Examining the Use of Region Web Counts for ESL Error Detection

Joel R. Tetreault
Educational Testing Service
660 Rosedale Road
Princeton, NJ, USA
JTetreault@ets.org

Martin Chodorow
Hunter College of CUNY
695 Park Avenue
New York, NY, USA
martin.chodorow@hunter.cuny.edu

Abstract

Significant work is being done to develop NLP systems that can detect writing errors produced by non-native English speakers. A major issue, however, is the lack of available error-annotated training data needed to build statistical models that drive these major systems. As a result, many systems are trained on well-formed text with no modeling of typical errors that non-native speakers produce. To address this issue, we propose a novel method of using geographic region-specific web counts to detect typical errors in the writing of non-native speakers. In this paper we describe the approach, and present an analysis of the issues involved when using web counts.

1 Introduction

In recent years, much NLP work has been devoted to detecting errors in the writing of non-native speakers learning English as a Second Language (ESL). These efforts have focused primarily on the main errors that ESL writers typically make, such as determiner usage, e.g. "We read *a* same book" (Han et al., 2006; Lee and Seneff, 2006; Nagata et al., 2006), preposition usage, e.g. "She is married *with* John" (Felice and Pullman, 2007; Gamon et al., 2008; Tetreault and Chodorow, 2008), and collocations, e.g. "We purchased a *strong* computer." (Sun et al., 2007).

While early grammatical error detection systems used a collection of manually-constructed rules (such as (Eeg-Olofsson and Knuttson, 2003)), recent ones are largely statistically-based. They work by first developing a model of correct usage based on well-formed text produced by native writers (usually news text). Next, the system flags a usage as an error if it has a low probability given the model. In essence, the system diagnoses as an error any usage that seems statistically unlikely given the probability of the correct usage. Optimally, statistical models should be trained on examples of incorrect usage as well as on examples of correct usage. However, the few annotated corpora of learner writing that do exist are either not freely available or are very small in size and thus insufficient for training large models.

There are, of course, problems that arise from training exclusively on error-free, native text. First, some errors are more probable than others. For example, in the ESL literature it is noted that many English learners incorrectly use "married with John" instead of "married to John". These observations are commonly held in the ESL teaching and research communities, but are not captured by current NLP implementations. Second, it is well known that ESL learners from different first languages (L1s) make different types of errors (Swan and Smith, 2001). For instance, a writer whose L1 is Spanish is more likely to produce the phrase "*in* Monday" while a German speaker is more likely to write "*at* Monday". Without errors in the training data, statistical models cannot be sensitive to such regularities in L1 error patterns.

In the absence of a large corpus of annotated non-native writing, we propose a novel approach which uses the "region" search found in both the Google and the Yahoo search APIs to compare the distribution of a certain English construction in text found on web pages in an English-speaking country to the distribution of the same English construction on web pages in a predominantly non-English speaking country. If the distributions differ markedly, this is a sign that the English construction may be problematic for speakers of that L1.

Consider the example in Table 1 of "depends on" and "depends of". Native writers typically use the preposition *on* in "depends on". It should be

Region	*on*	*of*	Ratio	RR
US	92,000,000	267,000	345:1	
France	1,500,000	22,700	66:1	5.22:1

Table 1: Region Counts Example for "depends *preposition*"

noted that one can construct examples with *of* such as "it depends *of* course on other factors..." though these happen much less frequently. This distribution is reflected in the region counts for the United States. The more common usage "depends on" is used 345 times more frequently than "depends of." However, when performing the same queries with France as the region, the ratios are considerably different: 66 to 1. This means that the ratio of ratios (RR) comparing the US to France is about 5.2 to 1. We hypothesize that if speakers of a particular L1 had no problem with the construction, then the distribution would look similar to that of the US, but that a large RR, such as the one obtained for "depends of" signals a potential error. If enough L1s have distributions that deviate from the native English distribution, then that provides additional evidence that the construction may be problematic for non-native speakers in general.

Knowing what constructions are problematic can allow us to tune a system trained on native text in different ways. One approach is to adjust internal thresholds to make the system more sensitive to known errors. Another is to augment the training data for the statistical model with more examples of correct usage of the construction.

This paper makes the following contributions:

- A novel approach to detecting common errors by non-native speakers of English that uses the "region search" in search engine APIs. To our knowledge, this is the first NLP approach to use the region-dependent search. (Section 2)
- A preliminary validation study of the approach (Section 3)
- An empirical analysis of the issues involved when using region counts (Section 4)

Although this is a general method for discovering errors, here we will discuss its use with respect to preposition error detection in which the context licenses a preposition, but the writer used the incorrect one.

2 Region-Counts Approach

More formally, the approach works in the following manner. Given a construction (such as "married *preposition*" or "they used *determiner* stone"), do:

1. Select a gold standard region to compare against (either the US or the UK).
2. Select a set of non-native regions to query.
3. For each region, query the construction in its variant forms (e.g., "married to", "married of", "married with"; "they used stone", "they used a stone" and "they used the stone.") using a search engine and save the counts.
4. Upon completion of step 3, find the most frequently occurring variant in the gold standard distribution and calculate the ratio of that variant compared to every other variant in the region.
5. Using the variant form that was most frequent in the gold standard distribution (e.g., "married to"), calculate for every other region the ratio of that variant's frequencies compared to each of the other variants' frequencies.
6. Calculate the RR by comparing the ratios in the non-native region to the corresponding ratios in the gold-standard region.
7. Use a threshold function on the RRs to flag a construction as problematic in a specific region or problematic in general. For details on setting the threshold function see Section 5.

To illustrate how the approach works, we will use the example construction "married *preposition*" using the Yahoo search engine API, three prepositions (*to*, *for*, *with*), the UK as the gold standard region, and three non-native regions (China, Russia, France). Table 2 shows the results of the approach with this construction's three variants. The columns labeled "Count" show the Yahoo web counts for that region and variant. In this example construction, *to* is the most frequent variant in the gold standard region, so for each region, the ratios are calculated: *to*:*for* and *to*:*with*. The figures are shown in the columns labeled "Ratios." Next, the RR is calculated between the non-native ratios and the gold-standard ratios. For example,

	to	*for*			*with*		
Region	**Count**	**Count**	**Ratio**	**RR**	**Count**	**Ratio**	**RR**
UK	6,200,000	1,050,000	5.90:1		1,890,000	3.28:1	
China	417,000	62,300	6.69:1	0.88:1	92,900	4.49:1	0.73:1
Russia	378,000	57,100	6.62:1	0.89:1	185,000	2.04:1	**1.61:1**
France	191,000	23,600	8.09:1	0.73:1	162,000	1.18:1	**2.78:1**

Table 2: Example of Approach on "married *preposition*" where *to* is the most frequent gold standard preposition

RR for "married for" (China) is 5.90:1 to 6.69:1, or 0.88:1.

A RR greater than 1 signals that the region uses that particular variant relatively more than the gold-standard region. The larger the RR, the greater the "over" usage of that form. For example, France's ratio of "married with" versus "married to" is 2.78 times that of the UK. This is not surprising since many speakers of Romance languages have difficulties with the preposition *of*. Determining a threshold function for the RR (or any other metric one can derive from the relative frequencies) is an area we are currently exploring. One approach is to flag an entire construction if several regions have RRs markedly over 1.00, or if one variant has values over 1.00 in several regions. An example of this is "married with" which has a RR greater than 1.00 in two of the three regions in Table 2.

To put this approach into practice, one first needs to generate a list of constructions (and then variants), and use the region counting approach above to iterate through the list. In the case of preposition error discovery, one could take a large corpus of student writing and extract all bigrams (or any n-grams or skip-grams) that start with a preposition or end with a preposition, and treat those as constructions.

3 Proof of Concept

3.1 Validation with Examples of Known Errors

To test how well the approach described in Section 2 fares, we conducted a simple pilot study in which we checked to see if it was able to "discover" common errors described in the ESL literature. We collected 20 examples of common preposition errors from ESL research websites and second language acquisition papers. The examples consisted of the error commonly made, as well as the correct form. For the sake of space, we will focus on 5 of the 20 examples (see Table 3). The results for these 5 were representative of the larger set.

Correct Usage	**Incorrect Usage**
depends on	depends of
surprised by	surprised with
married to	married with
arrive at	arrive to
worried about	worried with

Table 3: Typical ESL Error Constructions

For each example, we collected region counts via Yahoo for 12 non-native regions, as well as counts for the US, which served as the gold-standard region. In all 20 examples, at least one region had a RR greater than 1.00. In 10 of the examples, over half of the regions had RRs greater than 1.00. Finally, in 15 of the 20 examples, at least one region had an RR greater than 2.00.

3.2 Validation with Student Data

Next, we checked to see if these errors actually occur in a large corpus of student writing and then quantified the need for error data in a preposition error detection system.

We extracted sentences which contained the target construction variants from 530,000 essays written for the Test of English as a Foreign Language (TOEFL®). The essays were written by non-native speakers representing 40 different L1s. Next, a trained annotator rated each construction variant, judging it as correct usage or incorrect usage, and then these judgments were reviewed by another trained annotator. Table 4 shows the corpus analysis and annotation statistics; for each construction the correct variant is listed first, and the incorrect variant second. The Frequency column shows the count for the variant in the entire

corpus, and the Errors column gives the percentage of those cases that were judged to be an error by the annotator. For constructions with hundreds of cases, the annotator rated a randomly selected sample of 150.

Variant	Frequency	Errors
depends on	18,675	0.6%
depends of	813	97.3%
surprised by	221	3.3%
surprised with	61	34.4%
married to	82	9.8%
married with	134	93.3%
arrive at	1,201	12.6%
arrive to	871	95.3%
worried about	2,857	2.7%
worried with	36	91.7%

Table 4: TOEFL Corpus Analysis

All 20 constructions appeared in the corpus of student essays. More importantly, the corpus analysis validates what the ESL literature (and the region-counts approach) predicted: in four out of the five cases listed above, the "incorrect" variant was an actual error over 90% of the time.

3.3 System Performance

Next, we used a preposition error detection system to determine how many of these errors the system currently detects. If it correctly identifies most of the incorrect cases as errors, there is no need to augment the system with this procedure. On the other hand, if a system performs poorly on these errors, this then shows the extent to which the approach can potentially improve performance.

For this analysis, we used our preposition error detection system (Tetreault and Chodorow, 2008) trained on 7 million preposition examples from native text. The system has been shown to be among the best performing systems. Over all of the constructions, the system missed on average about 80% of the errors. Table 5 ("Original Model") shows the results for five of the constructions. While the system had very high precision, its recall was very poor. For example, for the "married with" variant, it missed 88% of the errors in the annotated corpus. We believe that this shows the potential benefit of increasing the sensitivity of the system to errors which are known to occur frequently in ESL writing.

One method of using the approach to improve a system is to build small models specifically tuned to handle those constructions. If the variant is encountered, the system uses the tuned model, otherwise, it uses the more general, original model. For each construction, we extracted 50k examples from native text and trained a model in the same manner as the original model. We then evaluated this model on the error variants ("Tuned Model" in Table 5). Recall improved for four out of five cases, and substantially for "depends of" (45.2% to 80.1%) and "married with" (12.4% to 48.7%). This is, of course, a very simple way of leveraging the region-counts approach into a system; there are more sophisticated machine learning approaches one could use to tune a smaller model or augment the original model, though this is outside the scope of the current paper. However, we believe that the gains from this straightforward model tuning show the potential benefit of increasing the sensitivity of the system to constructions in which errors are known to occur frequently in ESL writing.

4 Reliability of Web Counts

While web counts have the advantage of being free, Kilgariff (2007) observed that there are limitations associated with their use: (1) there is no lemmatizing or part-of-speech tagging, (2) search syntax is limited, (3) the number of queries per day is constrained by the search engine and (4) web counts are for pages, not for unique instances (a page could have more than one instance of the query term). Despite these problems, previous work (such as (Keller and Lapata, 2003; Lapata and Keller, 2005; Nakov and Hearst, 2005; Nakov, 2007)) has shown that different NLP applications can be improved by using web counts. In this section, we examine the extent to which the limitations commonly associated with general web counts also affect region web counts and thus our approach. In 4.1, we examine how variable the region counts are over the course of one week, and in section 4.2 we look at a sample of web pages that the region search method returns and assess the quality of the sample with respect to our approach.

4.1 Variability of Web Counts

Web counts tend to vary from week to week, and sometimes even from hour to hour. This can be a problem for any approach, such as ours, which

			Original Model		Tuned Model	
Variant	**Frequency**	**# of Errors**	**Precision**	**Recall**	**Precision**	**Recall**
arrive to	149	142	100.0%	20.4%	100.0%	35.2%
depends of	150	146	100.0%	45.2%	100.0%	80.1%
married with	122	113	100.0%	12.4%	99.1%	48.7%
surprised with	61	21	85.7%	27.3%	100.0%	27.3%
worried with	36	33	100.0%	57.0%	100.0%	60.0%

Table 5: System Performance on Error Constructions

assumes that the counts are fairly stable. A frequency spike or dip in one region count could skew a RR and thus an error may be missed or spuriously flagged.

To assess the variability of the counts, we took the 20 examples from the previous section and collected the respective region counts (with UK as a gold standard and 12 non-native regions) using both Yahoo and Google. The process was repeated for seven consecutive days allowing us to track the variability of 520 region counts[1]. For each region and variant combination, we calculated its coefficient of variation (CV) over the 7 days (i.e., σ/μ, the result of dividing the standard deviation of its counts by its mean count) and then averaged all 520 coefficients of variation. Yahoo and Google had average CVs of 0.02 and 0.08, respectively, suggesting that the Yahoo search engine's region counts were somewhat more consistent over that one week period.

The most variable Yahoo searches were "insisted on" (Sweden) with a CV of 0.23, "disgusted with" (China), with a CV of 0.21, and "confronted with" (France), with a CV of 0.20. Google's most variable searches were "love with" (Japan) with a CV of 0.92, "confronted with" (Poland) with a CV of 0.74, and "surprised by" (Russia) with a CV of 0.72.

Taken as an aggregate, the CVs look acceptable, however there were several individual queries that showed wide variation when repeated. In the Google experiments, 10% of all the queries had an average CV greater than 0.20. These results suggest that our approach will likely miss some potential errors (or produce false positives on others). One way of dealing with this is to repeat the experiment several times over the course of a week or month and select the constructions which are consistently flagged as an error across those days. Of course, while this approach has the advantage of flagging errors more reliably, it has the drawback of having to use one's daily search quota on repeating experiments, thus slowing the pace of discovering new errors.

[1]There are 20 examples of 26 queries each: each example has a correct and incorrect construction, and 13 regions are queried for each.

4.2 Web Page Quality

While the variability of web counts can be an issue, the quality of the web pages counted in those hits can also impact the usefulness of the approach. For instance, it is possible that a variant with a high RR may not really be used incorrectly and that the high RR may be caused by missed punctuation, spam sites which repeat English phrases over and over, or American or British websites being hosted in a non-native region.

To determine the quality of the web counts, we randomly selected 10 variants with a very high RR and then examined the top 50 web pages that contained the variant, and another randomly selected 50, for a total of 100 web pages per variant. We annotated each web page using the scheme shown in Table 6. The third column of Table 7 lists the RR as well as the web counts for that variant.

The final tag distributions for each variant are shown in Table 7. Several of the variants: "confronted to", "consist by", "depend from", and "key-of", showed very high error counts (all 25% or more) which shows that for these cases the Ratio of Ratios metric is finding preposition usage examples that are problematic for non-native speakers. However, there are several other variants that were ranked highly that had very few errors. For example, "arrive on" had only four incorrect usages, and the remainder were either acceptable or language issues. Interestingly, many of the web pages in the set were tourism websites dealing with traveling to France. Another French example that only had a few errors was "nice on". We found that the overwhelming ma-

Tag Name	Code	Description
Error	Err	The variant in the gloss is an example of an incorrect preposition usage
Acceptable	Acc	The variant in the gloss is an example of correct preposition usage
Garbage	Gar	Web page is a spam site or listed as an attack site by Firefox
Language	Lang	Variant is actually an acceptable string in the native language and is included in the count though page is composed of mostly English sentences
Repeated	Rep	Gloss appears in another website
English	Eng	Site appears to be an American or British site hosted in that region
Punctuation	Punct	The variant in the gloss has punctuation in the middle that was skipped over by the search engine, or there should have been punctuation between the two words.

Table 6: Web Quality Annotation Scheme

Variant	Region	RR	Count	Err	Acc	Gar	Lang	Rep	Eng	Punct
arrive on	France	5.65	629,000	4	75	1	16	3	1	1
confront of	China	7.55	186	15	17	34	0	23	0	2
confront to	Japan	15.64	1,470	15	22	30	0	14	0	8
confronted to	France	20.41	32,800	98	1	0	0	1	0	1
consist by	China	23.55	1,660	32	2	50	0	15	1	2
depend from	Russia	4.35	3,630	81	4	7	0	5	0	1
dreamt for	France	17.15	12,400	9	12	1	0	78	0	0
dreamt in	Poland	39.76	4,290	9	22	0	0	68	0	1
key of	Korea	6.26	507,000	25	61	0	0	11	0	1
nice on	France	8.81	199,000	5	84	6	3	3	0	7

Table 7: Quality of Sample Web Pages

jority of acceptable cases were actually about the French city *Nice* and not the adjective. Other variants showed other peculiarities, and thus highlights the danger of using the raw web counts blindly. The variant "dreamt for" received a high "repeated" count because it is the title of a music album ("Dreamt for Light Years in the Belly of a Mountain"), and many French websites that were counted were either selling or reviewing the album. A similar trend happened when searching the string in the US or UK regions, but the ratio was larger for the French site because the counts for the other "dreamt *preposition*" variants were relatively smaller.

Overall, this quality experiment showed that for all ten cases, there were indeed some errors in each of the 100 glosses. However, some of the cases were very weak and were affected by problems with repeated website, punctuation and language issues.

Next, we checked how often each of the ten constructions appeared in our corpus of 530,000 student essays and, as in Section 3, rated each case as correct or incorrect preposition usage. Table 8 shows the frequency and error rates. The rightmost column shows the percentage of glosses that had the construction as an error (from Table 7, column 5). The chart shows that in 8 of the 10 constructions, a majority of the cases were actually errors. And in the remaining two, at least 20% of the cases were errors. It is also notable for those two cases that the web error counts were quite low: 4.0% and 5.0% respectively. This probably means that L1s other than French also use those phrases incorrectly.

5 Related Work

The "region counts" approach is just one method of trying to enhance current error detection models. For instance, Foster and Andersen (2009), created a system (GenERRate) to insert errors into native corpora to create large amounts of artificial non-native-like corpora. The advantage of their system is that it allows the user to create style sheets that control the type and number of errors. However, the performance impact from using artificial corpora in the error domain has yet to be ex-

Construction	Freq.	% TOEFL Errors	% Web Errors
arrive on	70	27.2%	4.0%
confronted to	100	100.0%	15.0%
confront of	11	72.8%	15.0%
confront to	21	90.5%	98.0%
consist by	8	100.0%	32.0%
depend from	94	91.5%	81.0%
dreamt for	8	75.0%	9.0%
dreamt in	3	100.0%	9.0%
key of	96	96.0%	25.0%
nice on	22	22.3%	5.0%

Table 8: Corpus Analysis of Discovered Errors

amined closely. Hermet and Désilets (2009) also developed a novel method of using roundtrip Machine Translation techniques to improve a standard preposition error detection system. Although their evaluation corpus was limited to 133 prepositions, the hybrid system outperformed their standard method by roughly 13%.

6 Discussion

In this paper, we have presented an approach to detecting common grammatical errors found in the writing of ESL speakers. The approach involves using the region search function found in the Yahoo and Google search APIs to gather statistics on the distribution of potentially problematic constructions in different non-native regions. These distributions are then compared to the distribution of a native English region. In addition, we presented results from a pilot study that showed the approach can detect common ESL errors noted in the literature, and we also verified that these errors do in fact appear in a large corpus of varied student writing, but that a state of the art preposition detection system fails to detect a significant portion of these errors. Finally, we demonstrated that these systems can easily be improved by training models that target the specific constructions. We believe that this demonstrates the potential impact such an approach can have on a system which detects common ESL errors.

While the preliminary results appear encouraging, our analysis showed that problems with variation as well as the quality of English web pages counted in non-native region searches may reduce the effectiveness of the approach. As a result, our future work will focus on the following areas:

Similarity Function In this work, we have used the RR metric to compare one region's variant ratios to the gold standard's, but other measures of distributional similarity are also available, such as Cosine similarity and Kullback Liebler (KL) Divergence.

Thresholding Function Another area to explore is how to threshold the similarity function. One could flag a whole construction or variant if several regions have RRs over a set value. This function can be empirically determined by comparing the distributions of constructions known to have errors with those that are known to be non-problematic for non-native speakers.

Collapsing Regions The variability in the region counts has the effect of potentially skewing the results of the thresholding function. False positives can arise if one uses a threshold function that flags the whole construction as an error if only a few regions have a very high RR. One way of reducing the impact of variable regions is to collapse regions into different language groups: East Asian (Japan, Korea, China), Slavic (Russia, Poland), and Romance (France, Spain, Italy, etc.). One can carry this aggregation even further and group all non-native regions into one class. The advantage of this approach is that it is less sensitive to the usual variations from the search engines, but the effects due to smaller regions may be less detectable and thus the system will miss these cases.

Finally, it should be noted that although we have focused on preposition error detection in this paper, this is a general approach that can discover problematic constructions for other types of errors. The method also has applications beyond grammatical error detection. For instance, it can form the foundation of a system which automatically generates test items for ESL students.

Acknowledgments

We would like to first thank Tomonori Nagano for his technical assistance and Sarah Ohls for her annotation work. In addition, we thank Michael Gamon, Detmar Meurers and the three anonymous reviewers for their comments and feedback.

References

J. Eeg-Olofsson and O. Knuttson. 2003. Automatic grammar checking for second language learners - the use of prepositions. In *Nodalida*.

R. De Felice and S. Pullman. 2007. Automatically acquiring models of preposition use. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*.

J. Foster and Øistein Andersen. 2009. Generrate: Generating errors for use in grammatical error detection. In *The 4th Workshop on Building Educational Applications Using NLP*.

M. Gamon, 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 *IJCNLP*.

N-R. Han, M. Chodorow, and C. Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12:115–129.

M. Hermet and A. Désilets. 2009. Using first and second language models to correct preposition errors in second language authoring. In *The 4th Workshop on Building Educational Applications Using NLP*.

F. Keller and M. Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3).

A. Kilgarriff. 2007. Googleology is bad science. *Computational Linguistics*, 33(1).

M. Lapata and F. Keller. 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing (TSLP)*, 2(1):1–31.

J. Lee and S. Seneff. 2006. Automatic grammar correction for second-language learners. In *Interspeech*.

R. Nagata, A. Kawai, K. Morihiro, and N. Isu. 2006. A feedback-augmented method for detecting errors in the writing of learners of English. In *Proceedings of the ACL/COLING*.

P. Nakov and M. Hearst. 2005. Using the web as an implicit training set: application to structural ambiguity resolution. In *HLT-EMNLP*.

P. Nakov. 2007. *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. Ph.D. thesis, EECS Department, University of California, Berkeley.

G. Sun, X. Liu, G. Cong, M. Zhou, Z. Xiong, J. Lee, and C.-Y. Lin. 2007. Detecting erroneous sentences using automatically mined sequential patterns. In *ACL*.

M. Swan and B. Smith, editors. 2001. *Learner English: A teacher's guide to interference and other problems*. Cambridge University Press.

J. Tetreault and M. Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *COLING*.