# AUTOMATIC GENERATION OF CONFUSABLE SETS IN SMART SPELL CHECKING FOR KOREAN LEARNERS OF ENGLISH

Kong Joo Lee & [i]Jee Eun Kim
Dept. of Information Communications Engineering
Chungnam National University, Korea
kjoolee@cnu.ac.kr
[i] Dept. of English Linguistics
Hankuk University of Foreign Studies, Korea
jeeeunk@hufs.ac.kr

**Abstract:** This paper presents an automatically generated English confusable word set to be used for smart spell checking. A confusable set includes pairs or subsets of frequently misused English words. When Koreans learn English as L2, they produce various types of errors, some of which are caused by negative language transfer. The language system of Korean interferes with English which results in creating a peculiar system. In particular, Korean and English present distinct phonetic and phonemic inventories from each other. The distinctions influence not only the pronunciation of a word, but also its spelling. For certain types of spelling errors, smarter suggestions can be provided when a confusable set is modified for Korean learners. The Double Metaphone algorithm is adopted and revised to implement the phonetic and phonemic properties of Korean. The result is used to automatically generate a confusable set which provides customized suggestions to be used in spell checking.

**Keywords:** Smart Spell checking, Korean learner, English

## Introduction

Spell checking techniques become smarter as human language technology develops. Its fundamental algorithm is to check whether a word is known and correctly spelled. A word in its use is supposed to be identical to the lemma or one of its morphologically legitimate variants. When the word is not identified as one of those forms, it is determined as an error. When an error is detected, the spell checker suggests a list of candidates to correct the error, which is generated based on the dictionary entries and various heuristics. Semantically anomalous errors, however, cannot be identified as an error because they are spelled correctly and the error detecting scope is beyond the capacity of the spell checker. These errors are legitimate dictionary entries and occur in a morphologically grammatical form. They are classified as usage errors which occur when a word does not fit in the context. This type of errors can be resolved effectively by utilizing a confusable word set.

Usage errors are produced by the following causes: people choose an inappropriate word because a) they are unfamiliar with the word implying the intended meaning, b) they are not certain of the contextually appropriate form of the word, and c) they are confused with the spelling or the pronunciation of the word in question. These types of errors can be detected only when the scope of spell checking expands to the context from a single word.

1) Our holiday house is among/between the mountains and the sea.
The ancient fountain was hidden among/between the trees.

2) Can I borrow/lend your dictionary?
I never borrow/lend my books to anyone.

In example 1) and 2), only one of the underlined words is grammatical in the context although they are semantically related and occur in an appropriate form with the correct spelling. The underlined pair of words in example 1) implies the same meaning, but one has to be selected depending on the quantity of the following noun phrase. The pair in example 2) is also semantically related, but in antonymic relation. A word of the pair can be chosen only when the intended meaning of the sentence is comprehended. In other words, selecting an appropriate word is context-sensitive in both examples. In order to detect these types of errors, the speller needs to figure out the contextual meaning of the word, which requires expanding the scope of the speller, from a single word to a bigger unit such as a phrase or a sentence.

3) My old diary/dairy called some unpleasant memories to mind.
The most bizarre/bazaar thing has happened.
Her necklace is too expensive to appraise/apprise the value.
You always need to cite/site your sources.

Example 3) presents four pairs of words which contain a word to trigger a usage error. Similarly to example 1) and 2), a context-sensitive error occurs when the second word of the pairs is selected. Each pair looks alike in their spelling and their pronunciations are either similar or identical while all of them are presented in a morphologically grammatical form. Unlike example 1) and 2), the words in each pair are not semantically related. Their spellings or pronunciations are similar or identical, which is confusing enough to cause an error.

This paper focuses on the last type of usage errors presented in example 3). Providing a set of confusable words is known as one of the popular approaches to resolve usage errors. The set is added to the suggestion list from which users choose a contextually appropriate word. Pertaining to the errors created by Korean learners of English, however, a different set of confusables should be derived from examining the errors due to the discrepancies in the phonetic and phonemic inventories between Korean and English. Transferring the native language to L2 leaves Korean traces in English, which affects not only pronunciations but also spellings when Koreans learn English as L2. Since a cause of these errors involves the sound systems, the Double Metaphone algorithm is adopted to generate the pairs of phonetically confusable words. Since the original algorithm is devised for English, the distinctions between English and Korean language systems are examined and implemented in the algorithm. The customized algorithm is expected to resolve the errors caused by Korean specific influences.

## The Study

When Koreans learn the English language as L2, their native language interferes with acquiring correct pronunciations of English since sound systems of the two languages are distinct from each other. In language acquisition, fossilization of the pronunciation system proceeds rather rapidly in the case of L1. As a result, the phonetic and phonemic inventories of L1 easily transfer to L2, which creates a peculiar system and triggers speech production errors.

**Table 1**: English Phonemes - Consonants

| | Bilabial | Labiodental | Interdental | Alveolar | Palatal | Velar | Glottal |
|---|---|---|---|---|---|---|---|
| Stop | p / b | | | t / d | | k / g | ʔ |
| Nasal | m | | | n | | ŋ | |
| Fricative | | f / v | θ / ð | s / z | ʃ / ʒ | | h |
| Affricate | | | | | tʃ / dʒ | | |
| Glide | ʍ/w | | | | j | ʍ/w | |
| Lateral Liquid | | | | l | | | |
| Central Liquid | | | | r | | | |

**Table 2**: Korean Phonemes - Consonants

| | Bilabial | Labiodental | Interdental | Alveolar | Palatal | Velar | Glottal |
|---|---|---|---|---|---|---|---|
| Stop | pʰ/ p'/ p | | | tʰ / t'/ t | | kʰ/ k'/ k | |
| Nasal | m | | | n | | ŋ | |
| Fricative | | | | s'/ s | | | h |
| Affricate | | | | | tɕʰ/ tɕ'/ tɕ | | |
| Glide | | | | | j | w/ɰ | |
| Liquid | | | | l(or r) | | | |

Table 1 and Table 2 display consonant phonemes of English and Korean respectively while they indicate clear differences in their types. One of the noticeable distinctions pertains to the phonemes which English includes but Korean lacks or vice versa. For example, English has dental sounds, labiodental and interdental [f, v, θ, ð] which map to null in the Korean phonemic table. Those phonemes are often replaced by the phonemes available in Korean, such as [p/h, b/p, t/s, t/d]. Another null mapping in the two systems can be found in voiced obstruent consonants; English [b v g □] are absent in Korean. In addition, two distinctive English phonemes function as the allophones of a single phoneme in Korean. The English phonemic table lists two types of liquid, lateral [l] and central [r], both of which are mapped to a single phoneme, the alveolar liquid in Korean. One of allophones is articulated depending on the context.

Because of the differences in the language systems, the Korean language interferes with learning English, which is known as negative language transfer in L2 acquisition. Flege(1987) claims that the L2 learners classify the sounds of L2 based on the L1 sound system, which may result in a new and peculiar system. Accordingly, the resultant system induces various types of production errors. Jang(2005) has categorized the sounds produced by negative language transfer which progresses while Koreans learn English. Since he focuses on speech including phonetic and phonemic representations, the classification has to be revised to include the relation between the pronunciations and the spellings.

The main goal of this research is to provide a set of Korean pronunciation influenced confusable words for spell checking. Since a confusable set by definition is composed of approximately suitable words to be used as suggestion candidates, mapping between the scripts and their pronunciations has to be redefined, also considering the errors produced by Korean learners. In order to customize the set for Koreans, the Double Metaphone algorithm is adopted and revised to satisfy the research need. Double Metaphone refers to the second generation of phonetic encoding algorithm designed for indexing English words according to their pronunciations. The goal of the original algorithm is to reduce the discrepancies between the spelling and pronunciation of English words. The algorithm is found to be effective in identifying words with similar pronunciations. In addition, Double Metaphone is known for working best with recognizing proper names. Even native speakers of English produce different pronunciations for the same name, which frequently results in spelling errors when converting the speech to the text.

More than one acceptable pronunciation can be mapped to a proper name. The Double Metaphone algorithm is designed to allow dual scripts for a single string, primary and secondary encodings in order to capture those script variants and an ambiguity which the string may presents in terms of its pronunciation. The algorithm adopts a single representative script for similar sounds. It attempts to encode all the English words using four consonant metaphones to the maximum, ignoring vowels although the number of metaphones can be adjusted as necessary.

**Table 3:** English Primary Encoding in Double Metaphone

| ALPHABET | PRIMARY | ALPHABET | PRIMARY |
|---|---|---|---|
| *a/e/i/o/u/y* | A (at first position)<br>ignore (others) | *n* | N |
| *b* | P | *p* | P or F(*ph*) |
| *c* | K(*c*horus) or S(*c*aesar) or<br>X(*c*hair) or KS(a*cc*ident) | *q* | K |
| *d* | J(e*d*ge) or TK(ed*g*ar) or T(wi*d*th) | *r* | R |
| *f* | F | *s* | X(*s*ugar) or S(*s*mith) or SK(*s*chool) |
| *g* | K(ba*gg*y) or J(a*g*ile) or F(cou*gh*) or<br>KN(i*g*nite) or N(*g*narl) or KL(*g*lide) | *t* | T or X(*t*ion) or 0(*t*h) |
| *h* | H | *v* | F |
| *j* | J(*j*et) or H(*J*ose) | *w* | R(*w*r-) or A(*W*omo) or<br>TS(filipo*w*icz) or F(Arno*w*) |
| *k* | K | *x* | S(first character) or<br>ignore or KS(breau*x*) |
| *l* | L | *z* | S or J(*Z*hang) |
| *m* | M | | |

Table 3 presents the representative primary encoding for each English alphabet while summarizing the encoding rules of the algorithm. Some of the alphabets such *k* and *m* are mapped to a single pronunciation. In other cases, an alphabet is represented with multiple metaphones encoding similar pronunciations. For example, a letter *c* is represented by 4 different metaphones including K, S, X or KS depending on its pronunciation within a word.

**Table 4:** English Double Metaphone Examples

| METAPHONE PRIMARY ENCODING | EXAMPLES |
|---|---|
| FNRL | venereal, funereal, funeral |
| FN | fawn, feign, finny, vino, vain, vine, faun, fain, fine, funny, van, fen, fin, von, phony, phone, fan, vein, venue, fanny, fun, vane, fauna |
| PST | basset, past, best, bust, beside, baste, boost, biocide, bast,<br>paucity, bestow, beast, pasta, post, paste, beset, peseta, boast, pest, posit, pasty |
| FTLT | futility, vitality, fatality, fidelity, feedlot |
| KKTR | cockatrice, coquetry |
| FNN | vinyon, phonon |
| SNTF | centavo, cenotaph, scientific, sendoff |

Table 4 displays examples generated utilizing Double Metaphone primary encoding described in Table 3. It also shows the pairs of words which share the same metaphone encodings. Native speakers of English often produce pronunciation errors by articulating the counterpart of a sound, with the same phonetic features but voicing. Two consonant phonemes, *p* and *b* are encoded as P since they are identical in their articulatory properties except for voicing. For example, *pest* is frequently pronounced when *bast* is intended. This confirms that native speakers of English tend to get confused with voicing.

## Findings

The coverage of the Double Metaphone algorithm has to be expanded to add a set of Korean influenced confusable words since the algorithm is intended for the sake of the native speakers or the experts of English. In order to add Korean specific rules to the algorithm, the errors produced by Korean learners have been examined. One of the most noticeable errors is incorrectly used English liquids. The Korean phonemic system includes a single liquid which allows two allophones mapped to the two distinctive liquids, lateral and central in the English system. Korean allophonic distinctions of the liquid can be recognized by considering the context; the central liquid is pronounced at the syllable initial position whereas the lateral is articulated elsewhere. Since the Korean liquids are not phonemically distinctive, Koreans often fail in distinguishing one English liquid from the other. For example, the pairs of words such as *right/light* and *lead/reed* are pronounced the same, selecting either one of the pair. Another difference of worth noticing is that voiced obstruent consonants are not available in the Korean sound system, which enforces the learners to replace them with their voiceless counterparts. For example, [p] is articulated for [b] as in [pai] pronounced for *buy*. Similarly, English dental sounds are replaced with Korean phonemes which are not very phonetically similar. For example, [v] in *veil* is articulated as [b] and [f] in *fight* is pronounced as either [p] or [hw].

Not all of the distinctions in the phonetic/phonemic inventories, however, result in spelling errors. Null mapping to English voiced obstruents causes speech errors depending on the position. Only a word initially occurring voiced obstruent is frequently replaced with its voiceless counterpart. For example, a word *gag* is often pronounced as [kag]; a word initially occurring [g] as tends to be pronounced as [k] whereas the word final *g* usually remains as [g]. However, this particular phonemic gap between the two systems does not seem to interfere with spelling L2 words.

With considering the characteristics of various Korean influenced pronunciation errors, the Double Metaphone algorithm is revised to provide an effective confusable set. In order to minimize the revision, only the secondary encoding is modified while adopting the primary as is.

**Table 5:** Korean Encoding in Double Metaphone

| ALPHABET | PRIMARY ENCODING (Original Metaphone) | SECONDARY ENCODING (Modified For Korean) |
|---|---|---|
| *f* | F | F |
| *v* | F | B |
| *b* | P | B |
| *p* | P | F |
| *l* | L | L |
| *r* | R | L |

Table 5 presents the changes in the secondary encoding to represent the influence of Korean pronunciation on English. A metaphone B is added to map *v* and *b* and L to represent *r*. Based on the revision with the secondary encoding, an algorithm is created to check whether a word is qualified for one of the confusable pair.

**Table 6:** Algorithm to Determine Confusables

```
IsConfusable(word a, word b)
{
    if   metaphone(a) == metaphone(b)
                        and   minimum_edit_distance(a, b) < Threshold
                              and   hasSameStem(a,b) == False:
        return True;
}
```

Table 6 presents a function called IsConfusable(word a, word b) describing the procedure to check whether the two input words, *a* and *b* compose a confusable pair. They are determined as a pair if they satisfy all of the following conditions: 1) their metaphone encodings are identical, 2) their minimum edit distance is less than the threshold, and 3) they do not share the same stem. The embedded function hasSameStem(a, b) checks their stems and returns 'True' when they have the same stem. For example, given with the two words, *replicate* and *replicator*, the function IsConfusable returns "False" determining that they are not a confusable pair. They have the same metaphone encoding and their minimum edit distance is below the threshold. However, they share the same stem *replicate*, which enforces the embedded function to return 'True'.

**Table 7:** Confusable Words for Korean Learners of English

| CONFUSABLES | FOR ENGLISH USING PRIMARY ENCODIGN | FOR KOREAN USING SECONDARY ENCODING |
|---|---|---|
| **COMMON** | {base, past} {bast, past}<br>{fine, vine}<br>{rebel, repel} | {fast, past} {fine, pine}<br>{bast, vast}<br>{rebel, revel} |
| **PRIMARY ONLY (English Speakers)** | {funereal, venereal}<br>{backing, packing}<br>{backoff, pickoff}<br>{fatality, vitality} {file, vile} | |
| **SECONDARY ONLY (Korean Speakers)** | | {bane, vane} {bang, vang}<br>{flat, plat} {read, lead}<br>{lace, race} {lasting, resting}<br>{lavish, ravish} {level, revel} |

Table 7 presents a set of confusable words produced as the output of the current research. The confusables are automatically generated using the algorithm described in Table 6. The first column lists the pairs of confusable words generated by implementing the original primary encoding of Double Metaphone while the second column displays the pairs for Koreans produced utilizing the secondary encoding. The row led by COMMON displays the pairs of confusables which can be shared by both English and Korean speakers. The pairs in the PRIMARY ONLY row are the words generated solely for English speakers. The row with SECONDARY ONLY presents the confusables which Korean speakers would benefit from. Table 7 suggests clear differences between English and Korean regarding the types of confusables. For example, English speakers mistake *fine* for *vine* or vice versa, from which the misused phoneme can be predicted using English phonetic/phonemic features. In addition to this confusable pair, Korean speakers get confused with *fine* for *pine* more often, which is created by the influence of Korean phonetic/phonemic system. Unlike the English pair, the replaced phoneme such as [p] for [f] cannot be predicted while causing frequent spelling errors. Accordingly, the customized confusables for Koreans are expected to be effective in providing refined suggestions for common spelling errors when they are added to the existing list in spell checking. Similarly, smart spell checking implemented with context-sensitivity can suggest *revel* over *repel* for an error *rebel* since, for Korean speakers, confusion between the pair, *rebel*/*revel* causes a usage error more frequently than the pair *rebel*/*repel*. In addition, the speller can provide an updated pair, *read*/*lead*, for another common usage error.

## Conclusions

This paper has introduced an automatically generated confusable set to be implemented in smart spell checking. The set is created utilizing the customized Double Metaphone algorithm for Korean learners of English. Double Metaphone is useful in capturing the representative scripts of an alphabet which may be mapped to multiple pronunciations caused by either phonological convention of English or speech production errors. Mapping multiple pronunciations to a single metaphone is effective in creating a set of confusables to be used as suggestion candidates in spelling checking. Since Koreans produce different sets of pronunciation errors from those made by native speakers of English, the secondary encoding in the Double Metaphone algorithm is complemented with a list of confusable words customized for Koreans. Implementing the modified algorithm in spell checking is expected to provide an improved set of suggestion candidates. Additionally, the revised Double Metaphone algorithm can be adopted to create a suggestion list for proper names whose spelling may be confusing or unknown. For example, a smart speller implemented with the algorithm can suggest *Pausini* for *Fausini* inputted by a Korean to search an Italian singer *Laura Pausini* while *Vausini* would be a more effective candidate to be paired with *Pausini* for English speakers.

As the next step of the research, the revised algorithm will be implemented to generate accurate spellings of proper names which are known only with their pronunciations. It is expected to be useful since spelling proper names is often difficult for the learners of English as well as the native speakers. The performance accuracy will be evaluated in comparison with different approaches.

**Acknowledgements**

This work was supported by Hankuk University of Foreign Studies Research Fund of 2015.

## References

Flege, J. E. (1987) The production of new and similar phones in a foreign language: evidence for the effect of equivalence classification, in Journal of Phonetics. 15: 47-65.

Fromkin, V., R. Rodman & N. Hyams (2011). *An Introduction to Language*. 9th ed. Wadsworth.

Jang, T. Y. (2005). Construction of an English speech database for Korean learners of English", Language and Linguistics, Vol 35: 293-310. Language Research Institute. Hankuk University of Foreign Studies. (in Korean)

Phillips, Lawrence (1990). "Hanging on the Metaphone", Computer Language, 7(12).

Phillips, Lawrence (2000). "The Double Metaphone Search Algorithm", C/C++ Users Journal, 18(6).

The Soundex Algorithm, available online at http://www.archives.gov/research_room/genealogy/census/soundex.html

UzZaman, N., & M. Khan (2005). A double metaphone encoding for Bangla and its application in spelling checker. In Natural Language Processing and Knowledge Engineering, 2005.

[i] Corresponding author.