Developing an Annotation Scheme for ELL Spelling Errors

DJ Hovermale and Scott Martin
Department of Linguistics
The Ohio State University
Columbus, OH
(djh|scott)@ling.osu.edu

Abstract

This paper describes an XML annotation scheme for English Language Learner (ELL) spelling errors in learner corpora which can be used to create standardized, annotated ELL error corpora for use by researchers who are developing spelling correction tools for ELLs. We also provide an error taxonomy (with examples of each error type) upon which the scheme was based.

1 Introduction

In 1997 there were 375 million native English speakers and 750 million people who spoke English as a second language (Crystal 1997). Over the past ten years there has been much development in the way of tools for these English Language Learners (ELLs). An increasingly large amount of energy and resources has been invested into developing Natural Language Processing (NLP) tools for Intelligent Computer-Aided Language Learning (ICALL). These NLP tools provide error diagnosis and feedback to individual users on a more frequent basis than tutoring from instructors is available. These tools often rely on the output of parsers, part-of-speech taggers, content matching modules, etc., which are unable to properly process input containing spelling errors. For instance, De Felice and Pulman (2008) analyze preposition usage by context, a process which is thwarted by misspellings. Nagata's (2002) Robo-Sensei tutor flags words that are not in its lexicon and asks users to correct the spelling errors themselves. Even outside of the context of ICALL there are many companies and researchers who are developing tools for automatically processing input from ELLs (c.f.e.g. Gamon et al. 2008). In developing these tools, researchers generally correct the spelling errors by hand and continue with development, leaving spelling correction to commercial spell checkers. However, the effectiveness of these NLP tools is limited when being used by actual learners, who often make spelling errors that commercial spell checkers, which are geared toward native speakers, are simply not equipped to correct. In fact, because the errors made by learners differ so much from those made by native speakers, more than one-third of spelling errors made by learners are not corrected by commercial spelling correction programs (Hovermale 2008).

2 Obstacles to ELL spelling correction research

Despite the pressing need for spelling correction which targets ELL errors, there are many obstacles faced by researchers who wish to work on this problem. First of all, although there are many learner corpora available, not all spelling errors made by learners are detectable automatically, and therefore must be labeled by hand. This process is very expensive and time-consuming. Secondly, there is currently no standard for annotating these errors, so researchers cannot efficiently share the efforts they put in to labeling the spelling errors in learner corpora. In ICALL applications diagnosis and feedback are just as important as providing the correct spelling. This poses a problem for computer scientists and computational linguists without a background in Second Language Acquisition who could potentially contribute to creating a spelling correction program which performs better on ELL errors. Lastly, although there are some errors that are made consistently by ELLs regardless of their first language, there are many phonological confusions that are specific to certain languages and/or language families. Spelling correction programs should take this into account for maximum effectiveness, which means handwriting rules for each individual language/language family - another process balked at by most researchers.

However, if learner English corpora were available that had the spelling errors annotated according to a standard annotation scheme, then many of these obstacles to developing ELL spelling correction tools would be eliminated. Researchers would be able to use tools to automatically detect the annotated errors, eliminating the need to hand-label their own data. Where the target word is clear it can be provided for proper evaluation of ELL spelling correction tools. The ideal annotation scheme would also provide information about error type which could be used for diagnosis and feedback. If there were a standard annotation scheme, then corpora annotated according to this scheme could be used as a gold standard for evaluation and comparison of various ELL spelling correction programs, as well as to measure incremental improvement of individual systems. Where the first language of the learners is available, it should be included in the annotation, which might provide the possibility to learn common errors made by individual learner populations by automatic means. These corpora would also provide statistics such as the frequency of each error type, which could prove very helpful in allocating resources, as high frequency error types could be targeted before less common ones.

3 Annotation Scheme

Our annotation scheme is formalized in an XML Document Type Declaration (DTD), and can be found in Appendix A. The hierarchical structure of XML allows several target corrections for each error, and several possible diagnoses for each target correction. Unfortunately, the task definition module of Callisto[1] only allows XML data with a single layer of structure, which precludes the DTD described here. The details of our XML annotation structure are as follows.

At the root element level, an `<annotation>` element encapsulates all the errors described by a given annotator for some text. Its attributes include the learner's native language (L1), the learner's proficiency level, and a unique identifier for the annotator. These are attributes that should be consistent within the entire document being annotated. The unique annotator identifier attribute is the only required attribute of the `<annotation>` element, as the leaner L1 and proficiency level are not available for all corpora. An `<annotation>` contains a series of `<error>` elements, each of which describes (in an attribute) whether the error is a non-word or a real-word error.

For each error the learner had in mind a specific target word. However, this target is not always apparent to the annotator, and in some cases there may be more than one feasible target word for a given learner production. For this reason any number of target corrections can be specified containing the text of the correction followed by a series of probable reasons (diagnoses) the annotator believes the error occurred. When the annotator has evidence for a preferring a specific target word over another the most probable target can be specified by means of the "rank" attribute in the `<target>` element. There may be instances when there is not a single clear candidate for a target, such as the following sentence:

(1) *Do you smell that anything is biring*[2]*?*

Were it not for the mere fact that this was a sentence from a translation task, the underlined word, *biring,* would have no clear target, and therefore the `<target>` element could be omitted. This is also the case when the context of the error provides no help in determining the target.

The optional `<how_determined>` element is a child of the `<target>` element, and allows the annotator to describe why they selected a specific target word. For instance, in (1) above we know what the target is because we have the target sentence from the translation task. This is not a luxury that is afforded in every corpus, but when we do have this information we should include it.

Although the error taxonomy suggests a finite number of diagnoses for ELL spelling errors, our DTD does not contain an exhaustive list of diagnoses, so that new diagnosis types can be added flexibly. Some example diagnosis types can be found in the next section, which contains a suggested error taxonomy that can be represented by our annotation scheme.

[1] The Callisto annotation tool is available for download at http://callisto.mitre.org/index.html

[2] All examples are taken from the HELC-2 Corpus unless otherwise specified.

4 Error Taxonomy

4.1 Non-word vs real-word errors

As stated in the previous section, the distinction between non-word and real-word errors is simply being explicit about whether the word is contained in the dictionary of the researcher's spellchecker or not. This is an important distinction to make when dealing with spelling correction, as detection of errors is treated as a separate task from correction of errors. Often spelling correction programs do not detect any errors that result in a learner production that is in the dictionary of the spellchecker. Making this distinction allows accurate reporting on the detection of words that are contained in the dictionary of the spellchecker currently being used by the researcher.

4.2 Non-word child elements

4.2.1 Morphological Overregularization

Morphological Overregularization (Marcus 1996) is a common error among ELLs. While native speakers of English typically only make such mistakes during first language acquisition, ESL learners frequently make such errors (Paradis 2005).

The HELC-2 Corpus is a collection of approximately 2400 unique learner responses (most of which are complete sentences) to a translation task. The students were given a sentence in Japanese and asked to translate it into English. Of the approximately 500 ELL spelling errors in the corpus, roughly 10% were morphological overregularizations. These can be split into subcategories based on which morpheme (or class of morpheme) is being overregularized. The data from the HELC-2 corpus suggest the following distinctions:

(2) *PAST - The speech contest will be <u>holded</u> on Saturday of this week.* /held

(3) *DERIVATIONAL- I'm sorry for my <u>latery</u>.* /lateness

We propose that this be an open class, as other forms of morphological regularizations can be imagined, for example, *PLURAL* (i.e. **childs* for children or **mans* for men).

4.2.2 Phonological Confusion

The phonology of the learner's L1 can affect the way that they spell words in English, usually resulting in words that are out of the vocabulary of the spellchecker. These words should be set apart from simple typos, so we label them with the diagnosis "phonological confusion" where possible. Some researchers are already working on L1 specific spell checkers for English (Mitton and Okada 2007; Hovermale 2008) and this information is vital to the development of such tools. An example of a phonological confusion resulting in a non-word is listed below:

(4) *I'm <u>solly</u> too late .* /sorry

This should be classified as a phonological confusion rather than a typo because this is a documented phonological confusion of Japanese learners of English (Hansen and Arslan 1995).

4.2.3 Typographical Error

There are some errors that are not distinct from the errors which are made by native speakers of English. We simply label these as "typographical errors" so that all errors in the corpus are labeled, thereby allowing researchers who are developing spelling correction tools which target learner errors to also monitor their performance on errors which are commonly made by native speakers. An example is provided below:

(5) *I have <u>sorked</u> all time since six.* /worked

This should be labeled as a typographical error because confusion of the phonemes [s] and [w] is not typical of Japanese learners.

4.2.4 Unknown

This diagnosis is generally reserved for instances when the target word is unclear, however, there may be cases where a target word is very probable from the context, but no diagnosis seems to fit. Consider the following example, where the underlined word is very probably supposed to be '*please*', but the reason the learner produced the underlined word does not fit with any of our diagnoses:

(6) *Tell me What should I wear to go to the perty, <u>plcne</u>.* /please

Or this example, where there is no clear target:

(7) *Please tell me inhe the train at a time.* /????

It is unreasonable to expect any spellchecking tools to correct the overwhelming majority of these types of errors. These errors should therefore be distinguished from normal typographical mistakes, as outlined in previous work on native English speaker spelling errors (c.f.e.g. Damerau 1964; Pollock and Zamora 1984).

4.3 Real-word child elements

4.3.1 Homophone

These errors are the result of confusing words that sound alike in standard English. We used the Carnegie Mellon Pronouncing Dictionary[3] to identify approximately 13,000 sets of spellings with shared pronunciations. If a real-word error shares a pronunciation with the target word, then it is determined to be a homophone in English and labeled as such. Examples follow:

(8) *I'm solly too late.* /to

(9) *He said to me I came Japan 10 year's ago.* /years

(10) *He said to me, "Ten years have past since I came to Japan."* /passed

(11) *Please tell me what to ware for the party.* /wear

4.3.2 Phonological Confusion

As stated above, phonological confusions often result in words that are not in the dictionary of the spellchecker. There are instances, however, when these phonological confusions result in words which *are* in the spellchecker's list of words, making them impossible to detect by a standard dictionary lookup. These should be kept distinct from the "homophones" spoken of in section 4.3.1, since these errors are the result of phonological confusions, and are therefore specific to learners. These are labeled as "real-word" errors and the diagnosis of "phonological confusion" is applied. Several examples follow:

(12) *He sat with looking at the liver.* /river

(13) *I heard a bard singing.* /bird

(14) *I heart that a bird was singing.* /heard

4.3.3 Typographical Error

Typos can sometimes result in words which are in the spellchecker's dictionary. When a real-word error does not share a pronunciation with the target word and is not a phonological confusion it is labeled as a typographical error. Examples follow:

(15) *I saw then enter the restaurant.* /them

(16) *He set seeing the river.* /sat

5 Conclusion

In this paper we have set forth a scheme for annotating ELL spelling errors in learner corpora. We have also suggested an error taxonomy based on the errors found in the HELC-2 corpus. While we recognize that this annotation scheme is not perfect, we do believe that it is a viable starting point and a key step in creating a standard for the annotation of ELL spelling errors.

6 Future Work

Despite this paper containing what we feel to be a fairly clear description of this annotation scheme, detailed guidelines for annotation are still needed.

Because of the nature of the various learner corpora available and the copyrights involved with them, a standalone version of the annotation scheme needs to be created for use with corpora that cannot be redistributed directly with inline annotation.

There is currently no annotation tool known to the authors that is suitable for this task. Therefore, there is a need for an annotation tool that is able to represent the structure that we set forth in our DTD. Either the functionality of the Callisto annotation tools can be expanded, or a new annotation tool can be created

[3] The Carnegie Mellon University Pronouncing Dictionary is available for download at http://www.speech.cs.cmu.edu/cgi-bin/cmudict

References

Crystal, D. 1997. *Global English.* Cambridge University Press.

De Felice, R. and S. Pulman. 2008. *Automatic detection of preposition errors in learner writing.* ICALL Special Interest Group pre-conference workshop, CALICO conference, March 18 –22.

Damerau, F. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3), 171-176.

Gamon, M., J. Gao, C. Brockett, A. Klementiev, W. B. Dolan, D. Belenko, and L. Vanderwende. 2008. *Using Contextual Speller Techniques and Language Modeling for ESL Error Correction.* In Proceedings of IJCNLP, Hyderabad, India.

Gui S. and H. Yang. 2003. *Zhongguo Xuexizhe Yingyu Yuliaoku. (Chinese Learner English Corpus*). Shanghai Waiyu Jiaoyu Chubanshe, Shanghai. (In Chinese).

Hansen, J. & Arslan, L. (1995). Foreign accent classification using source generator based prosodic features. *Proc. ICASSF, Detroit,* 1:836-839.

Hovermale, DJ. 2008. *SCALE: Spelling Correction Adapted for Learners of English.* ICALL Special Interest Group pre-conference workshop, CALICO conference, March 18 –22.

Marcus, G. 1996. Why do children say “breaked”? *Current Directions in Psych. Science.*, 5: 81-85.

Miura, S. *Hiroshima English Learners' Corpus No. 2*. Available online at http://home.hiroshima-u.ac.jp/d052121/eigo2.html

Mitton R., Okada T. 2007. The adaptation of an English spellchecker for Japanese writers, *Symposium on Second Language Writing*, Nagoya Gakuin University, Nagoya, Japan, 15-17 September

Nagata, N. 2002. BANZAI: An Application of Natural Language Processing to Web based Language Learning. *CALICO Journal*. 19(3), 583–599.

Paradis, J. 2005. Grammatical morphology in children learning English as a second language, in *Language, Speech, and Hearing Services in Schools*, 36: 172-187

Pollock, J. J., & Zamora, A. 1984. Automatic spelling correction in scientific and scholarly text. *Communications of the ACM*, 27(4), 358-368.

Appendix A - DTD

```
<!ELEMENT annotation (error+)>
<!ATTLIST annotation
    annotatorId CDATA #IMPLIED
    learner_level CDATA #IMPLIED
    learner_L1 CDATA #IMPLIED>

<!ELEMENT error (learner_production, target+)>
<!ATTLIST error status (non-word|real-word) #REQUIRED>

<!ELEMENT target (how_determined?, text, diagnosis+)>
<!ATTLIST target rank CDATA #IMPLIED>

<!ELEMENT how_determined (#PCDATA)>
<!ELEMENT text (#PCDATA)>

<!ELEMENT diagnosis EMPTY>
<!ATTLIST diagnosis type NMTOKEN #REQUIRED>
```

Appendix B - Examples

```
<?xml version="1.0" ?>
<annotation annotatorId="126" learner_level="intermediate" learner_L1="Japanese">

    <error status="nonword">
          <learner_production>kew</learner_production>
          <target>
                <how_determined>translation task target sentence</how_determined>
                <text>knew</text>
                <diagnosis type="typographical_error" />
          </target>
    </error>

    <error status="nonword">
          <learner_production>flied</learner_production>
          <target>
                <text>flies</text>
                <diagnosis type="typographical_error" />
          </target>
          <target>
                <text>flew</text>
                <diagnosis type="morph_overreg_past" />
          </target>
    </error>

    <error status="nonword">
          <learner_production>latery</learner_production>
          <target>
                <text>lateness</text>
                <diagnosis type="morph_overreg_deriv" />
          </target>
    </error>

    <error status="nonword">
          <learner_production>weare</learner_production>
          <target>
                <text>wear</text>
                <diagnosis type="typographical_error" />
                <diagnosis type="phonological_confusion" />
          </target>
    </error>

</annotation>
```