Annotation of Spelling Errors for Korean Learner Corpora
Seok Bae Jang, Sun-Hee Lee, & Markus Dickinson
AALL-09; Tempe, AZ; March 11, 2009

Introduction
Learners frequently make spelling errors in Korean (Ko et al 2004)
Errors reflect aspects of linguistic knowledge (cf. Rimrott & Heift 2008, and references therein)
In Korean: frequent mismatches between a syllable and a character, due to phonological rules and morphological boundaries
Errors are computationally problematic (e.g., van Rooy & Schäfer 2002, de Felice & Pulman 2008)
Goal: Provide taxonomy & annotation scheme for Korean learner spelling errors
Useful for automatic error diagnosis
Useful for feedback instruction in a Korean ICALL system

Overview
Background on Korean
Korean writing system
Korean word formation
Error taxonomy
Corpus Study
Interannotator agreement with pilot corpus
Abilities of current spelling checkers
Spelling Error Analysis with 100 Learner corpus

Korean writing system
Syllabic alphabet with one umjeol ‘syllable’, as a basic unit
A syllable is composed of at most 3 components:
- choseong, the first sound, e.g., [h]
- jungseong, the second sound(s), e.g., [a]
- jongseong, the last sound(s), e.g., [k]
- Syllable is represented combinatorically, e.g., [hak]
  unlike the more linear representation for English

Korean word formation
As an agglutinative language, word formation in Korean has complex morphological combinations
- Morphemic boundaries tend to be maintained in spite of application of a phonological rule.
- As 졅-으시-었-것-습니다 japon-eus-eoss-kess-seupnita
  Pronounced: ja-beu-si-eo-kkey-sseum-ni-da
- Composed of 5 morphemes & orthographically maintains the basic dictionary form for each morpheme

Syllabic representation of the Korean writing system requires learners to acquire:
- Linguistic knowledge of exact syllable compositions
  - e.g., for [hakki] ‘semester’:
    - the first [k] combines with the preceding vowel [a]
    - the second [a] attaches to the following vowel [i]
- Specific orthographic knowledge, including sound-letter relationships
  - e.g., [k] and [g] sounds both correspond to [k]
- Positional constraints of letter patterns
  - e.g., double consonant [n] does not appear word initially
Korean word formation

- With a lack of linguistic awareness for morphological combinations, Korean language learners tend to depend on sound.
- Errors stem from lack of morphological knowledge.
- For English, phonological confusion plays a crucial role in error production (Hovermale, 2008).

Classifying errors

- Spelling errors range from simple mistakes to more linguistically complex errors (Kukich, 1992).
- Mistakes result from inattention or physical conditions, including typos.
- Systematic errors reflect a lack of linguistic knowledge & require more informative feedback.
- Lack of phonological awareness: phoneme discrimination, identification, or segmentation.
- Lack of morphological awareness: morpheme identification or segmentation (words, particles, inflected verbal endings).

Error taxonomy

- An annotated corpus analysis can show actual range of spelling errors of Korean learners.
- How each type of error is related to linguistic knowledge.
- Related to learner’s native language.
- Related to deficit of phonological/morphological knowledge of Korean (cf., e.g., Rimrott & Heft, 2008).
- We classify 5 categories of spelling errors:
  - Phonological, morphological, typographical, incomprehensible, foreign word.
  - Similar to Hovermale (2008), except errors with foreign words are separately marked.

Phonological errors

- Phonological errors are based on:
  - Incorrect mapping between a sound & a letter.
    - Consonant mismatch.
      - Plain-tense-aspirated: [p, p', pʰ]; [k, k', kʰ]; [t, t', tʰ]
      - Ex. correct form: 예쁨니다 incorrect form: 예웁니다.
    - Vowel mismatch.
      - Ex. correct form: 남씨가 incorrect form: 남시가
    - Ex. correct form: 노래 incorrect form: 녀래
- These types of errors are generally restricted to differences between Korean & English.

Morphological errors

- Morphological errors include:
  - Failures of morpheme identification.
    - Example: 막아+으러다 [matʌ+kə] correct form: 막아옵니다 Incorrect form: 막어요
  - Overgeneralizations.
- These errors are related to inflection, word syllabification, & syllable boundaries.

Phonological distance between sounds can be remote.

- Consonants: [ŋ[k]] vs. [ŋ[p]]
- Example: correct [wesungmo] vs. incorrect [wesupmo]
- Vowels: [a] vs. [eu]
- Example: correct [’app] vs. incorrect [’app]
- This perceptual confusion is rare in native Korean.
Foreign word errors

- Foreign words in Korean borrowed from other languages often have non-predictable spelling
  - For example, the proper name New York
    - correct standard form: 뉴욕 nyuyok
    - commonly learner innovation: 뉴요크 nyuyokhu
  - Closely related to phonological confusion
    - But hard to determine the exact match between a sound and a letter or identify the exact phonological rule

Typographical and incomprehensible errors

- Typographical errors
  - Criterion: student marked it correctly at other points

- Incomprehensible errors
  - It was unclear what was meant

- Both kinds of errors are similar to native speakers errors

Interannotator agreement with pilot corpus

- We evaluated the Kappa statistic to measure interannotator agreement
  1) Error Type: $K=0.83 / 1281$ pairs
     - $P$ for phonological, $M$ for Morphological, etc.
  2) Correction: $K=0.73 / 1281$ pairs
     - Wrong word: Correct word pairs
     - e.g., 아름답습니다: 아름답습니다 --> to be beautiful
  3) Feedback: $K=0.75 / 1281$ pairs
     - e.g., $u:u$ --> consonant $u$ (p) should be replaced with consonant $u$ (m)

- All 3 scores show high correlation between 2 annotators: positive results for our guidelines

Distinguishing phonological & morphological errors

- Morphological variations caused by phonological rules have been treated as morphological errors
  - Examples.
    - Assimilation: correct form 웰래 incorrect form 웰례

- Phonological variants:
  - correct form부모님과 [kwa] incorrect form 부모님의 [wa]
  - correct form 한국말을 [eul] incorrect form 한국말을 [teul]

- Sometimes phonological variation disappears and become morphological variations.
  - Example. Sound distinction loss among native speakers
    - $h[ae]$ vs. $h[e]$

- All 3 scores show high correlation between 2 annotators: positive results for our guidelines

Our approach

- Pilot study: Gather a small learner corpus of 10 people
  - 10 non-heritage intermediate learners with 1 article each
  - Test interannotator agreement for spelling annotation
  - Test accuracy of existing spelling checkers

- Actual annotation of data for 100 people
  - 25 heritage beginners, 25 non-heritage beginners
  - 25 heritage intermediates, 25 non-heritage intermediates

Spelling checking

- Spell checkers for Korean do not adequately handle learner errors
  - Hypothesis: more morphological in nature
  - Closely related with word-spacing errors
  - Learner errors:
    - require different error diagnostic tactics
    - need to support feedback

- Two spelling checkers used
  - HWP (Korean Word Processor)
  - HAM v.5 (Hangeul Analysis Module)
Accuracy of spelling checkers with a Pilot Corpus

- HWP (Korean word Processor)
  - Raw Corpus
    - Precision: 0.717
    - Recall: 0.551
    - F-measure: 0.623
  - After Word-spacing Correction
    - Precision: 0.927
    - Recall: 0.551
    - F-measure: 0.691

- HAM v.5
  - Raw Corpus
    - Precision: 0.500
    - Recall: 0.667
    - F-measure: 0.571
  - After Word-spacing Correction
    - Precision: 0.611
    - Recall: 0.638
    - F-measure: 0.624

Corpus of 100 Learners

- Heritage Beginner: 2,669 words
- Heritage Intermediate: 2,496 words
- Non-heritage Beginner: 1,659 words
- Non-heritage Intermediate: 3,163 words

25 learners in each group

Basic statistics

<table>
<thead>
<tr>
<th></th>
<th>H1</th>
<th>H2</th>
<th>F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>26</td>
<td>10</td>
<td>26</td>
<td>16</td>
</tr>
<tr>
<td>I</td>
<td>10</td>
<td>19</td>
<td>19</td>
<td>12</td>
</tr>
<tr>
<td>M</td>
<td>266</td>
<td>168</td>
<td>207</td>
<td>236</td>
</tr>
<tr>
<td>P</td>
<td>289</td>
<td>150</td>
<td>203</td>
<td>560</td>
</tr>
<tr>
<td>T</td>
<td>9</td>
<td>3</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>366</td>
<td>966</td>
<td></td>
</tr>
</tbody>
</table>

Analysis by background

<table>
<thead>
<tr>
<th></th>
<th>Heritage</th>
<th>Non-heritage</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
<td>#</td>
</tr>
<tr>
<td>F</td>
<td>26</td>
<td>4.33</td>
<td>26</td>
</tr>
<tr>
<td>I</td>
<td>10</td>
<td>1.67</td>
<td>19</td>
</tr>
<tr>
<td>M</td>
<td>266</td>
<td>44.33</td>
<td>168</td>
</tr>
<tr>
<td>P</td>
<td>289</td>
<td>48.17</td>
<td>150</td>
</tr>
<tr>
<td>T</td>
<td>9</td>
<td>1.50</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>366</td>
<td>966</td>
</tr>
</tbody>
</table>

Analysis by language level

<table>
<thead>
<tr>
<th></th>
<th>Beginner</th>
<th>Intermediate</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
<td>#</td>
</tr>
<tr>
<td>F</td>
<td>36</td>
<td>6.91</td>
<td>16</td>
</tr>
<tr>
<td>I</td>
<td>17</td>
<td>3.26</td>
<td>12</td>
</tr>
<tr>
<td>M</td>
<td>227</td>
<td>43.57</td>
<td>207</td>
</tr>
<tr>
<td>P</td>
<td>236</td>
<td>45.30</td>
<td>203</td>
</tr>
<tr>
<td>T</td>
<td>5</td>
<td>0.96</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>521</td>
<td>445</td>
<td>966</td>
</tr>
</tbody>
</table>
Spelling checking (HWP) on corpus of 100 learners

<table>
<thead>
<tr>
<th></th>
<th>H1</th>
<th>H2</th>
<th>F1</th>
<th>F2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>Checked w/ C</td>
<td>77</td>
<td>22.19</td>
<td>67</td>
<td>26.91</td>
<td>34</td>
</tr>
<tr>
<td>Checked w/o C</td>
<td>184</td>
<td>53.03</td>
<td>95</td>
<td>38.15</td>
<td>70</td>
</tr>
<tr>
<td>Not checked</td>
<td>86</td>
<td>24.78</td>
<td>87</td>
<td>34.94</td>
<td>54</td>
</tr>
</tbody>
</table>

Analysis of Phonological Errors

<table>
<thead>
<tr>
<th># of Error Types</th>
<th>H1</th>
<th>H2</th>
<th>F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consonants</td>
<td>17</td>
<td>24</td>
<td>17</td>
<td>30</td>
</tr>
<tr>
<td>Vowels</td>
<td>25</td>
<td>34</td>
<td>24</td>
<td>21</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>58</td>
<td>41</td>
<td>51</td>
</tr>
</tbody>
</table>

Phonological Error Samples: Consonants & Vowels

<table>
<thead>
<tr>
<th></th>
<th>H1</th>
<th>H2</th>
<th>F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
<td>#</td>
</tr>
<tr>
<td>H : ñ</td>
<td>5</td>
<td>11.90</td>
<td>30</td>
<td>34.09</td>
</tr>
<tr>
<td>ɾ : -</td>
<td>9</td>
<td>21.43</td>
<td>10</td>
<td>11.36</td>
</tr>
<tr>
<td>Ŋ : ñ</td>
<td>4</td>
<td>9.52</td>
<td>1</td>
<td>1.14</td>
</tr>
<tr>
<td>Total</td>
<td>42.86</td>
<td>46.59</td>
<td>44.44</td>
<td>24.24</td>
</tr>
</tbody>
</table>

Challenges

- For automatic spelling checking:
  - Need to account for learner errors, by a better morpheme:phoneme mapping?
- For corpus annotation development:
  - Collect & annotate more data
  - Integrate annotation with other types of annotation
  - Use multi-layered annotation (cf. Lüdeling et al. 2005)

Conclusions

- With this annotated corpus:
  - Can improve evaluation for particle error detection (Lee, Eom, & Dickinson 2009)
  - Can test several methods of spelling error detection to determine their effectiveness for each error type
  - Methods involving POS tagging, string similarity, machine learning, etc.
- Future: conduct experiments to find out the most effective feedback for each type of spelling errors

Acknowledgements

Thanks to the following annotators:

- Sung-eun Choi (BYU)
- Jae-young Song (Wellesley)

Thanks for the support from:

- Brigham Young University
- Wellesley College


