A Computational Approach to Detecting Collocation Errors in the Writing of Non-Native Speakers of English

Yoko Futagi, Paul Deane, Martin Chodorow and Joel Tetreault

Abstract

This paper describes the first prototype of an automated tool for detecting collocation errors in texts written by non-native speakers of English. Candidate strings are extracted by pattern matching over POS-tagged text. Since learner texts often contain spelling and morphological errors, the tool attempts to automatically correct them in order to reduce noise. For a measure of collocation strength, we use the rank-ratio statistic calculated over one billion words of native-speaker texts. Two human annotators evaluated the system's performance. We report the overall results, as well as detailed error analyses, and discuss possible improvements for the future.

1 Introduction

Worldwide, there are an estimated 750 million people who use English as a second language, as compared to 375 million native English speakers (Crystal, 1997). In addition, in U.S. alone, there are an estimated 4 to 5 million students with limited English proficiency in public schools (National Center for Educational Statistics, 2002), as well as a large number of international students in American colleges and universities (Burghardt, 2002). These numbers highlight a growing need for support for non-native speakers of English who must perform academically in English, though their English language proficiency is limited.

There are different types of errors non-native speakers make, some of which are usage errors, such as prepositions or collocations. Collocation (e.g. "strong tea", "hold an election") has been defined in a variety of ways, for example, "words co-occurring within a short space of each other" (Sinclair, 1991), "arbitrary and recurrent word combinations" (Benson, 1990), or "a sequence of words or terms which co-occur more often than would be expected by chance" (Wikipedia).[1] The essential points are that (1) the words of the collocation, or collocates, occur close to, but not necessarily adjacent to, each other, (2) the combinations are arbitrary/conventional, and (3) they occur more frequently than by chance.

Various studies have found that knowledge of collocations is an important part of language learning. Zhang (1993) used a fill-in-the-blank collocation test and a pen-and-paper writing test with 30 native and 30 non-native speakers of English, all of whom were college freshmen. In both tests, native speakers significantly outperformed the non-native speakers. Zhang concluded that "collocational knowledge is a source of fluency in written communication among college freshmen," and that quality of collocations in terms of variety and accuracy is indicative of the quality of college freshmen writing." Al-Zahrani (1998) examined the relationship between the knowledge of English lexical collocations and the general English proficiency of 81 Saudi EFL[2] students at four

[1] http://en.wikipedia.org/wiki/Collocation

[2] English as a Foreign Language

academic levels at a Saudi university. A collocation test consisting of 50 verb + noun lexical collocations, a writing test, and an institutional version of the paper-and-pencil TOEFL®[3] were administered. The study found that as the students' academic level increased, so did their knowledge of lexical collocations. He also found that the knowledge of lexical collocation strongly correlated with the TOEFL® score, and that the writing test was a slightly better predictor of the students' knowledge of lexical collocations than the TOEFL® score. In a more recent study, Hsu (2007) found a significant positive correlation between the holistic score given by Criterion℠ 7.1[4] and the frequency and variety of lexical collocations in essays written by 62 Taiwanese college students. Wible, Kuo, Tsao, Liu and Lin (2003) reported that lexical miscollocations are among the most prevalent error types marked by teacher in the essays submitted through the web-based language learning system, Intelligent Web-based Interactive Language Learning (IWiLL).[5]

Even though the density of collocations in a text or speech can vary,[6] an atypical combination of words, such as "powerful tea" instead of "strong tea", can disrupt communication (Howarth 1998, Martyńska 2004, Wible et al., 2003). Since collocations are not rule-governed, and vary from language to language,[7] each one needs to be learned. Considering the great number of collocations that are in common use, teaching them to English language learners (ELLs) is a formidable task - as Wray (2000) states, "knowing which subset of grammatically possible utterances that is actually commonly used by native speakers is an immense problem for even the most proficient of non-natives, who are unable to separate out and avoid the grammatical but non-idiomatic sequences." Dictionaries and thesauri, common aids for an English learner, are often of limited value when one does not know the appropriate collocation and must sort through a list of synonyms to find contextually appropriate one. A computer program which automatically identifies collocation errors and suggests corrections would be a more context-sensitive lexical resource for those who are learning to write in English. Recent advances in the field of corpus and statistical linguistics make it feasible to build such a tool now. This paper discusses the overall design and evaluation of an automated system for detection of collocation errors in non-native speakers' writing.

2 Existing Approaches

There are a number of corpus-based tools focused on collocation errors, among which are Pantel and Lin (2000), Shei and Pain (2000), Chang, Chen, and Chang (2004), and Wible et al. (2003). Pantel and Lin (2000) construct a database of word-word collocation patterns based on corpus statistics, as a part of an effort to build a word-for-word glossing algorithm for parallel corpora in machine translation. An examination of

[3] Test of English as a Foreign Language

[4] *Criterion℠* is a Web-based application that evaluates a student's writing skills and provides score reporting and diagnostic feedback to both writing instructors and students. See http://www.ets.org/criterion.

[5] http://www.iwillnow.org/

[6] Cowie (1991, 1992) reports that between 37.5% and 46% of all occurrences of the verb + direct object construction are collocations in native speaker journalistic prose. Howarth (1998) found an average of 33% of occurrences of this construction to be collocations in native speaker social science texts.

[7] For example, the equivalent of English collocation "*strong tea*" would be something like "*thick tea"* in Japanese; "*take medicine*" would be "*drink medicine*".

Pantel and Lin's system for finding similar collocations[8] reveals that it is designed to retrieve all potentially similar collocations, allowing language learners to search a long list for potentially more useful phrases, but that it is not specifically designed to detect or correct collocation errors automatically. However, its basic strategy – using a corpus to build a database of corpus usage – has clear value for the problem of identifying collocation errors.

Shei and Pain (2000) specifically addressed the task with which we are concerned: the recognition of collocation errors in text produced by English language learners. Their system picks out the elements of target structures (they give V and N of a VP, and an Adv followed by an Adj as examples), which are first checked against a reference database to see if they make up a valid collocation. If they are not found in the reference database, they are looked up in an error database, which consists of common ELL collocation errors. If the structures are found in this database, they are marked as definitely anomalous. If the structures are still undecided, they use a synonym dictionary to see if any synonym would form a legitimate collocation with the other collocate(s). If an appropriate synonym is found, it is suggested to the user as an alternative. Finally, the entire structure may be replaced by a more native-like collocations listed in what they call "Definition Dictionary", which was created from paraphrases of collocations given by the learners. Many of their methods can be applied, with modifications, to address the error detection and error correction problems. However, they do not report their system's overall levels of performance, and it is unclear to what extent their system depends on automated methods purely, as opposed to the manually constructed database of common errors that it incorporates.

Chang et al. (2004) reported a web-based automatic collocation-detection system which makes use of bilingual corpora. They, like Shei and Pain (2000), specifically targeted verb-noun miscollocations which resulted from L1 interference. Their system checks a verb-noun collocation in user input, then derives a list of candidate English verbs that share the same Chinese translations using bilingual corpora. Next the system checks a reference English corpus to choose the proper collocation, which is given to the user as feedback.

Wible et al (2003) extracted 177 miscollocations from the IWiLL learner corpus which is a collection of Taiwanese student essays, some of which are marked by teachers. They found an overwhelming proportion to be verb-noun collocation errors (145). They also found that in all but three of the verb-noun miscollocations, it was the verb that was inappropriate. Based on these observations, they created a list of nine nouns, each of which has its own list of corresponding miscollocated verbs which were inappropriately used with it in the corpus. The grammar checker automatically marked a collocation error when one of the nine nouns was found with a corresponding miscollocated verb. They found this approach to have a high precision rate (95.5%), but its obvious limitation is that it would require a huge amount of human-annotated learner writing in order to cover a wider variety of miscollocations, and such an annotation effort would be very costly and labor-intensive.

None of the studies cited above provides extensive evaluation of system performance, but they do provide useful information about the typical distribution of collocation errors (at least for Chinese learners of English) and what approaches may be

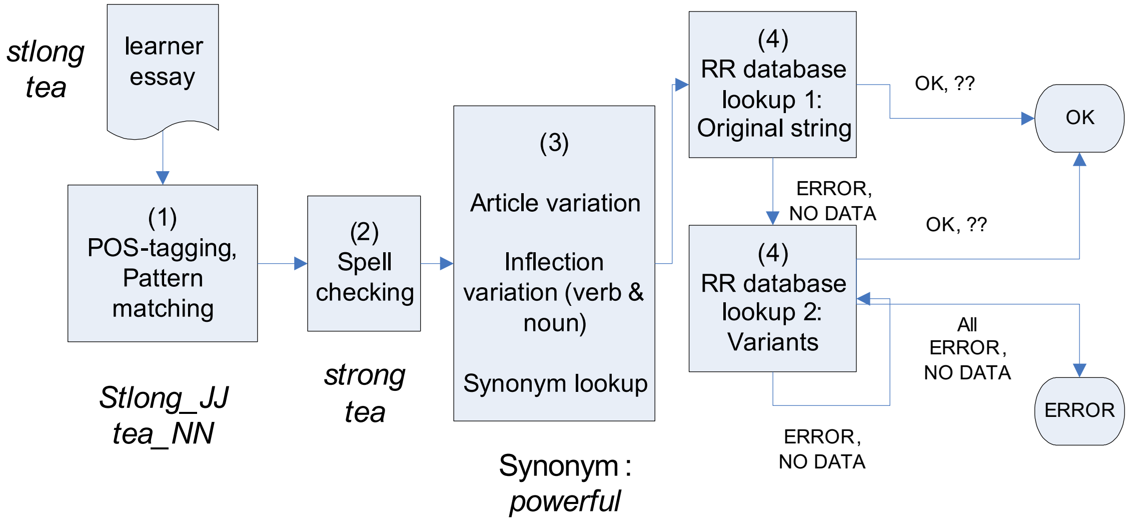
[8] http://www.cs.ualberta.ca/~lindek/demos/chooseword.htm

useful. One of the primary purposes of our study is not only to build such a collocation error detection system, but also to evaluate its performance and report on it in detail, in the hope of providing insight into the complexity of the problem at hand and of finding effective solutions. The next section describes the system architecture of our prototype, and section 4 describes how its performance was evaluated tool, and our evaluation method is described in section 4.

3 System Architecture

For detection of collocation errors, the tool does the following: (1) extract candidate word strings by tagging for part-of-speech (POS) and using regular-expression pattern matching, (2) check for spelling errors, (3) prepare the word string for database lookup by constructing variants of it with different articles and inflections and with synonyms, then (4) look up the word string in a co-occurrence database (the RR database). Please see Figure 1 for a flow-chart representation of the system architecture:

Figure 1: System architecture



3.1 Candidate String Extraction

The first step is to identify strings of words in learner data which match target syntactic patterns. Every word sequence is a potential collocation, but checking every single one of them would be inefficient. It is more practical to select the sequences that are more likely to be collocations. The BBI Combinatory Dictionary of English (Benson, Benson, and Ilson, 1997) identifies seven kinds of lexical collocations:

Table 1

Seven kinds of lexical collocations

Label	Syntactic Pattern(s)	Examples
L1	Verb of creation + Noun Phrase or Prepositional Phrase	*come to an understanding, launch a missile*
L2	Verb + Direct Object (our VERB-noun)	*reject an appeal*
L3	Adjective + Noun, Noun + Noun (our ADJ-noun and NOUN-noun)	*strong tea, house arrest*
L4	Subject + Verb	*bees sting*
L5	Noun of Noun (our NOUN-of-noun)	*a swarm of bees*
L6	Adverb + Adjective	*sound asleep*
L7	Verb + Adverb (our verb-ADVERB and ADVERB-verb)	*argue strenuously*

For this initial study, we focused on the syntactic patterns in the shaded rows. In each pattern, one word (shown in the table in the upper case, as in VERB-noun) is checked against the reference database to see how typically it appears in the context of the rest of the string. For the example in the table, the database would be used to determine (how typical "reject" is in the context of "_____ an appeal". In an ideal world, we would look for these syntactic patterns using syntactically parsed text. However, existing parsers are trained on native speakers' data, but ELL writing contains a variety of errors that are not typical of native speakers, making the use of parsers impractical. Thus, we decided to use a POS tagger,[9] then use regular expressions to extract matching strings.[10] To minimize noise, we filtered out a list of words which typically do not participate in collocations. These stopwords included some lexical categories, such as numbers, quantifiers and pronouns, as well as generic words such as "thing". We also filtered out verbs which take sentential complements, such as "think", "believe", etc.

3.2 Spell Checking

One of the challenges of dealing with language learners' data is misspelling. ELLs often produce misspellings which are not typically made by native speakers. Candidate strings which contain misspellings cannot be found in the reference database, since we need the exact strings (n-grams) to look up co-occurrence information. Thus,

[9] We used a maximum entropy POS tagger, which has accuracy of approximately 97%. Please see Ratnaparkhi (1998) for details.

[10] Our system takes into account that the collocates may not be immediately adjacent to each other, and allows for some syntactic variation. For example, the verb + direct object pattern allows the object noun to be modified by a determiner and one or more adjectives.

the tool incorporates a set of simple heuristics to automatically correct misspellings in the extracted candidate strings. Each candidate word in a text is checked against an alphabetically-sorted list of about 150K real words. If the word is not found in the list, a misspelling is presumed and a simple lookup algorithm is used to find the closest word in that list within a minimum edit distance of the original word. To find the point in the list from which to start the look-up, the tool uses a binary search to find the first listed word whose initial 3-letter sequence matches that of the misspelled word. From there, the search proceeds word-by-word, in each direction, until either a good-enough match is found, or the initial three letters no longer match. The problem with this system is that if a misspelling occurred in the first 3 letters, e.g. "ploblem" as misspelling of "problem", the system cannot find the right place to start the search.

3.3 Working with Reference Database

Two other learner data issues that we corrected in order to maximize the tool's performance were article errors and inflection errors. If the spell-checked original candidate string was not found in the data, the tool generates variants of the string by varying the article, and by varying verb or noun inflection. For example, "have knowledges" is not found in the database because of an inflection error. The tool then generates article variants ("have the knowledges"), inflectional variants ("have knowledge"), and the combination ("have a knowledge" and "have the knowledge"). These variants maximize the chance of finding the co-occurrence information for the word combination.

Our reference database, which we refer to as the "RR database", was created from 1 billion words of text from the Lexile corpus (informational and literary texts; collected by the MetaMetrics Corporation[11]); and the SourceFinder corpus (literary and scientific journal articles; Sheehan, Kostin, Futagi, Hemat, and Zuckerman, 2006). It uses the rank-ratio (RR) statistic (Deane, 2005) as a non-parametric measure of the likelihood of a word in a given context (please see Appendix for further information on rank-ratio statistic). RR database specifies a word+context, and returns the RR and the frequency of the given word in the queried context.

The tool also queries close synonyms of the word in the same context, and if synonyms with a better RR exist, their RR and frequencies are also returned. The rationale is that it is not always enough to know how often a word appears in a particular context; we need to know how much more often than its synonyms the word appears in that context. For example, "powerful tea" is a miscollocation, not just because "powerful" is rare in this context, but also because there is a very frequent alternative, "strong". Therefore, the tool needs to look up synonyms in the RR database to see if one or more of them is a stronger collocate than the original word.

We use direct lookup in WordNet and in an electronically available version of Roget's Thesaurus to find synonyms, then filter them by the POS tag of the original word so that the synonyms for the irrelevant senses of a polysemous word are excluded from the query. For example, if we are looking up the word "kind" as in "a kind person", we do not want synonyms for the noun "kind" as in "a kind of vegetable".

RR database lookup proceeds as follows:

(1) A synonym list is retrieved for the target word (e.g. "strong" in "strong tea" above).

[11] www.metametricsinc.com

(2) The RR and frequency of the original string (with spell-correction if any) are retrieved from the database, along with those of the synonyms.

(3) The RR of the original string is checked against a threshold (which was set experimentally); if RR is higher than the threshold and none of the synonyms has a significantly higher RR, the tool judges the original string to be a good collocation and the system returns OK.

(4) If the original string was not OK, the database lookup continues with each of the article and inflection variants, until either: one of them is deemed OK, the tool judges the candidate string to be a bad collocation and returns ERROR; or none of the variants were found in the database ("ND (no data)"). If any of the synonyms has a significantly higher RR than the target word in the context, it is returned as a possible alternative.

4 Evaluation

To evaluate the performance of the tool, we compared its decisions with those of two speakers of English. The decision to use human raters instead of dictionaries was made because we wanted to know how well the tool's performance reflects native speakers' knowledge of collocation usage. We used two raters instead of one, because of the inherent danger of a rater bias concerning language use. This point is presented in detail in Tetreault, Futagi and Chodorow (2008). The raters were trained to identify and judge collocation strings on 200 strings detected by the system.[12] Approximately 15% of these strings were marked as errors, which is higher than the error rate of the actual data they later annotated (shown in Table 2 below).[13]

Potential collocations were extracted from 300 randomly selected essays written by non-native English examinees on TOEFL®.[14] A total of 1,446 target collocation candidate strings were extracted and scored by our collocation tool (some strings appeared more than once as they occurred in different sentences and in different essays). Two native speakers of English were asked to judge whether each collocation was OK, a collocation error, or a string extraction error resulting in a false candidate string. The tool itself returned 3 responses, OK, ERROR, and NO_DATA which were, of course, not shown to the human evaluators. NO_DATA was grouped with ERROR for the purpose of this evaluation. Table 2 shows the results of the human and collocation tool judgments:

[12] Training set is not a part of the actual annotation set.

[13] For discussion of learner corpora error tagging in general, see Díaz-Negrillo and Fernández-Domínguez (2006) and works cited therein.

[14] TOEFL CBT (computer-based test) administered between April 2001 and March 2003.

Table 2

Summary of human and Collocation Tool judgments

	Rater 1	Rater 2	Collocation Tool
Collocation Error	225	206	50
OK	1051	1162	1396
False Candidate String	170	78	n/a

The collocation tool marked far fewer candidate strings as errors than did either human rater.

The ‘false candidate string’ category reflects judgments about cases where the collocation tool could not possibly make a correct judgment, so we decide to exclude them from our analysis of both inter-rater and tool-rater agreement. Examples of false candidate strings are:

Table 3

Examples of false candidate strings

False candidate string	Original sentence
say which subject	I can't *say which subject* is more important.
scientist study	This happened when the *scientist study* more and more about science and mathematics not about history and literature.

Table 4 shows examples which were proper candidate strings:

Table 4

Examples of the system response compared to human judgment

	System response	Examples
System and human agree	OK	*“wild guess*” *“wider range”*
	ERROR	*“do strong support* (for *give strong support*)” *“community animal* (for *social animal*)”
System and human disagree	OK	*“children development*” “*get the knowledge”*
	ERROR	*“have a successful career”* *“neglect the real potential”*

Table 5 is a summary of agreement on collocation judgments after those strings were excluded. In the first row, “Agree/Total”, “Agree” is the number of strings on which the raters or rater and the system had the same judgment, and “Total” is the number of strings at least rater deemed to be false candidates.

Table 5

Agreement with inappropriate strings excluded

	Rater 1 vs. Rater2	Rater 1 vs. ColTool	Rater 2 vs. ColTool
Agree / Total	1106 / 1260	1006 /1267	1057 /1368
Proportion Agree	0.878	0.788	0.773

The proportion of inter-rater agreement was somewhat higher than tool-rater agreement for either rater.

Two measures of performance, precision and recall, are commonly computed to evaluate natural language processing applications. They compare the system’s output to the gold standard or “true” classifications. For the collocation tool, precision and recall can be computed separately for OK and ERROR categories. Precision for OK (see Equation 1.1) is the number of true acceptable collocations that the tool detected divided by the total number of strings that it classified as acceptable collocations. Recall (Equation 1.2) is the number of true acceptable collocations that it detected divided by the total number of true acceptable collocations. Precision and recall for collocation errors are defined in an analogous manner in Equations 2.1 and 2.2.

Eq 1.1 Precision = [illegible]

Eq 1.2 Recall = [illegible]

Eq. 2.1 Precision = [illegible]

Eq. 2.2 Recall = [illegible]

Eq. 3.1 F_1-measure = [illegible]

Table 6 summarizes inter-rater agreement:

Table 6

Inter-rater agreement

Gold standard	OK judgments	Error judgments
Rater1	0.93	0.57
Rater2	0.84	0.63

Because agreement between the two human raters is not perfect, we calculated three sets of precision and recall values, first using Rater 1 as the gold standard, then using Rater 2 as the gold standard, and finally taking as the true classifications only those on which Rater 1 and Rater 2 agreed and eliminating collocations on which they disagreed. Table 7 shows the results for the categories OK and ERROR.

Table 7

Precision and recall of the Collocation Tool for categories OK and ERROR, based on three gold standards

	OK			ERROR		
Gold Standard	Precision	Recall	F_1-measure	Precision	Recall	F_1-measure
Rater 1	0.80	0.87	0.83	0.30	0.40	0.34
Rater 2	0.85	0.84	0.84	0.26	0.38	0.31
Raters' Consensus	0.89	0.93	0.91	0.28	0.43	0.34

We observe in these tables the same pattern between inter-rater agreement and tool-rater agreement. That is, both inter-rater and rater-tool agreement are much higher for OK judgments than for ERROR judgments. This pattern indicates that it is much easier to recognize good collocations than bad ones. This makes sense when one thinks about what collocations are: they are conventional combinations of words. It is much easier to recognize something one has already encountered many times before, than to recognize something that rarely or never occurs.

5 Discussion

5.1 Overview of Error Analysis

There were several sources of the miscategorization (false OK or false ERROR) made by the system. The three most common sources were: false candidate strings caused by string extraction problems; NO_DATA responses from the database; and misspelling. There were also other system issues, such as POS-tagger errors. The following table shows the number of miscategorizations (strings on which the tool disagreed with both human raters) attributable to each source:

Table 8

Collocation Tool miscategorization by source

	Source						
Pattern	False candidate string	'No Data'	Spelling	Other	System Issues	Total Mis-catego-rizations	Proportion of mis-categorization within the pattern
ADJ-noun	8	4	18	0	3	33	0.069
NOUN_of_noun	14	4	4	0	0	22	0.250
NOUN-noun	20	1	4	0	0	25	0.287
VERB-noun	20	47	9	8	3	87	0.243
ADVERB-verb	6	5	0	0	2	13	0.283
Verb-ADVERB	7	6	2	0	0	15	0.273
Overall	75	67	37	8	8	195	0.176
Proportion of Tool Miscategorizations	0.384	0.344	0.190	0.041	0.041		

Except for the ADJ-noun pattern, the proportion of miscategorization is fairly consistent across the patterns. By far, the two most frequent sources of disagreement between the human raters and the tool come from string extraction problems of the collocation tool (the column labeled "false candidate strings"), and strings that are unattested in the RR database (the column labeled "No Data"). Another major source of disagreement is misspelling. System issues other than string extraction problems, such as insufficient variant generation, were relatively few. "Other" includes cases where the sentence was so ungrammatical that the system could not be expected to process it correctly, and cases where the strings would be good collocations in other contexts. We discuss each miscategorization source in more detail below, starting with string extraction problems, followed by misspelling and other sources.

5.2 String Extraction Problems

As shown in Table 8, the false candidate string category accounts for just over a third of all the miscategorizations the tool made. This category consists of cases where the tool extracted a candidate collocation from a sequence that did not actually match one of the target grammatical constructions (examples in Table 3 above). This type of miscategorization is partly the result of using POS tagger output as the basis for syntactic pattern identification, since surface matches with target patterns can yield false candidate strings (i.e., any verb + noun incorrectly classified as verb + object). Efforts were made to filter out many of these miscategorization types, but the noise was large enough to produce a significant class of mismatches. The solution to this problem is to increase the number of filtering patterns to cause the system to ignore inappropriate strings such as those which contain verbs that take sentential complements, including "think" and "believe" (e.g. "*think good teachers*" was extracted as verb + object from a sentence "*I think good teachers are more important…*"). Increasing precision in this fashion will require a relatively small amount of additional development time, to add filter rules for preventing most syntactically inappropriate strings from being falsely recognized by the tool.

A less frequent source of extraction problems is mistagging by the POS tagger.[15] For example, the string "take care" was incorrectly extracted for the pattern Nn because the verb "take" had been tagged as a noun. The obvious solution is to use a better tagger, and in our future research, we plan to evaluate other POS taggers for this purpose. No matter which tagger is used, some tagging mistakes are inevitable because of the many grammar errors and misspellings typically found in non-native writing.

5.3 Misspelling

In the evaluation, human raters were asked to flag misspellings in the target string but to ignore misspellings when making their judgments so long as the intended word was clear (e.g. "childen" for "children" should be ignored). The tool was also programmed to detect misspellings and correct them, if possible, before deciding whether the string was OK or an ERROR. The results showed that misspellings rarely posed problems for human raters' judgments. In fact, they often failed to notice misspellings in the target strings and proceeded to make their judgments as usual.

The tool incorporates several fairly simple heuristics for spelling correction as described above. This method corrected the majority of misspellings in the input data, but even the remainder caused serious problems for the tool. Whenever the tool failed to correct simple misspellings, the misspelled sequence usually could not be found in the rank ratio database, and the tool returned a NO_DATA classification. For example, *"educational ploblems"* or "*current depelopment*" was not corrected by the system because the misspelling is in the first 3 letters. When the spelling tool worked correctly, it provided a very effective context-dependent spelling correction. If a hypothesized spelling-corrected sequence could be found in the Rank Ratio database, it was correct, whereas sequences that could not be found in the Rank Ratio database usually still contained a misspelling. Examination of these cases suggests that the spelling correction module can be improved considerably by using a more robust spell-checking module and considering multiple spelling corrections.

5.4 Other Miscategorization Sources

The problems under "System Issues" include various sources of miscategorizations in which the extracted string is appropriate for the given pattern and the misspelling detection/correction was performed properly, but the tool's classification nonetheless differed from that of the two human raters. The main problem lies in the algorithm which generates alternative forms of the extracted string in order to compensate for minor grammatical errors produced by non-native speakers (e.g. missing determiners), as well as for some of the "creative" aspect of human writing. For

[15] There were 8 cases, or less than 0.6% of all strings, in which mistagging unrelated to misspelling in the candidate string itself, occurred. Assuming that every misspelling is a source for mistagging, 37 remaining misspellings (under "Spelling" in Table 8) and the 8 mistagged cases account for 23% of all miscategorizations. Van Rooy and Shäfer (2003) compared the performance of 3 taggers, TOSCA, Brill, and CLAW, on learner texts. They found that correcting misspelling significantly improved tag accuracy of each tagger. Our plan to upgrade the spelling correction portion of the system, as mentioned in subsection 5.3, and to re-tag the spell-corrected candidate strings, are likely to improve this problem.

example, the string "*neglect the real potential*" was judged OK by the human raters, but was judged an ERROR by the tool. Although the system found synonyms in this context, it did not find the target word (*neglect*) in this context. If a determiner had been added to the "minimal" variant, then "neglect the potential" would have been found, and the tool would not have classified it as an error. Simple changes to the algorithm that generates alternative forms of the collocation ought to correct most of these problems.

Similar problems accounted for many of the NO_DATA cases as well, which was the second most frequent source of miscategorizations. This category is almost as frequent as string extraction problems. Together, they account for nearly 3/4 of all the miscategorizations made by the tool. The system classifies a string as an ERROR if the target word is not attested with above-threshold values in the rank ratio database, but at least one of the synonyms of that word is attested above a threshold rank ratio value. However, many words and their synonyms were not attested at all in the contexts that appeared in the student essays. For example, contexts containing the word "roommate" (e.g. "choose a roommate") were returned as NO_DATA by the tool because none of these combinations were found in the database.

A significant proportion of the residual NO_DATA cases can be handled by relatively small modifications to the method used to search the Rank Ratio database. In the evaluated implementation of the tool, longer sequences appearing in possible collocations were tested first, then shorter combinations were searched, possibly with minor grammatical alternatives. This search did not exhaustively consider all combinations, and in some cases did not proceed to consider minimal combinations of essential words (since these often were ungrammatical). For example, in the case of the Vn pattern, it is often only the verb and the head noun that make up the collocation. For example,

String:	"develop the children's ability"
Essential words:	"develop", "ability"

The possessive noun "the children's" is not essential to judging whether the verb "develop" can be used with the head noun "ability", and a more effective search of the RR database would examine several possible combinations, including "develop ability", "develop abilities", "develop the ability", "develop their ability", and a few other combinations. We have implemented some of the necessary changes, and the modified system is able to account for 29 of the 67 'No Data' cases.

However, the fundamental problem underlying these issues is essentially one of data sparsity: even with a database built from a corpus containing more than one billion words of running text, there is often very little statistical data about particular word combinations. One logical solution, therefore, is to increase the amount of data used. One source of such data is Google's n-gram database corpus (released through the Linguistic Data Consortium[16]), which was built from a collection of one-trillion words of English web pages, and contains every 2- through 5- word sequence that occurred at least 20 times in the corpus. Since the current Rank Ratio database was built from a much smaller corpus, this data source may provide significantly greater coverage and greatly reduce the number of unattested strings.

[16] www.ldc.upenn.edu/Catalog

6 Conclusions

In this paper, we presented a system that has near-human performance on valid collocations, while the performance on miscollocations needs improvement for operational use. The long-term goal of the system is to detect collocation errors and suggest alternatives. In this context, precision of error detection, that is, how many of the strings that the tool marks as ERRORs are actually miscollocations, is of paramount importance, while recall, or how many actual miscollocations the system detects, may be sacrificed to boost precision if necessary. One possibility to improve performance, then, is to experiment with the threshold of rank ratio to find the point which produces an acceptable level of precision (e.g., 90%) while sacrificing as little recall as possible. Incorporating a modified version of the approach used by Wible, et al (2003) may also help enhance the efficiency of our system. That is, to build a database of candidate strings which are judged to be miscollocations by the annotators.

Another area which needs improvement is the synonym finding, which was not formally evaluated at this stage, in order to give good suggestions to the user. We found that some of the "synonyms" the system uses are not appropriate (e.g. "accept" as a suggestions to "acquire" in the string "acquire logic"). While this does not happen all the time, it is a serious issue for a learner's tool. Work is planned to make a fuller use of WordNet, and also to investigate the possibility of using other thesauri.

The final point of discussion is the evaluation. We decided to use the tool's agreement with human raters as our benchmark. This was because we wanted to know how well the tool's collocation judgment works compared with that of native speakers. We used two trained human raters in order to prevent the bias which is inevitable when dealing with language usage. Our concern for bias seemed to be warranted in the case of miscollocation judgments where precision and recall were around 0.6. While this result may provide some insight into the difficulty of language usage annotation, for the purpose of evaluating the tool's accuracy, we are considering the use of collocation dictionaries in the future, in addition to the human raters.

There is also a sampling issue. As seen in Table 2, there were much greater numbers of OK judgments made by the system than ERROR judgments. Each rater also found approximately 5 times as many "good" collocations as "bad" ones in the same sample. The small number of collocation strings tagged ERROR means that we have less information about collocation errors to use to improve the tool's performance. Using a sample set with a more balanced distribution, as discussed in Tetreault et al. (2008), will help with this particular problem.

Once our system achieves the level of accuracy sufficient to be included in *Criterion*^{SM}, it will help learners write more natural English by detecting anomalous collocations in their writing and suggesting alternatives.

References

Al-Zahrani, M. S. (1998). Knowledge of English lexical collocations among male Saudi college students majoring in English at a Saudi university. Unpublished doctoral dissertation, Indiana University of Pennsylvania, Pennsylvania.

Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography, 3(1),* 23-25.

Benson, M., Benson, E., & Ilson, R. (Eds.). (1997). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations.* Amsterdam & Philadelphia: John Benjamins Publishing Company.

Burghardt, L. F. (2002, Spring). Foreign applications soar at universities. *New York Times*, Late Edition (East Coast), April 7, 2002.

Chang, R., Chen, T-P., & Chang, J. S. (2004, September). *An automatic collocation writing assistant for Taiwanese EFL Learners: Using Corpora for language teaching and learning based on NLP Technology*. Paper presented at the annual conference of European Association for Computer-Assisted Language Learning, Viena, Austria.

Cowie, A.P. (1991). Multiword units in newspaper language, in S Granger (ed.), *Perspectives on the English Lexicon*. Cahiers de l'Institut de Linguistique de Louvain, Louvain, 101-116.

Cowie, A.P. (1992). Multi-word lexical units and communicative language teaching, in P. Arnaud & H. Bejoint (eds.), *Vocabulary and Applied Linguistics*. Macmillan, London. 1-12.

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

Deane, P. (2005). A nonparametric method for extraction of candidate phrasal terms. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, MI, 605-613.

Díaz-Negrillo, A. and Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista Española deLingüística Aplicada*, 19, 83-102.

Howarth, P. (1998). The phraseology of learners' academic writing. In A.P. Cowie, (Ed.), *Phraseology: Theory, Analysis, and Application*. Oxford: Clarendon Press.

Hsu, J.-Y. (2007, May). Lexical collocations and their impact on the online writing of Taiwanese college English majors and non-English major. Paper presented at the International Conference on English for Specific Purposes, Taipei, Taiwan.

Martynska, M. (2004). Do English language learners know collocations? *Investigationes Linguisticae, 11*. (http://www.staff.amu.edu.pl/~inveling)

National Center for Educational Statistics. 2002. *Public School Student Counts, Staff, and Graduate Counts by State: School Year 2000-01* (NCES 2002-348).

Pantel, P., & Lin, D. (2000). Word-for-word glossing with contextually similar words. *Proceedings of the 1st Annual Meeting of the North American Chapter of Association for Computational Linguistics*, 4, 78-85.

Ratnaparkhi, A. (1998). Maximum Entropy Models for natural language ambiguity resolution. Unpublished doctoral dissertation, University of Pennsylvania.

Sheehan, K. M., Kostin, I., Futagi, Y., Hemat, R. & Zuckerman, D. (2006). Inside SourceFinder: Predicting the Acceptability Status of Candidate Reading Comprehension Source Documents. Educational Testing Service Research Report Series (RR-06-24). Princeton: Educational Testing Service.

Shei, C. C., & Pain, H. (2000). An ESL Writer's Collocational Aid. *Computer Assisted Language Learning*, 13(2), 167-182.

Sinclair, J. (1991). *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Tetreault, J., Futagi, Y. and Chodorow, M. (2008, March). *Reliability of human annotation of usage errors in learner text.* Paper present at the annual conference of the Computer Assisted Language Instruction Consortium, San Francisco, CA.

Van Rooy, B. and Schäfer, L. (2003). The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies 2002,* 20, 325–335.

Wible, D., Kuo, C. H., Tsao, N. L., Liu, A., & Lin, H. L. (2003). Bootstrapping in a language learning environment. *Journal of Computer Assisted Learning*, 19(4), 90-102.

Wray, A. (2000). "Formulaic sequences in second language teaching." *Applied Linguistics,* 21.4.

Zhang, X. (1993). English collocations and their effect on the writing of native and non-native College Freshmen. Unpublished doctoral dissertation, Indiana University of Pennsylvania, Pennsylvania.

Appendix: Rank Ratio statistic example

The following is a simplified example of rank-ratio statistic:

Example phrase: *east end*
Word: *east*
Context: ___ *end*

Suppose that the set if contexts and their *local* frequency (frequency when *east* fills the blank) for *east* and *global* frequency (overall frequency of the context, regardless of which word fills the blank) is as follows (all numbers are fictional for the purpose of illustration):

Context	Local frequency	Global frequency
___ *end*	38	470
___ *side*	46	910
the ___	98	12,000
toward the ___	24	1,300

The *actual rank* of ___ *end* is calculated by sorting all the contexts by their local frequency:

Context	Local frequency	Actual rank
the ___	98	1
___ *side*	**46**	**2**
___ *end*	38	3
toward the ___	24	4

The *expected rank* of ___ *end* is calculated by sorting all the contexts by their global frequency:

Context	Global frequency	Expected rank
the ___	12,000	1
toward the ___	1,300	2

___ *side*	**910**	**3**
___ *end*	470	4

The rank ratio of *east* in ___ *end* is the ratio between the expected rank (here, 3) and the actual rank (here, 2), that is, 1.5.