilation". In: *Phonology* 18.01, pp. 81–108.

Ko, Kwang-Mo (1989). "Explaining the noun-final change t> s in Korean". In: *Eoneohag* 11, pp. 3–22.

Mandera, Paweł et al. (2014). "Subtlex-pl: subtitle-based word frequency estimates for Polish". In: *Behavior research methods*, pp. 1–13.

Nakagawa, Shinichi and Holger Schielzeth (2013). "A general and simple method for obtaining R2 from generalized linear mixed-effects models". In: *Methods in Ecology and Evolution* 4.2, pp. 133–142.

Rhee, Seok-Chae and Yoo-Kyung Choi (2001). "A statistical observation of vowel epenthesis in English loanwords in Korean and its significance". In: *Studies in Phonetics, Phonology and Morphology* 7.1, pp. 153–176.

Tagliamonte, Sali A and R Harald Baayen (2012). "Models, forests, and trees of York English: Was/were variation as a case study for statistical practice". In: *Language Variation and Change* 24.02, pp. 135–178.

Tang, K. (2012). "A 61 Million Word Corpus of Brazilian Portuguese Film Subtitles as a Resource for Linguistic Research". In: *UCL Working Papers in Linguistics* 24, pp. 208–214.

Zuraw, Kie Ross (2000). "Patterned exceptions in phonology". PhD thesis. University of California Los Angeles.

국립국어원 (1990). 외래어 사용 실태 조사. 서울: 국립국어원.

— (2004). 표준 발음 실태 조사 *3*. 서울: 국립국어원.