K-SPAN (Korean Surface Phones and Neighborhoods)

Jeffrey J. Holliday, Rory Turnbull and Julien Eychenne

September 1, 2016

The K-SPAN database provides surface phonetic forms derived from a publicly available orthographic corpus of Korean, namely the Modern Korean Usage Frequency Survey 2 (Kim, 2005; Korean title: "현대 국어 사용 빈도 조사 2"), which can be freely downloaded from the website of the National Institute of the Korean Language (NIKL) (see below). In addition, the K-SPAN corpus provides neighborhood density statistics for each word in the corpus. The surface phonetic forms are rendered in an ASCII-encoded scheme, which allows users to explore and query the database without having to read Korean orthography. This database is described in detail in the following paper:

> Holliday, Jeffrey J., Rory Turnbull and Julien Eychenne. Surface phonetic forms and phonological neighborhood density statistics from an orthographic Korean corpus, manuscript.

The file named `kspan_base.csv` contains the phonetized database. It can be merged with the original NIKL corpus by running the script named `merge_kspan.py`, as explained below.

1 Content of the K-SPAN corpus

The K-SPAN database contains the following columns, in this order:

- Word: the word form, in Hangeul
- Disambiguation: an optional disambiguation, in Hanja for Sino-Korean words or in the source language for other borrowings
- pos: part-of-speech
- Frequency: frequency count
- Rank: frequency rank
- WordNum: word number in the original NIKL corpus
- ModernKey: modern pronunciation in keystrokes (e.g. "o" represents the letter " ㅐ")

- ConservativeKey: conservative pronunciation in keystrokes
- OrthographyKey: orthographic representation in keystrokes
- ModernPhon: modern pronunciation in WorldBet
- ConservativePhon: conservative pronunciation in WorldBet
- SyllableCount: number of syllables in the phonetized forms. This is identical to the syllable count in the orthographic form except for the name of Hangeul letters (e.g. ㄱ phonetized as [kiʌk]).
- OrthographyNumNeighbors: number of neighbors for the word in the orthographic representation
- OrthographyMeanNeighborFreq: mean neighbor frequency for the word in the orthographic representation
- ModernNumNeighbors: number of neighbors for the word in the modern pronunciation
- ModernMeanNeighborFreq: mean neighbor frequency for the word in the modern pronunciation
- ConservativeNumNeighbors: number of neighbors for the word in the conservative pronunciation
- ConservativeMeanNeighborFreq: mean neighbor frequency for the word in the conservative pronunciation
- OrthographyNumNeighborsSyll: number of neighbors for the word in the orthographic representation (syllable-based)
- OrthographyMeanNeighborFreqSyll: mean neighbor frequency for the word in the orthographic representation (syllable-based)
- ModernNumNeighborsSyll: number of neighbors for the word in the modern pronunciation (syllable-based)
- ModernMeanNeighborFreqSyll: mean neighbor frequency for the word in the modern pronunciation (syllable-based)
- ConservativeNumNeighborsSyll: number of neighbors for the word in the conservative pronunciation (syllable-based)
- ConservativeMeanNeighborFreqSyll: mean neighbor frequency for the word in the conservative pronunciation (syllable-based)

If the database is merged with the NIKL corpus (as explained in Section 2), the following five columns will be added as the first five columns:

- Word: the word form, in Hangeul

- Disambiguation: an optional disambiguation, in Hanja for Sino-Korean words or in the source language for other borrowings
- pos: part-of-speech
- Frequency: frequency count
- Rank: frequency rank

Table 1 provides the keystroke, WorldBet (Hieronymus, 1994), IPA (International Phonetic Alphabet), and Hangeul jamo (Korean alphabet) representation for each segment present in the corpus (including both modern and conservative forms). Note that the diphthongs ㅘ ㅝ ㅚ ㅟ ㅙ ㅞ ㅢ, which correspond to a combination of two keystrokes in Hangeul, were rendered with a single pseudo-keystroke (digits from 1 to 7) for the sake of consistency.

2 Merging the NIKL and K-SPAN corpora

The K-SPAN database solely provides surface phonetic forms and neighborhood statistics. For the sake of convenience, we provide a Python script that can merge the K-SPAN and NIKL corpora into one. As a side effect, this script automatically converts forms from the NIKL corpus to Unicode (UTF-8), which makes it much more convenient to work with the data on platforms other than Windows with a Korean locale. The original NIKL corpus, which is distributed under an open share-alike license, is available at the following address:

`http://korean.go.kr/front/reportData/reportDataView.do?mn_id=45&report_seq=1&pageIndex=1`

The raw corpus is available as a ZIP archive entitled `현대 국어 사용 빈도 조사 2.zip`. Uncompressing the ZIP file will create a directory entitled 현대+국어+사용+빈도+조사+2, which contains several files in TXT, Excel and PDF format. The relevant file, which contains the full list of lexical items from the corpus, is entitled `일반어휘통계.txt`. Note that on Linux, Mac and Windows systems with a non-Korean locale, file names may not be displayed properly. Should that be the case, the file can still be identified thanks to its size: it is the largest TXT file in the directory, weighing about 2 megabytes. We suggest renaming this file to `nikl_original.txt`. It contains the following columns: rank, word frequency, word form, disambiguation and part of speech.

The K-SPAN databse is provided as a CSV file named `kspan_base.csv`. This file does not contain any information from the original NIKL corpus; however, the number in the first column corresponds to the word number in the NIKL corpus. For example, the word number for the 8th word in this file is 10, which means that this line corresponds to the 10th word in the NIKL corpus, which is the word 가가호호. To merge the two corpus, put the script named `merge_kspan.py` in the same directory as the two text files. The script assumes that the NIKL corpus is named `nikl_original.txt` and that the phonetized corpus is named `kspan_base.csv`, as indicated above. The resulting file will be named kspan.csv. In order to run this script, users will need to have a working installation of Python 3, which can be obtained at the following address: www.python.org.

To run the script on Windows, simply put the three files (*i.e.* the Python script, the NIKL corpus and the phonetized corpus) together inside the same directory, right-click on the Python script and select Edit with IDLE. A new window will open: the script can be run by clicking on Run Module from the Run menu in that window. This will automatically create the merged file in the same directory. In addition, it will create a file named discarded.csv, which contains all the words in the NIKL corpus which were not included in K-SPAN because they were absent from the dictionary used for phonetization.

The simplest way to run the script on Linux and Mac (or any other UNIX-like system) is to put the three files in the user's home directory, and to run the following command from a terminal window:

```
chmod +x merge_kspan.py && ./merge_kspan.py
```

This will create the merged corpus and the file containing discarded words in the user's home directory.

References

Hieronymus, J. L. (1994). ASCII phonetic symbols for the world's languages: Worldbet. Technical report, AT&T Bell Laboratories.

Kim, H. (2005). *Hyeondae Gugeo Sayong Bindo Josa 2.* National Institute of the Korean Language, Seoul.

keystroke	WorldBet	IPA	jamo
q	p	p	ㅂ
Q	p*	p*	ㅃ
v	ph	p^h	ㅍ
e	t	t	ㄷ
E	t*	t*	ㄸ
x	th	t^h	ㅌ
r	k	k	ㄱ
R	k*	k*	ㄲ
z	kh	k^h	ㅋ
w	tc}	ʨ	ㅈ
W	tc}*	ʨ*	ㅉ
c	tc}h	ʨ^h	ㅊ
t	s	s	ㅅ
T	s*	s*	ㅆ
a	m	m	ㅁ
s	n	n	ㄴ
d	N	ŋ	ㅇ
f	l	l	ㄹ
g	h	h	ㅎ
k	a	a	ㅏ
j	^	ʌ	ㅓ
h	o	o	ㅗ
n	u	u	ㅜ
m	ix	ɨ	ㅡ
l	i	i	ㅣ
o	e	e	ㅐ
p	E	ɛ	ㅔ
1	wa	wa	ㅘ
2	w^	wʌ	ㅝ
3	we	we	ㅙ
4	7	ø	ㅚ
5	y	y	ㅟ
6	wE	wɛ	ㅞ
7	ixi	ɨi	ㅢ
O	je	je	ㅒ
P	jE	jɛ	ㅖ
i	ja	ja	ㅑ
u	j^	jʌ	ㅕ
y	jo	jo	ㅛ
b	ju	ju	ㅠ
:	:	ː	NA

Table 1: Conversion table for keystroke, WorldBet, IPA, and jamo representations