Non-Native Writers' Errors – a Challenge to a Spell-Checker

Björn Hammarberg and Gintarė Grigonytė
Department of Linguistics, Stockholm University, Sweden
ham@ling.su.se, gintare@ling.su.se

Abstract

Spell checkers are widely used and if they do their job properly are also highly useful. Usually they are built on the assumption that the text to be corrected is written by a mature native speaker. However non-native speakers are in an even greater need of using spell checkers than native speakers. On the other hand current spell checkers do not take the linguistic problems of learners into account and thus they are poor in identifying errors and supplying the adequate corrections. There is a number of linguistic complexities specific to non-native learners that a spell-checker would need to handle in order to be successful.

1. Introduction

This extended abstract was motivated by the earlier study (Grigonytė and Hammarberg, 2014) on the impact of pronunciation factors on the spelling by non-native writers of Swedish. Our findings have shown that a significant portion of misspellings were conditioned by pronunciation factors and therefore created problems for the automatic spell-checker described in the same paper.

The purpose of this paper is to give examples of complexities that a spell-checker needs to handle in the case of non-native writers.

Our data comes from the analysis of the out-of-dictionary words found in a corpus of learner's Swedish, the ASU Corpus (Hammarberg, 2010).

2. ASU Corpus

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.

We use the part of written essays in Swedish, produced by adult learners of the ASU corpus. The learners' written part of the ASU corpus data comprises of 220 text units (10 persons x 11 sessions x 2 texts) totalling 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.

3. Spell-checker prototype

Initially we use the SALDO (Borin et al., 2008) 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. 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 (Brants (2007), Östling (2012)). The n-gram language model calculates the probability of a word in a given context. The maximum-likelihood probability estimates for the n-grams are calculated by their relative frequencies.

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

Original word	**Intended word**	**Probability in given context**
försöka att förstå hela **menningen**	meningen *(En. sentence)*	5.06e-05
(En. try to understand all senntence)	mynningnen *(En. mouth)*	4.71e-09

Table 1. An example case of the spelling correction for the word *menningen*.

4. Defining learners' errors

Native users and non-native learners display different problems. The picture is more variable, more complex, more ambiguous, often more difficult to analyse with non-native learners. Learners also differ with respect to stages of proficiency and types of learning conditions. In many cases the spell-checker that was relying on the edit-distance method and local context has failed to produce a relevant correction (Table 2.).

This is due to the fact that the errors of non-native speakers are caused by more complex factors than orthographic deviation of a string and the corrected word fitting or not to the context it stands in.

A distinction should be drawn between the definition of an error (i.e. determining the kind of deviance from norm) and the interpretation of the cause of the error.

The definition of the error determines the correction (i.e. determining what norm is violated). Often alternative corrections are possible and it may be difficult to rule out alternatives; hence a unique definition and correction may not be possible to establish.

Produced form	Intended word	Spell-checker correction
fallade	föll	fallande
asiker	åsikter	avsikter
entusiasthet	entusiasm	entusiaster

Table 2. Examples of inadequate spell-checker corrections.

Although determining the cause of error is logically distinct from defining the norm violation, it is often necessary to consider plausible causes in order to establish a plausible definition (and correction).

Causes may have to be sought in cross-linguistic influence or in target-language internal factors. Underlying problems are related to a variation of linguistic levels:

(a) lexical shape of words,

Produced form	Intended word	Spell-checker correction
motsidig	ömsesidig	motstridig
oilen	oljan	ollen

Table 3. Examples of lexical shape errors.

The first example is an incorrect lexical form made up hypothetically on the basis of the German form *gegenseitig*. The second example is formed on the basis of English *oil* and the Swedish definite suffix *-en*.

(b) morphology,

Produced form	Intended word	Spell-checker correction
sittade	satt	siktade
feodaliskt	feodalt	feodalism

Table 4. Examples of morphology errors.

The first example follows the regular pattern of creating a weak past form of the verb *sitta*. However *sitta* is a strong verb. The second example is an adjective formation with the suffix *–isk* which is not used with the adjective *feodal*.

(c) phonology-phonetics,

Produced form	Intended word	Spell-checker correction
flygan	flugan	flyga
hella	hela	hylla

Table 5. Examples of phonological errors.

These errors occur due to difficulties of Swedish phonology or the phonological basis for spelling.

(d) orthographical norms,

Produced form	Intended word	Spell-checker correction
lexion	lektion	legion
sommna	somna	somna

Table 6. Examples of orthographical errors.

These examples are reasonable forms from the point of view of lexical shape, morphology and phonology but they violate norms of how these particular words are spelled.

(e) code switching,

Produced form	Intended word	Spell-checker correction
fashion	fashion	fusion
exciting	exciting	excitering

Table 7. Examples of code switching.

The words in Table 7 are not intended to be Swedish in the first place, but are temporary switches from Swedish into another language, in this case English.

(f) random spelling or typing lapses,

Produced form	Intended word	Spell-checker correction
utblidningen	utbildningen	utbildningen
envenemanger	evenemang	evenemangen

Table 8. Examples of random spelling errors.

(g) very often multiple errors due to various causes appear in the same word,

Produced form	Intended word	Spell-checker correction
skylnat	skillnad	skyltat
forberedat	förberett	förbereda

Table 9. Examples of words with multiple errors.

These examples contain several phonological and morphological errors.

(h) strongly deviant words which are difficult to interpret unless the context makes it clear.

Produced form	Intended word	Spell-checker correction
verstao	förstår	värsta
showte	duschade	shorts

Table 10. Examples of extreme deviations.

To interpret examples like the ones in Table 10, it is necessary to inspect a wider context than one sentence/paragraph. This poses a problem for automatic spell-checkers.

5. Discussion

A basic problem with the types of errors shown in the tables above is that they require linguistic information that an automatic spell-checker does not possess. For the purpose of checking text written by a non-native writer, we envisage a semi-automatic checking and correcting method based on knowledge of a number of well-known and frequent morphological, phonological and orthographic problems in learner Swedish, large enough to capture an essential part of the errors that usually occur in Swedish written learner production.

We suggest three main strategies that need to be taken into account in order to deal with L2 errors and extend limitations of the prototype spell-checker.

Distinction between languages. Code switching cases pose the challenge of language distinction in L2 learner's text. Methods that deal with language detection (Lui et al., 2014) could be beneficial for discriminating between L2 and L1 words.

Dealing with morphological errors. For instance morphological errors like strong verb conjugation errors. These errors could be treated with the help of morphological segmenters (Grönroos et al., 2014) and would cover the following issues:

- verb conjugation
- noun and adjective inflection

For instance, the morphological error *sittade* in Table 4 can be discovered by recognizing the word-final string *ade* as a past tense suffix and have the correct conjugation being picked on the basis of the root.

Dealing with phonological errors. Phonological errors like failure of distinguishing separate phonemes or segment quantity i.e., long vs. short, and orthographic conventions i.e. like the doubling of the consonants after the stressed short vowel, could hypothetically be solved by, hard-wired rules, e.g.:

- double consonants
- u-y
- o-u

The replacement of *flygan* by *flugan* in Table 5 is a likely correction because of the well-known tendency to confuse the vowel phonemes pronounced.

If fact many of these phonological errors are common to writers of different language backgrounds. Errors which are frequent with L2 language learners are such because they reflect problems which are frequent with many language backgrounds. Normally frequent errors are those that mark typologically unusual target language structures. We argue that if one can capture large enough error set of L2 learners, it is possible to cover a large proportion of L2 errors indifferent to L1 backgrounds.

Finally in the case of checking and correcting the non-native text the semi-automatic procedure would be more sensible. As with the earlier spell-checker that we have used, the tool has access to a dictionary and reacts to each word form not found in the dictionary by automatically selecting the most likely correction. Preferably a L2 word correction tool could present the text sentence by sentence for manual inspection, and for each deviant word one or more alternative suggestions for corrections could be presented. For a human reader who knows the target language, the relevant alternative will usually be easy to recognize, and the choice can be made with a click. Although the procedure will certainly require some time, it can be beneficial for L2 error annotation and/or correction purposes.

6. Conclusions

We have presented various types of errors found in Swedish learners' texts. The analysis of such errors shows that not all deviations from the target written norm are *spelling* errors in a strict sense. Distinguishing between these types of errors and identifying what the deviation from the target norm is, is a challenge for a spell-checker in order to provide a relevant correction.

Reference

L. Borin, M. Forsberg and L. Linngren. 2008. The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology. Studia Linguistica Upsaliensia, p.21–32.

G. Grigonytė and B. Hammarberg. 2014. Pronunciation and Spelling: The Case of Misspellings in Swedish L2 Written Essays, In proceedings of the 6th Baltic HLT conference, IOS Press. p. 95-98.

Grönroos, S.-A., Virpioja, S., Smit, P., and Kurimo, M. 2014. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In proceedings of the 25th International Conference on Computational Linguistics, p. 1177-1185.

B. Hammarberg. 2010. Introduction to the ASU Corpus, available from: http://spraakbanken.gu.se/swe/itg

M. Lui, J.H. Lau, T. Baldwin. 2014. Automatic Detection and Language Identification of Multilingual Documents. TACL 2: 27-40.

R.Östling, SLME tool inspired by the Stupid Backoff Model (Brants et al., 2007), (2012), http://www.ling.su.se/english/nlp/tools/slme