On using context for automatic correction of non-word misspellings in student essays

Michael Flor
Educational Testing Service
Rosedale Road
Princeton, NJ, 08541, USA
mflor@ets.org

Yoko Futagi
Educational Testing Service
Rosedale Road
Princeton, NJ, 08541, USA
yfutagi@ets.org

Abstract

In this paper we present a new spell-checking system that utilizes contextual information for automatic correction of non-word misspellings. The system is evaluated with a large corpus of essays written by native and non-native speakers of English to the writing prompts of high-stakes standardized tests (TOEFL® and GRE®). We also present comparative evaluations with Aspell and the speller from Microsoft Office 2007. Using context-informed re-ranking of candidate suggestions, our system exhibits superior error-correction results overall and also corrects errors generated by non-native English writers with almost same rate of success as it does for writers who are native English speakers.

1 Introduction

Misspellings are ubiquitous in student writing. Connors and Lunsford (1988) have found that spelling errors accounted for about one quarter of all errors found in a random sample of 300 student essays. Desmet and Balthazor (2006) found that spelling errors are among the five most frequent errors in first-year college composition of US students. Lunsford and Lunsford (2008) found that spelling errors constituted about 6.5% of all errors found in a US national sample of 3000 college composition essays, despite the fact that writers had access to spellcheckers.

Misspellings are even more ubiquitous in texts written by non-native speakers of English, especially English Language Learners (ELL). The types of misspellings produced by L2 writers are typically different from errors produced by native speakers (Hovermale, 2010; Al-Jarf, 2010; Okada, 2005).

In the area of automatic assessment of writing, detection of misspellings is utilized in computer-aided language learning applications and in some automatic scoring systems, especially when feedback to users is involved (Dikli, 2006; Warschauer and Ware, 2006). Yet spelling errors may have a deeper influence on automated text assessment. As noted by Nagata, et al. (2011), sub-optimal automatic detection of grammar and mechanics errors may be attributed to poor performance of NLP tools over noisy text.

Presence of spelling errors also hinders systems that require only lexical analysis of text (Landauer, et al. , 2003; Pérez, et al., 2004). Granger and Wynne (1999) have shown that spelling errors can affect automated estimates of lexical variation, which in turn are used as predictors of text quality (Crossley, et al., 2008; Yu, 2010). In the context of automated preposition and determiner error correction in L2 English, De Felice and Pulman (2008) noted that the process is often disrupted by misspellings. Futagi (2010) described how misspellings pose problems in development of a tool for detection of phraseological collocation errors.

Given this state of affairs, it is only natural for automatic text assessment systems to utilize automatic spellchecking components. However, generic spellcheckers are typically oriented for errors produced by writers who are native speakers of a language. Rimrott and Heift (2008, 2005) have demonstrated that a generic speller has poor performance on data from German language learners.

The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pages 105–115,
Montréal, Canada, June 3-8, 2012.

Bestgen and Granger (2011) and Hovermale (2010) have demonstrated similar results on data from ELL.

Many researchers have suggested that spell-checkers for L2 users need to be adapted for the particular patterns of errors that characterize each native language (L1), by studying patterns of interference and influence from L1 to L2 (Mitton and Okada, 2007; Mitton, 1996; Rimrott and Heift, 2008, 2005; Bestgen and Granger, 2011; Hovermale, 2010). We have set up to explore a different path, in the context of automated text assessment. Our goal in the present study is to examine to what extent detection and automatic correction of non-word misspellings can be improved by utilizing essay context, for data from both native and non-native English speakers.

The rest of this paper is organized as follows. Section 2 provides a description of the corpus of texts and misspellings that was used in this study. Section 3 describes the ConSpel automatic spell-checking system. Section 4 presents results from a comparative evaluation of our system, ConSpel, the popular Aspell speller and the Microsoft Office 2007 speller. Section 5 compares our findings with some recent studies and discusses implications for further development of automatic spell-checking systems.

2 Corpus

The corpus used in this study is a collection of essays, annotated for misspellings by trained annotators. It is developed for evaluation of automatic spellcheckers, and for research on patterns of misspellings produced by both native English speakers and ELL.

2.1 Texts

The corpus comprises essays written by examinees on the writing sections of GRE® (Graduate Record Examinations) and TOEFL® (Test of English as a Foreign Language) (ETS, 2011a,b). The TOEFL test includes two different writing tasks: a short opinion essay, on a pre-assigned topic, and a summary essay that compares arguments from two different sources (both supplied during the test). GRE also includes two different writing tasks: one is a short argumentative essay taking a position on an assigned topic, the other is an essay evaluating the soundness of arguments presented in prompt. Both tests are delivered on computer (at test centers around the world and via Internet), always using the standard English language computer keyboard (QWERTY). Editing tools such as a spellchecker are not provided in the test-delivery software (ETS, 2011a). All writing tasks have time constraints.

In the current phase of the project, the corpus includes 3000 essays, for a total of 963,428 words. The essays were selected equally from the two tests (4 tasks, 10 prompts per task, 75 essays per prompt), also covering full range of scores (as a proxy for English proficiency levels) for each task. The majority of essays in this collection were written by examinees for whom English is not the first language (98.73% of TOEFL essays, 57.86% of GRE essays).

2.2 Annotation

Each text was independently reviewed by two annotators, who are native English speakers experienced in linguistic annotation. Annotators were asked to identify all non-word misspellings and provide the adequate correction for each one. Inter-annotator agreement was quite high - annotators agreed in 82.6% of the cases (Cohen's Kappa=0.8, $p<.001$). All disagreements were resolved by a third annotator (adjudicator). For details of the annotation procedure, see Flor and Futagi (2011).

The Annotation Scheme for this project provides three classes of misspellings, as summarized in Table 1. Classification of annotated misspellings was automatic.

Type	Description	Count in corpus
1	single token non-word (e.g. "businees", "inthe")	21,160
2	single token non-word for which no plausible correction was found	52
3	multi-token non-word misspelling (e.g. "mor efun" for "more fun")	383
	Total	21,595

Table 1. Classification of misspellings annotated in the study corpus.

The annotation effort focused specifically on misspellings, rather than on a wider category of orthographic errors in general. The annotation ignored repeated words, missing spaces[1] and improper capitalization. Many of the essays have inconsistent capitalization and essays written fully in capital letters are not uncommon (not only in our corpus). In addition, different spelling variants were acceptable. This consideration stems from the international nature of the two tests – the examinees come from all around the world, being accustomed to either British, American, or some other English spelling standard; so, it is only fair to accept all of them.

Overall, the annotated corpus of 3,000 essays has the following statistics. Average essay length is 321 words (the range is 28-798 words). 148 essays turned out to have no misspellings at all. Total spelling error counts are given in Table 1; 2.24% of the words in the corpus are non-word misspellings.

3 Spelling correction systems

3.1 Background

Classic approaches to the problem of spelling correction of non-word errors were reviewed by Kukich (1992). The typical approach for error detection is using good spelling dictionaries. The typical approach for correction of non-word errors is to include modules for computing edit distance (Damerau, 1964; Levenshtein, 1966) and phonetic similarity. These are used for ranking suggestions by their orthographic and phonetic similarity to the misspelled word. A more recent feature utilizes word frequency data for candidate ranking. Mitton (2009) and Deorowicz and Ciura (2005) describe state of the art approaches to non-word correction without contextual information.

The use of context for spelling correction was initially proposed by Mayes, et al. (1991) only for 'contextual spelling' – correcting real-word errors (e.g. writing '*fig*' instead of '*fog*'). A common strategy for this task is using pre-defined confusion sets, which makes it more amenable to classifier-based approaches (Golding and Roth, 1999). Several recent studies used a web-scale language model (Google Web1T n-gram corpus – Brants and Franz, 2006) for "context-sensitive" (i.e. real-words) spelling correction (Bergsma, et al., 2009; Islam and Inkpen, 2009; Carlson and Fette, 2007). Chen, et al. (2007) used a LM for pruning candidate corrections for non-words in web queries. Whitelaw, et al. (2009) used a LM for correcting non-word and real-word errors without a dictionary and using a statistically trained error model. Our study extends the use of language models to automatic correction of non-word errors, with a dictionary, but without any explicit error model.

3.2 ConSpel system

The ConSpel system was designed and implemented as a fully automatic system for detection and correction of spelling errors. The current version is focused on non-word misspellings. The system has two intended uses. One is to serve as a component in NLP systems for automatic evaluation of student essays. The other use is to facilitate automation for research on patterns of misspellings in ELL essays.

In ConSpel, detection policy is quite simple. A token in a text is potentially a misspelling if the string is not in the system dictionaries. A text may include some non-dictionary tokens that systematically are not misspellings. ConSpel has several parameterized options to handle such cases. By default, the system will ignore numbers, dates, web and email addresses, and mixed alpha-numeric strings (e.g. 'RV400'). The system can be instructed to ignore capitalized words (e.g. 'London') and/or words in all uppercase (e.g. 'ROME').

ConSpel spelling dictionaries include about 360,000 entries. The core set includes 245,000 entries, providing a comprehensive coverage of modern English vocabulary. This lexicon includes all inflectional variants for a given word (e.g. 'love', 'loved', 'loves', 'loving'), and international spelling variants (e.g. American and British English). Additional dictionaries include about 120,000 entries for international surnames and first names, and names for geographical places.

Dictionaries are also the source of suggested corrections. Candidate suggestions for each detected misspelling are generated by returning all dictionary words that have an edit distance up to a given threshold. With the default threshold of 5, a

[1] Annotation ignored missing spaces around punctuation (e.g. "*chairs,tables*", but all cases where missing spaces result in fused words were marked in annotation (e.g. "*inthe*").

misspelling can easily get hundreds of correction candidates. Since ConSpel is intended to work on ELL data, and ELL misspellings can be quite dissimilar from the intended words, starting with a large number of candidates is a deliberate strategy to ensure that the adequate correction will be included in the candidate set. Candidates are pruned during the re-ranking process, so that only a few candidates from the initial set survive to the final decision making stage.

Candidate suggestions for each detected misspelling are ranked using a set of algorithms. An edit distance module is used to compute orthographic similarity between each candidate and the original misspelling. Phonetic similarity is computed using the Double Metaphone algorithm (Philips, 2000). Word frequency is computed for each candidate using a very large word-frequency data source.

The main thrust of our new spelling correction system is the conjecture that non-word misspellings can be corrected better when their context is taken into account.

Local context (several words around the misspelled word in the text) provides lots of information for choosing the adequate correction. For each candidate, we check the frequency of its co-occurrence (in a language model) with the adjacent words in the text. This approach borrows from the family of noisy-channel error-correction models (Zhang, et al., 2006; Cucerzan and Brill, 2004; Kernigham, et al., 1990). With the advent of very large word n-gram language models, we can utilize large contexts (about 4 words on each side of a misspelling). Our current language model uses a filtered version of the Google Web1T collection, containing 1,881,244,352 n-gram types of size 1-5, with punctuation included.[2] Notably, ConSpel does not use any statistical error model.

A second context-sensitive algorithm utilizes non-local context in the essay. The idea is quite simple – given a misspelled token in a text and a set of correction-candidates for that word, for each candidate we check whether that candidate string occurs elsewhere in the text. Since content words have some tendency of recurrence in same text, the misspelled token might be such a case, and the candidate should be strengthened. The idea is somewhat similar to cache-based language model adaptation (Kuhn and De Mori, 1990), though there are considerable differences. First, our system looks not only in preceding context, but over the whole essay text. Second, and unique to our system, ConSpel looks not only in the text, but also into the k-best candidate correction lists of the other misspelled words. Thus, if a word is systematically misspelled in a document, ConSpel will strengthen a candidate correction that appears as a candidate for multiple misspelled instances.[3]

For each misspelling found in a text, each algorithm produces ranking scores for each candidate. We use a linear-weighted ensemble method to combine scores from different algorithms. First, scores for all candidates of a given misspelling are normalized into a 0-1 range, separately for each ranker. Normalized scores are then summed using a set of constant weights.[4]

The ConSpel system is implemented as a flexible configurable system. Configuration settings include choice of dictionaries, choice of algorithms and weights for computing the final ranking, and choice of the output formats.

4 Comparative evaluation

In this section we report the results of evaluation on data from our gold-standard corpus of 3,000 essays described in section 2. This evaluation focuses on detection and correction of the 21,212 single-token non-word misspellings (types 1 and 2 in Table 1) as well as false alarms raised by spell-checkers.

Evaluation included three systems. In addition to ConSpel, we tested Aspell (version 0.60.6), a popular open-source spell checking library (Atkinson, 2011). The third system is spellchecker included in Microsoft Office 2007 (hereafter 'MS Word').

All evaluations were performed "in full context" (rather than word-by-word) – each essay in the corpus was submitted to each system separately, as a simple text file. All evaluations used standard

[2] ConSpel system uses the TrendStream n-gram compression software library (Flor, 2012) for fast and memory efficient retrieval of n-gram data. As a result, the ConSpel system runs even on modest hardware (e.g. a 4GB RAM laptop), concurrently with other applications.

[3] A detailed comparative study of different context utilization methods is under way.

[4] The current weights were found experimentally, prior to the annotation effort described in this article. We intend to use machine learning methods in future research, using the annotated corpus for this purpose.

measures of recall, precision and F-score (Leacock, et al., 2010).

Evaluations for Aspell and MS Word were conducted twice – once with their original dictionaries[5] and once with the ConSpel spelling dictionary of about 360,000 word forms. Evaluations where Aspell and MS Word were bundled with ConSpel dictionary are marked below as Aspell+ and MS Word+.

4.1 Error Detection

Detection results for non-word misspellings are presented in Table 2. All systems show very strong recall rates, above 99%. There is more variability when precision of error detection is concerned. Both MS Word and Aspell benefit from using the larger dictionary – they raise much less false alarms than with original dictionaries (Aspell improves precision by about 4% and MS Word by about 6%). ConSpel shows best precision, the difference with second-best (MS Word+) is statistically significant at $p<.01$.

System	Recall	Precision	F-score
Aspell	99.45	86.66	92.62
Aspell+	99.14	90.92	94.85
ConSpel	99.40	98.43	98.91
MS Word	99.55	90.26	94.68
MS Word+	99.32	96.16	97.71

Table 2. Evaluation results: non-word error detection

4.2 Error Correction

For evaluating spelling correction, we again use the measures of recall, precision and F-score. Note that precision of error correction is defined as proportion of adequately corrected misspellings out of total number of misspellings that a system tried to correct (this excludes cases missed in detection).

We conducted error-correction evaluations with ConSpel in two variants. The baseline variant, ConSpel-A, ranks candidate suggestions using edit distance, phonetic similarity and word-frequency. The contextual variant, ConSpel-B, adds contextual information in the ranking process.

Results of error-correction evaluations are shown in Table 3. While MS Word speller provided the adequate correction (top ranked suggestion) in about 73% of annotated cases, its precision is only about 67-69%, due to large number of false alarms. Aspell has markedly lower accuracy – both in recall and precision. ConSpel-A has approximately same recall as MS-Word, but better precision (due to low rate of false alarms). ConSpel-B, which uses contextual information in ranking candidate suggestions, shows markedly better recall and precision than either ConSpel-A or MS Word (statistically significant at $p<.01$).

For Aspell, use of the larger spelling dictionary improved detection precision (fewer false alarms – see Table 2), but it has led to degradation in error correction – as shown in Table 3 (possibly ranking of candidates is affected by larger dictionaries).

System	Recall	Precision	F-score
Aspell	61.53	53.62	57.30
Aspell+	54.17	49.68	51.83
ConSpel-A	72.65	71.94	72.29
ConSpel-B	**78.32**	**77.55**	**77.93**
MS Word	73.34	66.49	69.74
MS Word+	71.71	69.44	70.56

Table 3. Evaluation results: non-word error correction (top ranked candidates only)

An additional way to evaluate automatic spelling correction is to consider how often the adequate target correction is found among the k-best of the candidate suggestions (Mitton, 2009; Brill and Moore, 2000). Figure 1 shows error-correction recall and precision results for four systems[6] using k-best values 1-5 and 10.

When two-or-more best-ranked candidates are considered for each misspelling, the baseline ConSpel-A system shows better performance than MS Word. Aspell results lag significantly below the other systems, although it catches up with MS-Word beyond k=5. ConSpel-B system outperforms all other systems, in both recall and precision. It

[5] Notably, both Aspell and MS Word in this evaluation came with respective default dictionaries for US spelling, and generated many false alarms when flagging words that are British and other international spelling variants. Such false alarm cases were discounted from the evaluation statistics.

[6] 'MS Word' and 'MS Word+' overlap for all values of k, except for k=1, thus only 'MS Word' is shown in Figure 1.

places the target correction among the top two candidates in 88% of cases, and among top three or more candidates in beyond 90% of cases.

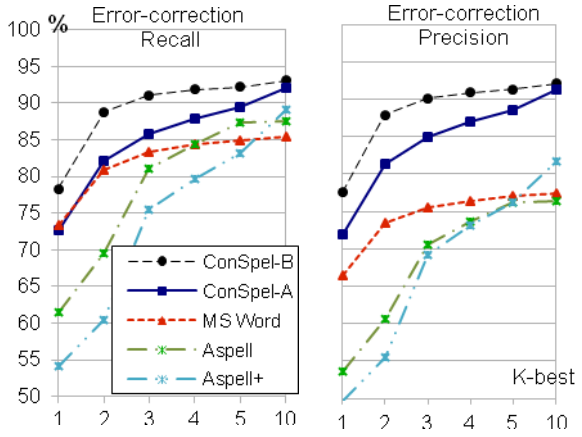

Figure 1. Error correction recall and precision for four systems, with different k-best cutoffs.

4.3 Evaluation with data from native and non-native English speakers

In this section we report the results of spell-check evaluation with data breakdown by native and non-native English speakers. Out of 21,212 single-token non-word misspellings in our corpus, 2,859 came from 570 essays written by native English speakers (NS) and 18,353 misspellings came from 2,282 essays written by test-takers who are not native speakers of English (NNS).

Comparison of error-detection for five systems is presented in Table 4. All systems show very strong recall results for both types of populations (all values are above 99%). The results are a bit different for error-detection precision. ConSpel achieves best results in both populations (the differences with second-best, MS Word+, are statistically significant at $p<.01$). MS Word has precision around 91%, approximately same in both populations. Compared to MS Word, MS Word+ has better recall rates, in both populations – due to a larger dictionary, it raises much less false alarms. Aspell lags behind in this comparison. Using a larger dictionary helps, as Aspell+ precision is better than that of Aspell in both populations; improvement is manifest for NNS data and only 2% for NS data. Aspell detection precision on NS data (77%) is lower than its precision on NNS data (88%). This may be due to Aspell having a problem with possessive forms (80% of the false alarms on NS data are possessives, but only 70% for NNS data).[7]

System		Recall	Precision	F-score
Aspell	ns: nns:	99.7 99.4	76.7 88.5	86.7 93.6
Aspell+	ns: nns:	99.6 99.3	78.7 93.3	87.9 96.3
ConSpel	ns: nns:	99.5 99.4	96.2 98.8	97.9 99.1
MS Word	ns: nns:	99.6 99.6	91.1 90.1	95.1 94.6
MS Word+	ns: nns:	99.2 99.3	94.4 96.5	96.7 97.9

Table 4. Evaluation results: percent correct for non-word error detection, with breakdown for data from native (ns) and non-native (nns) English speakers

Results of error-correction recall, with k-best levels 1-5 and 10, are presented in Figure 2. In comparisons of recall, with k=1, on NS data (right panel), MS Word (81.3%) and ConSpel-B (80.7%) show best results (the difference is not significant). For larger k-values, MS Word correction rate[8] improves to a ceiling of about 88.5% and both ConSpel variants have better improvement than MS Word. The context-informed ConSpel-B system has error-correction recall above 90% for $k\geq 2$ and reaches 94.2% at k=5.

On NNS data, ConSpel-B has a clear advantage over all other systems. At k=1, ConSpel-A and MS Word show equal correction performance (72%). For $k \geq 2$, ConSpel-A shows constant improvement, while MS Word improves to a ceiling of about 85%. For both NS and NNS populations, Aspell error-correction performance lags considerably behind the other systems, although it catches up with and even outperforms MS Word for $k\geq 3$. Interestingly, Aspell+ performs consistently worse than Aspell; the larger dictionary has detrimental effect on error-correction for Aspell, but not for MS Word.

[7] ConSpel dictionary does not contain possessive forms.
[8] Results for 'MS Word' and 'MS Word+' on this data overlap for all values of k, in both populations.

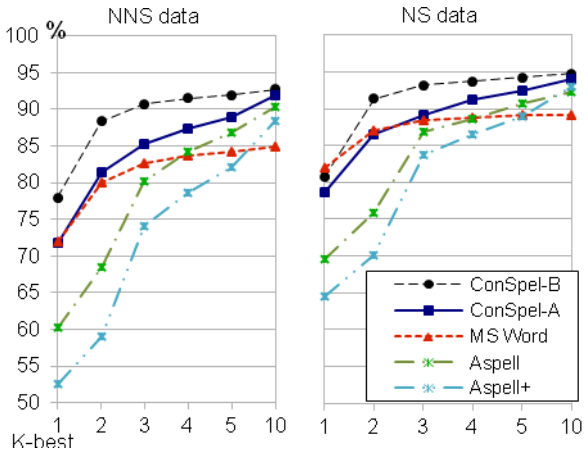

Figure 2. Error-correction recall for five systems, data from native (ns) and non-native (nns) English speakers.

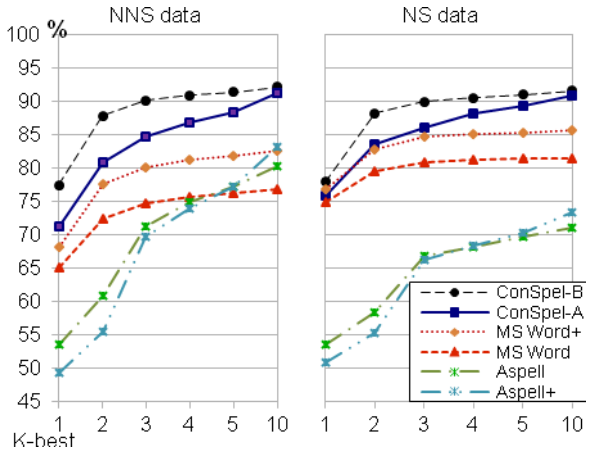

Figure 3. Error-correction precision for six systems, for native (ns) and non-native (nns) English speakers.

Error-correction precision results are shown in Figure 3. Overall, ConSpel-B outperforms all other systems, for both NS and NNS populations. On NS data, for k=1, MS Word+ (77%), ConSpel-A (76%) and MS Word (75%) are very close. For k≥2, ConSpel-A shows better improvement, reaching 89.4% at k=5, while MS Word+ reaches a ceiling of about 85% (81% for MS Word). Aspell performance lags clearly behind the other systems, although it also improves considerably with larger k-values. For NNS data, the separation between systems is even clearer. Aspell lags behind, although it catches up to MS Word at k≥5.

Except for ConSpel-B, all systems have manifestly better error-correction precision on NS data than on NNS data – misspellings made by non-native English speakers are harder to correct. ConSpel-B, with context-informed ranking of spelling suggestions, performs almost equally well for both populations. For k=1, its error-correction precision is 77.5% for NNS data and 78% for NS data. For k=2, precision is 87.9% for NNS and 88.2% for NS data. These differences are not statistically significant. For both populations, precision rises beyond 90% for k≥3. ConSpel-B also shows remarkably close error-correction recall in both populations: at k=1, recall is 77.9% for NNS and 80.7% for NS; at k=2, recall is 88.4% for NNS, 91.4% for NS (the differences are statistically significant). For k≥3, recall is beyond 90% for both populations, with about 2% advantage for NS population.

Table 5 presents F-scores for error-correction evaluation results, for six systems, for k-best values 1-5 and 10, for NS and NNS data. For each value of k, the ConSpel-B system has best values for both NS and NNS data. For each cell in Table 5, we calculated the absolute difference between the NS and NNS F-scores. The results are shown in Figure 4. Except for ConSpel-B, all systems have marked differences in performance on NS and NNS data. The differences tend to diminish for larger k-values. ConSpel-B is the only system for which the differences in error-correction between NS and NNS data are consistently below 2%, even for k=1.

K-best:	1	2	3	4	5	10
Aspell	60.6 54.9	65.9 62.3	75.6 72.8	77.0 76.5	78.9 78.9	80.3 81.8
Aspell+	56.9 49.6	61.8 55.6	74.0 69.8	76.4 74.0	78.5 77.3	82.4 82.9
MS Word	78.2 68.4	83.2 76.0	84.4 78.5	84.9 79.5	85.1 80.1	85.2 80.6
MS Word+	78.7 69.3	84.8 78.7	86.9 81.4	87.3 82.4	87.5 83.0	87.9 83.9
ConSpel-A	77.2 71.5	85.1 81.2	87.7 85.0	89.7 87.0	90.9 88.7	92.4 91.6
ConSpel-B	79.3 77.7	89.8 88.1	91.6 90.5	92.1 91.3	92.6 91.7	93.2 92.5

Table 5. Error-correction evaluation results: F-scores for six systems, data from native (upper value in each cell) and non-native English speakers

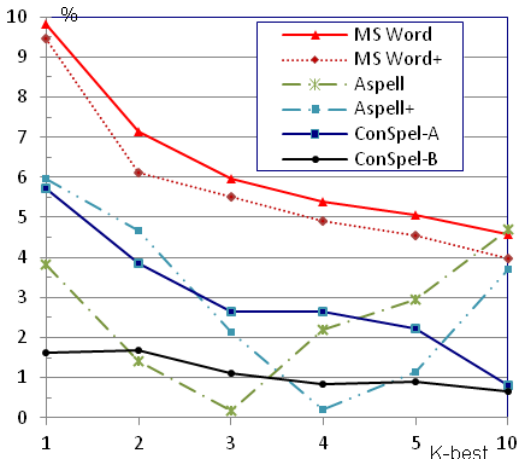

Figure 4. Error-correction F-scores absolute differences

A question we need to address is whether there are any real differences in misspellings produced by NS and NNS writers in our corpus. Our initial analyses show that there are some distinguishing characteristics.

One characterization is obtained when we look at the 'complexity' of the error, defining it as the edit distance between misspelling and correct word. The data is presented in Table 6. Native English speakers make significantly more simple errors (edit distance 1) than non-native speakers, while the latter make more complex errors (edit distance 4+).

Edit distance between misspelling and correct form	**NS**	**NNS**
1	83.3%	*79.9%
2	13.0%	14.0%
3	3.1%	3.9%
4+	0.6%	*2.1%

Table 6. Percent of non-word misspellings (tokens) by edit distance to correct word, for native and non-native populations. * difference significant at p<.01

Another difference we found in our data is the length (number of characters) of the correct word that was misspelled, for each population (Figure 5). For words of length 2 to 7, non-native speakers produce relatively more misspellings than native speakers. For words of length 8 and longer, native speakers produce relatively more misspellings than non-native speakers.

ConSpel-B performs about the same on NS and NNS data, and better than the other systems. Given the above differences of NS and NNS misspellings in our corpus, and given that all evaluated systems, except ConSpel-B, show better correction on NS data, we conclude that ConSpel-B shows this real advantage due to utilization of contextual data.

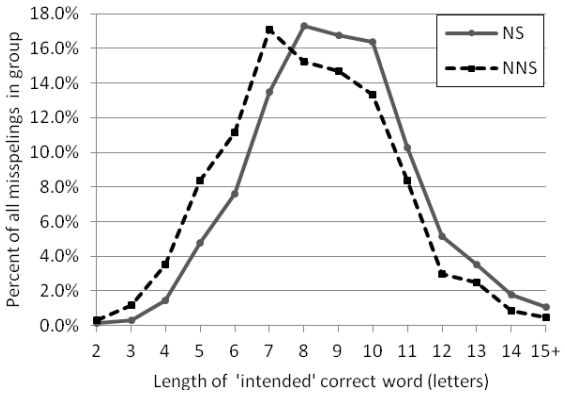

Figure 5. Percent of non-word misspellings (tokens) by length of the 'intended' correct word, for native and non-native populations.

5 Discussion

Large scale comparative studies of spellchecker performance on data from non-native language speakers are scarce, possibly due to large amount of effort required for expert annotation of data.

Hovermale (2010) analyzed 500 spelling errors from a corpus of essays produced by ELL in Japan. In that study, MS Word 2007 successfully corrected 72% of non-word errors, while Aspell had a success rate of 81% (presumably at k=1). In our study, with data from an international sample of non-native English speakers, Aspell error-correction precision rate is only 52% at k=1, and rises to 78% for k=5. MS Word and ConSpel-A (no-context) begin with precision of about 75-77% at k=1. At k=5, MS Word improves to about 85%, and ConSpel-A to above 89%.

Bestgen and Granger (2011) analyzed 222 argumentative essays from the ICLE corpus (Granger et al., 2009), written by European EFL students across different levels of English proficiency. Their sample included about 150,000 words and had 1,549 spelling errors. This amounts to spel-

ling-error rate of about 1%, compared to 2.2% in our data. In that study, MS Word 2007 had detection recall of 80.43%, and detection precision of 82.35%. In our study, MS Word had 99.6% recall and 90.1% precision in error detection. The difference may be attributed to the fact that we focus on single-token non-word misspellings, while Bestgen and Granger included other categories, specifically multi-token errors. Error-correction recall was 71% and precision 59% (at k=1). In our study, at k=1, MS Word achieved 72% recall and 65% precision, which is quite close to the above figures.

Given that our context-informed system has error-correction F-score of 77.9% at k=1, and 91.8% at k=5, it is obvious that the system picks up the right corrections. There is a potential for improvement, possibly by better ranking. Why doesn't the context help even more? Could the system perform with 90% at k=1? We have tentatively identified three major types of influences that detract the system from better performance. Those are a) local error density; b) poor grammar; and c) competition among inflectional variants. Local error density means simply that adjacent words are misspelled so there is not enough reliable context to use n-grams in such cases.

Poor grammar is also problematic for n-gram-based approach. In a fragment "*If docter want to operate, he...*", the intended word was '*doctor*', but '*doctor want*' is a subject-verb agreement error, which is not frequent in the normative n-gram data. Thus, under n-gram frequency influence, the system prefers '*doctors*' as top ranked candidate. There is competition of inflectional variants in presence of grammatical errors.

We have observed that even in absence of grammatical errors, sometimes an inadequate top ranked candidate is an inflectional variant of the adequate correction. For example: "*They received fresh air, interacte with other youth their age, solved problems...*". The adequate correction is '*interacted*', but ConSpel ranks it third, while '*interacts*' comes second and '*interact*' is ranked first. Notably, non-local context is not always beneficial – for a example, the presence of word '*interact*' elsewhere in the essay will strengthen the wrong candidate. Possibly, additional linguistic information could help improve ranking in such cases, e.g. by observing that all verbs in this sentence come in past tense.

Mitton (1996) suggested that it should be possible to adapt a spellchecker to cope specifically with L1-characteristic errors of English learners. Granger and Wynne (1999) analyzed misspellings produced by students with several different L1 backgrounds and have also suggested that it might be "useful to adapt tools such as spellcheckers to the needs of non-native users." Mitton and Okada (2007) have demonstrated a successful adaptation of a spellchecker (oriented for native English speakers) to Japanese learners of English.[9] However, adaptation to each specific L1 would require considerable resources. As noted by Hovermale (2010), it is not clear whether it is worthwhile to customize spellchecker heuristics for each learner population or better to just have one ELL spellchecker. Results from our study indicate that it is at least feasible to produce a general-purpose spellchecker that can successfully correct misspellings produced by non-native English speakers, almost as well as it does for native English speakers. A key for such development is utilization of essay context for re-ranking of spelling suggestions.

6 Conclusions

In this paper we presented a method for context-informed correction of single-token non-word spelling errors. Our results with ConSpel system demonstrate that utilizing contextual information helps improve automatic correction of non-word misspellings, for both native and non-native speakers of English, at least for essays written by test takers on standardized English proficiency tests. In future work we intend to produce a detailed study of the different ways of context utilization. We also intend to expand the system to handle multi-word spelling errors.

Acknowledgments

Many thanks to Chong Min Lee and Daniel Blanchard for assisting in evaluation with Aspell and Microsoft Office 2007; to our annotators, Nicole DiCrecchio, Julia Farnum, Melissa Lopez, Susanne Miller, Matthew Mulholland, Sarah Ohls, and Waverely VanWinkle. The manuscript has also benefited from the comments of three anonymous reviewers.

[9] Boyd (2009) used non-native (Japanese ELL) pronunciation modeling to improve a speller that uses just an orthographic error-model. Her combined system achieved 65% correction precision at k=1, and 82.6% at k=5. Our context-informed system achieves 77.5% and 91.4% respectively.

References

Reima Al-Jarf. 2010. Spelling error corpora in EFL. *Sino-US English Teaching*, 7(1):6-15.

Kevin Atkinson. 2011. GNU Aspell. Software available at http://aspell.net.

Shane Bergsma, Dekang Lin and Randy Goebel. 2009. Web-Scale N-gram Models for Lexical Disambiguation. In *Proceedings of International Joint Conference on Artificial Intelligence* (IJCAI-2009), pages 1507-1512.

Yves Bestgen and Sylviane Granger. 2011. Categorising spelling errors to assess L2 writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(2/3):235-252.

Adriane Boyd. 2009. Pronunciation Modeling in Spelling Correction for Writers of English as a Foreign Language. In *Proceedings of the NAACL HLT 2009 Student Research Workshop and Doctoral Consortium*, pages 31–36.

Torsten Brants and Alex Franz. 2006. *Web 1T 5-gram Version 1*. LDC2006T13. Philadelphia, PA, USA: Linguistic Data Consortium.

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of ACL*, pages 286-293

Andrew Carlson and Ian Fette. 2007. Memory-Based Context-Sensitive Spelling Correction at Web Scale. In *Proceedings of the Sixth International Conference on Machine Learning and Applications*, pages 166-171.

Qing Chen, Mu Li and Ming Zhou. 2007. Improving query spelling correction using web search results. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language* (EMNLP 2007), pages 181-189.

Robert Connors and Andrea A. Lunsford. 1988. Frequency of Formal Error in Current College Writing, or Ma and Pa Kettle Do Research. *College Composition and Communication*, 39(4):395–409.

Scott A. Crossley, Tom Salsbury, Philip McCarthy and Danielle S. McNamara. 2008. Using latent semantic analysis to explore second language lexical development. In Wilson, D. and Chad Lane, H. (Eds.): *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference*, pages 136–141.

Silviu Cucerzan and Eric Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP-2004), pages 293–300.

Frederick Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3): 659-664.

Rachele De Felice and Stephen G. Pulman. 2008. A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics* (COLING 2008), pages 69-176.

Sebastian Deorowicz and Marcin G. Ciura. 2005. Correcting spelling errors by modelling their causes. *International Journal of Applied Mathematics and Computer Science*, 15(2), pages 275–285.

Christy Desmet and Ron Balthazor. 2006. Finding Patterns in Textual Corpora: Data Mining, Research, and Assessment in First-year Composition. Paper presented at Computers and Writing 2006, Lubbock, Texas, May 25–29, 2006.

Semire Dikli. 2006. An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1):4-35. ejournals.bc.edu/ojs/index.php/jtla (last accessed on February 22, 2012).

ETS. 2011a. GRE®: *Introduction to the Analytical Writing Measure.* Educational Testing Service. www.ets.org/gre/revised_general/prepare/analytical_writing (last accessed on March 9, 2012).

ETS. 2011b. *TOEFL® iBT® Test Content.* Educational Testing Service. www.ets.org/toefl/ibt/about/content (last accessed on March 9, 2012).

Michael Flor. 2012. A fast and flexible architecture for very large word n-gram datasets. *Natural Language Engineering*. Available on CJO 2012 doi:10.1017/S1351324911000349 journals.cambridge.org/action/displayJournal?jid=NLE

Michael Flor and Yoko Futagi. 2011. Producing an annotated corpus with automatic spelling correction. Presented at the *Learner Corpus Research 2011 Conference*, 15-17 September 2011, Louvain-la-Neuve, Belgium. Submitted for publication.

Yoko Futagi. 2010. The effects of learner errors on the development of a collocation detection tool. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data (AND '10)*, pages 27-34.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier and Magali Paquot. 2009. *The International Corpus of Learner English*. Handbook and CD-ROM (Version 2), Presses Universitaires de Louvain, Louvain-la-Neuve.

Sylviane Granger and Martin Wynne. 1999. Optimising measures of lexical variation in EFL learner corpora. in Kirk, J. (Ed.): *Corpora Galore*, pages 249–257, Rodopi, Amsterdam.

Andrew Golding and Dan Roth. 1999. A Winnow based approach to Context-Sensitive Spelling Correction. *Machine Learning*, 34(1-3):107-130.

DJ Hovermale. 2010. An analysis of the spelling errors of L2 English learners. Presented at CALICO 2010 Conference, Amherst, MA, USA, June 10-12, 2010. Available electronically from http://www.ling.ohio-state.edu/~djh/presentations/djh_CALICO2010.pptx

Aminul Islam and Diana Inkpen. 2009. Real-word spelling correction using Google Web 1T n-gram with backoff. In *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering* (IEEE NLP-KE'09), pages 1-8.

Mark Kernighan, Kenneth Church and William Gale. 1990. A spelling correction program based on a noisy channel model. In *Proceedings of the 13th Conference on Computational Linguistics (COLING '90)*, pages 205-210.

Roland Kuhn and Renato De Mori. 1990. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583.

Karen Kukich, 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24:377-439.

Thomas K. Landauer, Darrell Laham and Peter Foltz. 2003. Automatic essay assessment. *Assessment in Education*, 10(3):295–308.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault, 2010. *Automated grammatical error detection for language learners.* Synthesis Lectures on Human Language Technologies, No. 9, Morgan & Claypool, Princeton, USA.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707-710.

Andrea A. Lunsford and Karen J. Lunsford. 2008. Mistakes Are a Fact of Life: A National Comparative Study. *College Composition and Communication*, 59(4):781-806.

Eric Mays, Fred J. Damerau and Robert L. Mercer. 1991. Context based spelling correction. *Information Processing & Management*, 27(5):517–522.

Roger Mitton. 1996. *English spelling and the computer*. Harlow, Essex: Longman Group. Available electronically from http://eprints.bbk.ac.uk/469

Roger Mitton. 2009. Ordering the suggestions of a spellchecker without using context. *Natural Language Engineering*, 15(2):173–192.

Roger Mitton and Takeshi Okada. 2007. The adaptation of an English spellchecker for Japanese writers. Presented at: *Symposium on Second Language Writing*, 15-17 Sept 2007, Nagoya, Japan. Available electronically from http://eprints.bbk.ac.uk/592

Ryo Nagata, Edward Whittaker and Vera Sheinman. 2011. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1210–1219, Portland, Oregon: Association for Computational Linguistics.

Takeshi Okada. 2005. Spelling errors made by Japanese EFL writers: with reference to errors occurring at the word-initial and the word-final position. In V. Cook and B. Bassetti (Ed.), *Second language writing systems*, pages 164-183. Clevedon: Multilingual Matters.

Diana Pérez, Enrique Alfonseca and Pilar Rodríguez. 2004. Application of the Bleu method for evaluating free-text answers in an e-learning environment. *Proceedings of the Language Resources and Evaluation Conference* (LREC-2004), pages 1351-1354.

Lawrence Philips. 2000. The Double-metaphone Search Algorithm. *C/C++ User's Journal*, June, 2000.

Anne Rimrott and Trude Heift. 2005. Language learners and generic spell checkers in CALL. *CALICO Journal*, 23(1):17-48.

Anne Rimrott and Trude Heift. 2008. Evaluating automatic detection of misspellings in German. *Language Learning & Technology*, 12(3):73-92.

Casey Whitelaw, Ben Hutchinson, Grace Y Chung and Gerard Ellis. 2009. Using the Web for language independent spellchecking and autocorrection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP 2009), pages 890-899.

Mark Warschauer and Paige Ware. 2006. Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research*, 10(2):157–180.

Guoxing Yu. 2010. Lexical diversity in writing and speaking task performances. *Applied Linguistics*, 31(2):236–259.

Yang Zhang, Pilian He, Wei Xiang and Mu Li. 2006. Discriminative reranking for spelling correction. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 64-71.