Design Methodology for Bilingual Pronunciation Dictionary

Jong-mi Kim

Department of English Language and Literature, Kangwon National University Hyoja 2 Dong, Chuncheon City, S. Korea

kimjm@kangwon.ac.kr

Abstract

This paper presents the design methodology for the bilingual pronunciation dictionary of sound reference usage, which reflects the cross-linguistic, dialectal, first language (L1) interfered, biological and allophonic variations. The design methodology features 1) the comprehensive coverage of allophonic variation, 2) concise data entry composed of a balanced distribution of dialects, genders, and ages of speakers, 3) bilingual data coverage including L1-interfered speech, and 4) eurhythmic arrangements of the recording material for temporal regularity. The recording consists of the triple way comparison of 1) English sounds spoken by native English speakers, 2) Korean sounds spoken by native Korean speakers, and 3) English sounds spoken by Korean speakers. This paper also presents 1) the quality controls and 2) the structure and format of the data. The intended usage of this "sound-based" bilingual dictionary aims at 1) cross-linguistic and acoustic research, 2) application to speech recognition, synthesis and translation, and 3) foreign language learning including exercises.

1. Introduction

This paper presents the design methodology for the "soundbased" bilingual pronunciation dictionary that covers the comprehensive, and yet, concise variations of

- 1) allophones,
- 2) male and female speech,
- 3) adult and child speech,
- 4) major dialects,
- 5) L1-interfered utterance of "Konglish,"¹ and
- 6) bilingual speech of Korean and English.

Most pronunciation dictionaries from Merriam-Webster, Cambridge, Longman, etc. are "lexicon-based," in that all lexical entries of different meanings are listed for audio registry. One selected speaker typically records them. The proposed dictionary is "sound-based," in that all sound entries are listed for the different allophonic environments, and all recording materials are read by various speakers of balanced distribution. Hence, when one seeks an answer to the question "how many representative variations are there for the phoneme [k] in English and Korean?", then only a sound-based dictionary as opposed to a lexicon-based dictionary can answer the variations from, for instance,

- 1) two major dialects of each language,
- 2) L1-interfered pronunciation, "Konglish,"
- 3) male and female genders,
- 4) adult and child speakers, and
- 5) comprehensive allophonic distributions.

The lexicon-based pronunciation dictionary does not present adequate pronunciation samples for the following reasons:

- 1) Insufficient numbers of the allophonic samples are represented, since the dictionary usually omits the audio recording of inflected forms of a word. Inflected forms, however, provide important sound variations; for example, we can find four consonants in a row in the syllable-final position as in the English word "text."
- 2) Excessive overlapping of allophones occurs because the lexical distribution inherently bears sound preponderance.
- 3) Dialectal variation is excluded by limiting the recording to a selected, and usually single, speaker.

The sound-based pronunciation dictionary, however, covers

- 1) comprehensive allophonic samples by positioning phonemes in all phonotactically possible locations,
- 2) reduced sound preponderance by deleting identical allophones due to the phonological changes, and
- 3) dialectal variations by selecting the speakers of representative dialects in each language.

The intended usage of this sound-based bilingual dictionary aims at the

- 1) cross-linguistic and acoustic research,
- 2) application to speech recognition, synthesis and translation, and
- 3) foreign language learning including exercises.

In order to accomplish the above objectives for the "soundbased" bilingual dictionary, the following sections will analyze

- 1) how sound can be presented in an encyclopedic manner,
- 2) how the quality of data is controlled, and
- 3) how the data are structured and formatted.

For practical needs in Korea, this paper considers the bilingual pronunciation dictionary of both Korean and English. These languages also provide academic insights, because their cross-linguistic variation ranges significantly. Moreover, the two languages are very different in both grammatical and social aspects. In the following sections, we will explore in detail how the difference affects the design of the lexical entry for the dictionary in question.

2. Comprehensive Sound Reference

The sound contents of the dictionary are enriched in both 1) the core-phonological and 2) the para-phonological aspects. The "core-phonological" aspect deals with the comprehensive collection of phonological variation. This type of variation depicts the differences in phonemes, allophones, syllables and stress. On the other hand, "paraphonological" aspect deals with the comprehensive collection of sound variations among speakers. This type of variation is a result of social, geographical or biological differences. Comprehensive coverage requires crosslinguistic, dialectal, and L1-interfered variations of sound along with the speech from different gender and age groups. Part of the methodology proposed here has been attested to in a speech translation database, and developed to ensure comprehensiveness and conciseness (Kim, Dyer & Day, 1998).

¹ The word "Konglish" refers to the English language spoken with Korean accent. It is a byproduct of English language modified, socially widely recognized, and typically used by non-fluent Korean speakers of English living in Korea. It ranges from accent to new-breed of vocabulary.

2.1. Core-phonological aspects

In the core-phonological aspect, the bilingual pronunciation dictionary provides the representative speech forms of 1) phonemes, 2) allophones, 3) syllables, and 4) stress.

2.1.1. Phonemic variation

Phonemically, all languages use different sets of phonemes. In an instance of the cross-linguistic variation, several phonemes in English are absent in the Korean language, which renders Korean learners of English to substitute them with the most similar Korean sound currently in existence in their phoneme inventory. For example, the English phoneme [th] is absent in Korean. They, therefore, would wrongly pronounce "think" more likely as "sink." The bilingual dictionary supplies all phonemes of both languages in such minimal pairs of sound as "think" and "sink."

2.1.2. Allophonic variation

Allophones are determined by language-particular phonological rules. For instance, the allophonic change of [l] in Korean induces the learners to utter the English word "light" as "right." The bilingual dictionary provides all possible sequences of sound in both languages.

In practice, the recording material is chosen from various templates of consonantal and vowel references. The following is an instantiation of CV and VC templates of English vowel reference [ih], which combines with various consonants: "C" stands for one consonant, and "V", one vowel. The transcription follows the conventions of TIMIT (1990).

No.	Pattern	CV Template	VC Template
207	[chihch]	chit	itch
208	[jhihjh]	jitney	cottage
209	[mihm]	mid	dim
210	[nihn]	nit	sin

Table 1: English vowel [ih] variation

In Table 1, the different consonants are combined with the given vowel [ih] in syllable-initial and syllable-final positions. These templates are taxonomically arranged for 14 vowels and 24 consonants in English totaling 336 samples. Vowels in the templates include both stressed and unstressed forms.

Koreans may use the recording material from a longer template than English, since artificial words can be used without inducing the speaker's hesitation. The Korean alphabet provides an accurate sound symbol correspondence to the extent that most dictionaries directly use the Korean alphabet for phonetic transcription. In concrete, the consonants are combined with various vowels in word-initial, syllable-initial and syllable-final positions. For instance, a recording sample [tilti] has word-initial [t], syllable-final [l], and syllable-initial [t]. Thus, the template covers five different allophones derived from three different phonemes. In particular, we find two allophones of /t/, one of the wordinitial position followed by [1], and the other in between [1] and [i].

2.1.3. Syllabic variation

Syllable forms are different from language to language. Due to the difference, the learners may pronounce English onesyllabic word "strike" as "su-tu-ra-i-ku" with five syllables. The bilingual dictionary provides all possible syllabic forms of sound in both languages. In concrete, all possible English consonant clusters are listed in the onset and coda positions. Maximum three consonants are listed syllable-initially; and four consonants, syllablefinally. In Korean, only one consonant is allowed in both syllable-initial and -final positions. Accordingly, the recorded materials include only two consonants at the maximum in a row; one for the coda of the preceding syllable, and the other for onset of the following syllable.

2.1.4. Stress variation

Stress placement is pre-determined in lexicon in English, but not in Korean. Korean rhythm depends on the number of syllables, while English rhythm on the number of stresses. Hence, the bilingual dictionary in the study provides the variation of both stressed and unstressed English vowels, but no variation of stress in Korean.

Another aspect of stress-timedness concerns artificial words. An English native speaker would have a difficult time reading an artificial word, puzzling as to where to put the stress and whether to utter the vowels tensed or laxed. The bilingual pronunciation dictionary uses real words in English to minimize the speaker's hesitation, and artificial words in Korean to maximize the coarticulation variation. The knowledge of stress serves to make a comprehensive coverage of sound variation.

2.2. Para-phonological aspects

Data entry is comprehensive in terms of language variation, dialectal variation, and biological variation of gender and age.

2.2.1. Cross-linguistic variation

The language variation includes

- 1) the cross-linguistic registry of English and Korean phonemes, and
- 2) the L1-interfered English spoken by the learners.

A "Konglish" database is required since the speakers of different languages try to use as many target language words as possible.

2.2.2. Dialectal variation

The dialect variation is classified by

- 1) social class in American dialects of English in the U.S., and
- 2) geographical region in the Korean language.

These parameters took into consideration are both the greatest difference between allophones and the largest number of speakers.

2.2.3. L1-interfered variation

The dictionary also provides the typical forms of erroneous pronunciation by the non-native speakers, which is useful for the so-called "error analysis" in the field of second language education. Error awareness benefits the learners for the improvement of phonetic performance.

The "Konglish" database is comprised of two Korean speakers reading English words. English speakers reading Korean words have not been modeled. Korean speakers using many English words in their speech, however, English speakers usually do not know any Korean.

2.2.4. Biological variation

Gender variation was achieved by assigning both male and female speakers for each dialect. Age variation was by the

adult speech modeled in all dialects and child speech added in the primary dialects of each language.

3. Quality Control of Data

The quality control of the data is achieved in terms of

- 1) reduction of redundant allophones,
- 2) eurhythmic arrangement of recording material,
- 3) tempo-controlled recording,
- 4) computer readable sound transcription,
- 5) aural evaluation of data and substitution, and
- 6) consistent time alignment.

3.1. Reduction of redundant allophones

To achieve precision, this study uses the technique of aligning not only the compared sound, but also the preceding and the following sounds. In order to do so,

- 1) CV templates are made to position the phonemes in different locations within different sequences of sound,
- 2) all possible combinations for different phonemes are arranged within the template, and
- 3) redundancy and preponderance of the allophones are eliminated when caused by phonological assimilation and deletion.

Preponderance in the allophonic entry is reduced on the basis of

- 1) segmental phonotactics,
- 2) stress effects, and
- 3) acoustic saliency of vowels.

Let us first consider how the segmental phonotactics is used to reduce the data size. It combines the patterns of phonological assimilation that merges into an output form. For instance, the Korean sample [minmi] combines two input forms of /mitmi/ and /minmi/. Similarly, the pattern [ninni] is from /nitni/ and /ninni/. This is because an obstruent is nasalized in front of nasal consonants in Korean. Other types of phonological assimilations in Korean and English (e.g. tensing, liquid assimilation, vowel reduction) also result in the reduction of data size.

Secondly, the stress variation is modeled only for English and not for Korean. This follows the presence and absence of stress in the underlying level of phonology. In English data, the stressed and unstressed vowel variations fall only on the following three places within words to reduce unwanted preponderance; that is, word-final syllable, penult syllable, and anti-penult syllable.

Thirdly, the acoustic saliency of vowels over consonants serves to reduce the combinable variation of vowels to vowel adjacency. Different vowel sequences for recording have not been listed. Unwanted vowel variations have been reduced when combined with a consonant. The consonantal template is filled within a closed syllable by

- 1) an identical consonant to the given consonant, if available, otherwise,
- 2) alveolar obstruents [d, t, s], which affect the formant transition the least.

In addition to the succinct design of the recording contents, other techniques are used in recording, sampling, transcription, labeling and phonetic decision of foreign sounds in order to enhance the standardization of samples and their compatibility to different database formats.

3.2. Eurhythmic arrangement of recording data

Rhythmic adjustments were depended on language typology. English is a stress-timed language and its recording words are arranged in terms of the similarity of the final rhymes. This minimizes the speakers' hesitation and regulates pitch and tempo. On the other hand, Korean is a syllable-timed language and its recording prompts are arranged in terms of the syllable numbers. Korean recording limits the number of syllables per breath group.

Thus some examples of English recording are *mid*, *lid*, *rid*, *hid*, *vivid*; *wit*, *nit*, *shit*, *chit* where *-id* and *-it* rhymes are arranged together with the pause position marked with semicolon in every 4-6 syllables.

Examples of Korean words are *tipti*; *tipti*, *tepte*, *tEptE*, *tIptI*, *tIptI*; *tAptA*, *tapta*, *tuptu*, *topto*, *topto*. Here the pause takes place in approximately every ten syllables. The dummy duplicates are inserted where the irregular pitch and tempo are expected as in the prompt-initial and pre-pause positions. The regulation of pitch and tempo was made in accordance with the rhythm of the two languages of different prosodic typology.

3.3. Recording quality

In order to achieve the standardization of data in timing and rhythm, this study uses an additional and common technique along with those linguistic considerations presented in the preceding section.

For psychological reasons, the practice-reading rehearsal is conducted to attain full coarticulation and to minimize hesitation. It is also important to provide every possible effort to lessen the speakers' anxiety and to comfort them.

For mechanical concerns, the recording took place in a sound treated room for clarity of the signal. Recording with a DAT recorder is recommended over a computer for the reduction of computer noise and for inducing of speakers' familiarity.

Recommended sampling rate is 48 kHz for DAT recording and 16 kHz for audio files in the computer. This combination reduces information loss when formatting the audio files into the 16 kHz samples of wide circulation.

3.4. Computer readable sound transcription

The bilingual pronunciation dictionary lists three types of transcription. They are

- 1) the International Phonetic Alphabets for accurate understanding of sound,
- 2) the computer readable phonetic alphabets for the convenience of label entering from a standard keyboard, and
- 3) the Korean alphabets "Hangul," since it is generally used for phonetic transcriptions of Korean.

I recently devised a new phonetic alphabet for Korean and proposed in consideration of computer readability (Kim, 2000). It features

- 1) use of only lower case letters on the computer keyboard,
- 2) use of similar symbols in the newly adopted romanization system by the government,
- 3) detailed representation of allophonic alternation,
- 4) systematic representation of the phonemic and allophonic distinction, and
- 5) expandability of symbol sets for system-specific phonetic details.

Others including Chung et al. (1994) also have devised the computer readable phonetic alphabets for Korean.

Using these sets of symbols that I proposed has the advantage of

1) increased computer readability by the simplest operation of keyboard when compared to IPA symbols,

- 2) easier mastery of symbols by Korean users who follow the official romanization system,
- 3) ampler expression of phonetic detail when compared to the official romanization system,
- 4) coherency of the phonemic and allophonic representations when compared to the official romanization system, and
- 5) flexibility for adaptation to the individual system of speech processing in the computer.

3.5. Aural evaluation and substitution

The data are aurally evaluated for detection of defects. Samples are deleted and replaced when the recording quality is unsatisfactory caused by the speakers or the machines.

To ensure the accuracy to match the dictionary entry to the corresponding audio files, the audio files are edited into the unit of words and the file names are written in the conventional orthography of the given language. For instance, the audio file of the English word "bat" is named in English orthography "bat," and Korean sample "mamma" is named in Korean orthography "마마". A longer sequence of numerical file coding risks the mismatch of a dictionary entry and its audio file.

Such edition of short and transparently-named files provides an additional advantage for the potential users, enabling them to easily identify and use the data for playing and copying.

3.6. Consistent time alignment

Time alignment is manually done for

- 1) representative samples of all speakers, and
- 2) all samples of one speaker per language.

Other samples are automatically labeled for practical reasons, although the manual correction is desired wherever applicable.

In order to secure consistency and accuracy of labeling and alignment, transcribers (the transcription workers) require

- 1) cross-checking of the others' alignment,
- 2) regular meeting on alignment, and
- 3) constant refinement of alignment guidelines.

Some specifications of the alignment detail, which my transcribers adopted, are given below as an example:

- 1) Word-initial pause is marked as "i-s" (initial-start), since the initial pause has distinct characteristics from non-initial pause. It often includes a burst, and is usually followed by a higher pitch and amplitude.
- 2) Word-medial voiceless stops include the independently labeled closure period that specifies the phonetic quality of the following release as in "pcl, tcl, kcl."
- 3) Sequence of adjacent sonorants is divided in 1/3 length for onset.

4. Data Structure and Format

The data is organized in the form of being 1) informative about the speech samples and 2) compatible to the user's computer environment.

4.1. Informative data structure

Information is needed in relation to 1) each speech sample as a whole and 2) individual allophones within the speech sample. In addition, 3) the CD-ROM contents supply full information about the data.

4.1.1. Information on audio files

Information on speech files is provided in terms of 1) language and dialect, 2) gender, age, region of the speakers, 3) date of recording and data source, 4) filtering, sampling rate, resolution and audio format. Each group of the information corresponds to the understanding of 1) dialectal variation, 2) idiolectal variation, 3) diachronic variation, and 4) difference in machine.

4.1.2. Information on allophones

Segmental information on the phonetic quality of the individual sound is provided in the record structure of the dictionary by listing the preceding and following phonemes. Consider the following examples in Table 2.

					(– pre	ceding phone	me, + followi	ng phoneme)
Sample	File	Korean	First	Second	Third	Fourth	Fifth	Sixth
Number	Name	Transcription	Phoneme	Phoneme	Phoneme	Phoneme	Phoneme	Phoneme
101311	딤디	딤디	t i m d I	t –pau +I	i –t +m	m —i +d	d –m +i	i –d +pau
101312	뎀데	템데	t e m d e	t –pau +e	e –t +m	m –e +d	d –m +e	e –d +pau
101313	댐대	댐대	t ae m d ae	t -pau +ae	ae -t +m	m –ae +d	d –m +ae	ae –d +pau
101314	臣	드	t eu m d eu	t -pau +eu	eu -t +m	m –eu +d	d –m +eu	eu –d +pau

Table 2: Coarticulation data entry

This is to instantiate how each phoneme in the audio files is annotated in terms of the preceding and following phonemes. The dictionary entry of each coarticulation of phonemes is in such a form.

4.1.3 CD-ROM contents

CD-ROM includes full documentation on the data structure, audio files and label files, including at least following descriptions:

- 1) related publication to provide the overview of the dictionary,
- 2) speaker's information,
- 3) phonetic transcription code,
- 4) list of words segmented by phonemes,
- 5) segmentation methods,
- 6) recorded lists of waveform files,

- 7) orthographic transcription of waveform files, and
- 8) phonetic transcription of the utterance variation of each speaker.

4.2. Widely circulated data format

The data format of the dictionary is a widely circulated form. Since the dictionary is distributed in CD-ROM, the format concerns the computer compatibility for the users. In particular, ensuring the compatibility is required on audio file format, word processor format and transcription convention.

4.2.1. Audio file format

The most accessible file format seems to be ".wav" of Window PCM waveform, 16 kHz, and 16-bit. For the

Windows PCM waveform (.wav), all ".wav" formatted files follow the RIFF (Resource Information File Format) specification. Most bits of the special information are saved with the wave file in this manner. The standard Windows PCM waveform contains PCM coded data, which is pure uncompressed pulse code of modulation-formatted data. This type of audio file form is also easily converted to other types of computers.

4.2.2. Phonetic transcription

The phonetic transcription adopts both 1) IPA (International Phonetic Alphabet) for the user's familiarity and 2) the computer readable phonetic alphabets for computational convenience. In addition, 3) Korean orthography is also used for the phonetic transcription. Some examples of the computer readable phonetic alphabets are provided for English words in Table 3.

Orthography	Phonetic Alphabets
teethes	t iy1 dh z
teethed	t iy1 dh d
thieve	th iy1 v
thither	th ih1 dh er0
twelfths	t w eh1 l f th s
thetic	th iy t ih0 k

Table 3: Computer Readable English Phonetic Alphabets

4.2.3. Time alignment

Time alignment is provided in two forms: 1) Xwaves format and 2) Excel format. The following is an Xwaves format, where the time is the end of segment, assuming to start at 0.

separator ; nfields 1 # 0.10000 26 pau ; 1.11500 26 n ; 1.29000 26 ow ; 1.30500 26 v ; ...

These can be transformed directly into Excel for the user's PC based programs so that there are two columns.

0.10000	pau
1.11500	n
1.29000	ow
1.30500	v

Otherwise, one would put all of the different segmentations such as utterance, word, and phoneme in paralleled and aligned columns into Excel.

4.2.4. Data size

The data size is limited to 600 megabytes economically to bit into a CD for a wider distribution. The tuned waveform is 16 kHz to save the disk space.

The sampling rate is 48 kHz for the digital audio tapes (DAT), and 16 kHz for the computer audio files. The sample length is 16-bit. This type of setting is optimal for the learners' personal computers.

5. Conclusion

To conclude, the "sound-based" pronunciation dictionary is proposed to make a comprehensive sound reference in a small disk space. At present, only the "lexicon-based" pronunciation dictionary is on the market, which provides not only the limited reference to sound variation, but also the excessively redundant recording of allophones. Detailed instructions are suggested for making

- 1) a comprehensive sound reference,
- 2) controlling the quality of the data, and
- 3) standardizing the data structure and formats.

First, both "core-phonological" consideration of phonemes, allophones, syllables, and stress, and "para-phonological" consideration of cross-linguistic, dialectal, L1-interfered, and biological variations, achieve the comprehensive sound reference.

Second, the quality control of the data is achieved by reduced preponderance of allophones, eurhythmic arrangement of recorded data, psychological and mechanical optimization of recording environment, computer readable phonetic transcription, aural evaluation of replaceable data, and consistent alignment of time.

Third, the data structure is informative and the format is compatible to the user's computer specification. Extensive information is provided for each sound sample, individual allophones within the sample, and CD-ROM contents. The audio file format and labeling convention are familiar to the users. The data size is limited to be under 600 megabytes to fit into one standard CD.

The proposed design methodology achieves

- 1) the completion of the major phonological aspects of the target language,
- 2) the balance of speech variation in terms of dialects, gender, and age-groups,
- 3) the parallel comparison of target sound with the native sound and L1-interfered sound,
- 4) the standardization of samples in timing and rhythm,
- 5) the informative data structure, and
- 6) the widely circulated data formats.

The intended usage of this "sound-based" bilingual dictionary aims at

- 1) the comprehensive reference to sound,
- 2) the cross-linguistic and acoustic research,
- 3) the acoustic phonetic research based on the actual utterance within a language,
- 4) the application to speech recognition, synthesis, translation of the given language, and
- 5) the foreign language learning including exercises.

Acknowledgements

This project is sponsored by the Korea Research Foundation. This bilingual pronunciation dictionary is part of the SORIDA (Sound Rich Database) series that was first introduced in LREC 1998 (Kim et al., 1998). Since then, SORIDA has developed eight different versions of the speech coarticulation database (Kim, 1999).

References

- Kenyon, J.S. & Knott, T.A. (1953). A Pronouncing Dictionary of American English. Springfield, MA: Merriam-Webster Inc.
- Cambridge Dictionary of American English on CD-ROM. (2000). Cambridge University Press.
- Chung, K., Koo, H-s., Han, S-h. & Kim, J-m. (1994). A Study of Korean phonetic and phonological properties for speech recognition and synthesis. Journal of the Acoustic Society of Korea, 13(6), 31--44.

- Kim, J-m. (1999). Speech coarticulation database of Korean and English. Journal of the Acoustic Society of Korea, 18(3), 17-26.
- Kim, J-m. (2000). Computerized phonetic alphabet of Korean. ms.
- Kim, J-m., Dyer, S.A. & Day, D.D. (1988). Construction of a Speech Translation Database. In Proceedings of First LREC Conference. (pp. 1071--1078). Paris: European Language Resources Association.
- Longman Interactive English Dictionary on CD-ROM. (2000). Harlow: Longman.
- TIMIT: Acoustic-Phonetic Continuous Speech Corpus CD-ROM. (1990). Distributed by LDC, UPENN.
- Wells, J.C. (2000). Longman Pronunciation Dictionary. Second Edition. Harlow: Pearson Education Limited.