Pronunciation and Spelling: the Case of Misspellings in Swedish L2 Written Essays

Gintarė GRIGONYTĖ a,1 and Björn HAMMARBERG a
aDepartment of Linguistics, Stockholm University, Sweden

Abstract. This research presents an investigation performed on the ASU corpus. We analyse to what extent does the pronunciation of intended words reflects in spelling errors done by L2 Swedish learners. We also propose a method that helps to automatically discriminate the misspellings affected by pronunciation from other types of misspellings.

Keywords. L2, misspellings, ASU corpus, written text, Swedish

Introduction

Misspellings normally occur due to typing errors or the lack of knowledge of the correct spelling. Presumably, when not being certain and trying to reproduce the correct spelling of a word, second language learners might also need to rely on the pronunciation of that word (among all possible sources which are relevant to the decision making of the spelling of that word).

In general a misspelled word can result in a string of characters that either represents: a) an out-of-dictionary word, i.e. incorrectly spelled word (like ”erata” instead of ”errata”); or b) a correct spelling of another word - heterograph (like ”too” instead of ”two”). In this paper we focus on the former type of misspellings.

The misspellings of L2 learners have received a lot of attention in L1 identification task. [1] suggest that spelling of L1 transfers to L2 and thus spelling errors in L2 can be helpful for automatically identifying writer’s native language. Based on the assumption that spelling errors are related to pronunciation in L1 authors have singled out 8 types of spelling errors and used them as features for L1 identification. In the Natural language Identification task 2013 [2] has shown that cognate and misspelling features are among the most important ones when identifying writer’s L1. In their study [2] have relied on orthography of misspelled words and mapped them with intended word in L2 and L1 cognates.

[3] demonstrated a 66% accuracy in language classification task by relying only on 200 most frequent character bigrams. Which was interpreted as the correlation between

1Corresponding Author: Gintarė Grigonytė, Department of Linguistics, Stockholm University, SE-10691 Stockholm, Sweden; E-mail: gintare@ling.su.se
L1 and the choice of words in L2 texts. Related to this study, a negative result has been reported by [4] showing that majority of indicative bigrams are independent of L1.

Another related research area to this paper is L2 error tagging. [9] study describes automatic error tagging of spelling mistakes in learner corpora. [8] presents how spelling errors can be used for assessing L2 writing.

1. Data and Method

This study proposes an L1 independent methodology for analysing L2 misspellings and detecting the pronunciation effect in spelling. We use the Swedish spell-checker system described in [5] and the large Swedish dictionary with pronunciation information [6]. First, we detect out-of-dictionary words in L2 written texts and suggest the most likely correct spelling candidate. Second, we perform an automatic phonologically-based mapping between the original L2 word and the pronunciation of the intended spelling. Finally, we measure the distances between the two mappings: a) the original spelling and the pronunciation of the intended word, b) the spelling of the intended word and the pronunciation of the intended word; and rank misspellings according to the phonological spelling bias.

We use the part of written essays in Swedish, produced by adult learners of the ASU corpus [11]. The ASU Corpus is a longitudinal corpus of transcribed audio-recorded conversations and written texts collected from adult learners of Swedish and supplemented by a comparable material from native Swedes.

The learners’ written part of the ASU corpus data comprises of 220 text units (10 persons x 11 sessions x 2 texts) ca 50,000 word tokens. The data ranges from the beginner stage up to a level where Swedish learners are studying in Swedish at university level.

1.1. Detection of Out-of-dictionary Words and Suggestion of Intended Spelling

Initially we use the SALDO [10] dictionary to perform a dictionary check-up for detecting possible spelling errors. The approach to Swedish spelling correction is orthographic and is based on the phonetic similarity key method combined with a method to measure proximity between the strings [5]. We use the Edit distance algorithm to measure the proximity of orthographically possible candidates and the Soundex algorithm to shortlist the spelling candidates which are phonologically closest to the misspelled word. Further, the spelling correction candidates are analyzed in a context by using the SLME n-gram model.

The SLME employs the Google Web 1T 5-gram, 10 European Languages, Version 1, dataset for Swedish, which is the largest publicly available Swedish data resource. The SLME is a simple N-gram language model, based on the Stupid Backoff Model [7]. The n-gram language model calculates the probability of a word in a given context Eq.(1). The maximum-likelihood probability estimates for the n-grams are calculated by their relative frequencies Eq.(2).

\[
P(w^L_i) = \prod_{i=1}^{L} P(w_i|w^{i-1}) \approx \prod_{i=1}^{L} \hat{P}(w_i|w^{i-n+1})
\] (1)
\[ r(w_i | w_{i-1}^{i+n-1}) = \frac{f(w_{i-n+1}^{i})}{f(w_{i-1}^{i+n-1})} \]  

(2)

The SLME n-gram model calculates the probability of a word in a given context: \( p(\text{word}|\text{context}) \) (Table 1). The highest probability determines the spelling correction.

<table>
<thead>
<tr>
<th>Original word</th>
<th>Intended words</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>försöka att förstå hela menningen</td>
<td>meningen (En. sentence)</td>
<td>5.06e-05</td>
</tr>
<tr>
<td>(En. try to understand all sentence)</td>
<td>myningen (En. mouth)</td>
<td>4.71e-09</td>
</tr>
</tbody>
</table>

### 1.2. Phonological Comparison of Original Spelling and Intended Spelling

In order to detect whether phonology has an effect on spelling mistakes we use the pronunciation information from the Swedish dictionary [6]. The pronunciation data encoded in the SAMPA format is simplified by: discarding markers of syllable boundaries, special punctuation characters (e.g. vet [ve:t] - [vet]), disregarding upper/lowercase characters (e.g. rätt [rEt] - [ret]), and replacing codes of numeric allophonic variants to the closest single or double alphabet characters (e.g. för [f9:r] - [fior]). This robust conversion is not ideal and looses subtle phonological information in the process, however it provides simple phonological transcriptions to the word-alike orthographic forms of the intended spellings. The simplified orthographic transcription of the intended spelling is then measured for the string similarity between the misspelled and the intended word (Table 2).

<table>
<thead>
<tr>
<th>Original word</th>
<th>Intended word</th>
<th>Pronunciation</th>
<th>SimOP</th>
<th>SimIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>snab</td>
<td>snabb</td>
<td>[snab]</td>
<td>1</td>
<td>0.889</td>
</tr>
<tr>
<td>ihog</td>
<td>ihåg</td>
<td>[ihog]</td>
<td>1</td>
<td>0.750</td>
</tr>
<tr>
<td>respektuos</td>
<td>respektlös</td>
<td>[respektlös]</td>
<td>0.857</td>
<td>0.857</td>
</tr>
<tr>
<td>interesserad</td>
<td>intresserad</td>
<td>[intresserad]</td>
<td>0.909</td>
<td>0.952</td>
</tr>
</tbody>
</table>

### 2. Analysis of the Results

We run the automatic procedure described in section 1.1. on the ASU corpus and detected 1677 out-of-dictionary words. The list was pruned by removing person and geographical names, also insertion words of the code switching (e.g., beef, buy, body, class, clean, etc.). After this step we were left with 1141 misspelled words with spelling corrections (the density of spelling errors in the ASU corpus is 0.023%).

For the phonological comparison described in section 1.2. we have used 725 unique pairs of misspellings and their spelling corrections (source 1411 misspelled words). We measured similarities between a) original spelling and pronunciation of intended spelling (SimOP) and b) intended spelling and pronunciation of intended spelling (SimIP).

All 725 spelling corrections were manually checked by a Swedish native speaker. Misspellings were annotated as phonologically (pronunciation) influenced (e.g. snell - snall), orthographically influenced (e.g. tva - två), influenced by cognate words from
other languages than Swedish (e.g. discotek - diskotek) or random spelling errors with no obvious motivation (e.g. upplysning - upplysning). The annotations also included learners’ own (invented) lexical and grammatical solutions which result in target-deviant word forms and hence not found in the dictionary, provided that they are correctly spelled according to Swedish orthography (e.g. motsidig - ömsesidig, sittude - satt).

The annotations were used as the gold standard to check whether it is possible to automatically differentiate between misspellings influenced by the pronunciation of the intended Swedish word and other types of misspellings. We have used the arithmetic difference between SimOP and SimIP as a binary classifier: if the difference was positive, the misspelling was categorized as phonological. The evaluation against the manually annotated 158 pronunciation related errors has shown 0.75 precision and 0.57 recall.

3. Conclusions and Future Work

The ASU corpus investigation has shown that around 21% of spelling errors are influenced by the pronunciation of intended Swedish words. We have attempted to model this phenomena and proposed the L1-independent method that combines Swedish pronunciation data and the statistical language model. The evaluation of the binary classification between the phonologically affected and other spelling errors has shown an encouraging result, however the large proportion of false negative classifications suggests that intended word and the pronunciation alone is not entirely sufficient for making such a distinction. The future work directions could include distinction and separate modeling of a) misspellings due to discrepancies between spelling and pronunciation in the target language (i.e. standard written and spoken Swedish), and b) misspellings due to differences between the target phonology and the learners interpretation of the pronunciation.

References