Automatic Detection of Preposition Errors in Learner Writing

Automatic Detection of Preposition Errors in Learner Writing

Rachele De Felice
*Educational Testing Service**

Stephen Pulman
Oxford University Computing Laboratory

ABSTRACT

In this article, we present an approach to the automatic correction of preposition errors in L2 English. Our system, based on a maximum entropy classifier, achieves average precision of 42% and recall of 35% on this task. The discussion of results obtained on correct and incorrect data aims to establish what characteristics of L2 writing prove particularly problematic in this task.

KEYWORDS

Preposition Errors, Automatic Error Detection, Preposition Usage, L2 English

INTRODUCTION

Preposition use is one of the areas of language that learners of English as a second/foreign language (henceforth L2 English or simply L2 for short) find most challenging. The *Cambridge Grammar for English Language Teachers* (Parrott, 2000), for example, defines prepositions as a "major problem" for learners, a finding confirmed by the analysis of a small error-tagged corpus we created in which prepositions account for 12% of the errors. The most common prepositions, for example, ***in***, ***of***, and ***to***, are also among the most frequent words in the language. Therefore, preposition errors are an ideal target for a study focusing on the possibility of automatically detecting and correcting errors in L2 writing because the frequency of the part of speech (POS), together with its high susceptibility to error, will ensure the availability of a sufficient amount of data.

This article presents our work on the preposition component of Dapper (Determiner And PrePosition Error Recogniser), a classifier-based system designed to automatically identify preposition errors in L2 writing. In the sections below, we first explain why preposition use is problematic and review related work. We then briefly introduce the classifier at the heart of the system. Next, we describe the L2 data used and discuss some of the issues related to using NLP tools with learner language. Following this, we present the results obtained and, before concluding, review the main obstacles encountered.

THE PROBLEM WITH PREPOSITIONS

Prepositions pose such a challenge to learners because they can appear to have no easily definable pattern which can be of assistance in making choices in novel contexts.[1] Indeed, the

*This work was completed while Rachele De Felice was a doctoral student at Oxford University Computing Laboratory.

requirements of the language often seem entirely idiosyncratic and unpredictable even across nearly identical contexts; for example, we say *I work **in** Verona* but *I work **at** Unilever*, despite the fact that the sentences have the same structure. Another potential pitfall is the fact that words with different POS but relating to the same lexical item will often require different prepositions, as in *independent **of*** versus *independence **from***. Similarly, words with related or near synonymous meanings cannot be relied on to follow the same prepositional patterns either: we say *reason **for*** but *cause **of***, for example.

It is therefore not surprising that learners encounter problems in mastering preposition usage, and it is often equally hard for native speakers to articulate the reasons for these differences or offer guidance on how to overcome such problems.

RELATED WORK

The body of work on automatic preposition error detection in L2 writing is not very large, although it has been experiencing a surge of interest recently. In work by Izumi, Uchimoto, and Isahara (2004), for example, a maximum entropy classifier is trained to recognize errors in a corpus of transcripts of L2 English spoken by Japanese learners. Their article reports results for various types of error (e.g., omission-precision 75.70%, recall 45.67%; replacement-P 31.17%, R 8%), but there are no figures for individual POS.

The work presented by Chodorow and Tetreault (Chodorow, Tetreault, & Han, 2007; Tetreault & Chodorow, 2008a, 2008b), on the other hand, focuses explicitly on preposition errors. The authors train a maximum entropy classifier to recognize correct usage of 34 prepositions, based on a set of 25 contextual features which include POS information, lexical items, and NP and VP chunks (e.g., 'preceding noun,' 'lemma of following verb,' etc.). These are used both as individual features and in combination. Data taken from the Google n-gram corpus[2] with regard to the most frequent sequences of nouns and verbs with the target preposition are also included, bringing the number of features to 41. These data are extracted from newspaper and high school texts and tested on both L1 and L2 texts. On the former, accuracy of 79% is achieved; on the latter, the figures are up to 84% precision and 19% recall. The authors also introduce several filters to minimize false alarms, that is, flagging an error where there is none. These include skipping misspelled words and accounting for cases where the preposition found in the text is either an antonym of the one given as correct (e.g., *to* vs. *from*) or is in a context in which more than one preposition may be correct.

Gamon et al. (2008) describe a complex system which uses both a decision tree, trained on text from the Encarta encyclopedia and Reuters data, and a language model, trained on the English Gigaword corpus, to detect preposition errors involving 13 prepositions. The feature set consists of several basic local features: a six-token window on either side of the potential preposition occurrence site is determined on the basis of POS tag sequences, and within this window the POS tags and lexical items present are taken into account. The classifier outputs a suggestion which is then scored by the language model. If the classifier choice receives a significantly higher score than the item found in the text, the former is given as a possible correction. Accuracy is just under 65% on L1 data and 56% on L2 data (a small set of texts written by Chinese learners of English).

Finally, Lee and Knutsson (2008) also address the issue of preposition generation to assist in grammar checking. They assess the contribution of a variety of lexical and syntactic features within a memory-based learning framework, training on the Aquaint corpus of news text. To the best of our knowledge, this is the only work, other than the present one, to in-

clude syntactic information derived from full parses in its feature set. Other features include the head of the PP phrase and the head of the object of the preposition. They do not, however, test their approach on preposition errors, but only on correct L1 data: on correct L1 data, the system achieves up to 71% accuracy in generating the correct preposition from a possible set of 10 prepositions. The inclusion of syntactic features is found to have a positive effect on the results.

The work we present here is similar to the articles described above in that it makes use of a classifier (a maximum entropy one) and a set of contextual features. However, we rely on a feature set which includes a wider range of syntactic and semantic elements, including a full syntactic analysis of the data. The other models do not incorporate any semantic information, nor, with the partial exclusion of Lee and Knutsson (2008), any deeper syntactic information such as preposition attachment as derived from parsing. While all approaches consider the nouns, verbs, and other lexical items in a preposition's context, here this information is extracted on the basis of syntactic relations (others choose to rely on linear ordering and chunking only). It is likely that the overlap between the type of information captured by these features and the POS/lemmas involved with the preposition will be high, especially where the object is concerned. However, it is also possible that some instances of more complex PP attachment will go undetected.

ACQUIRING CONTEXTUAL MODELS

What is a challenge for humans, whether L1 or L2 speakers, can also be a significant obstacle to the development of an automatic error detection system. As suggested above, it is very unlikely that one could manually craft all the rules necessary to describe correct preposition usage. Therefore, an alternative way of including this knowledge in a system must be found.

As anticipated above, our approach is based on a maximum entropy classifier. It relies not on handcrafted rules, but on exploiting the usage information that can be gathered from a large corpus of largely grammatically correct English. The full details of the development and structure of the system are given in De Felice (2008) and De Felice and Pulman (2008); here we will provide only a brief outline of the process.

Despite the issues discussed above, we believe there is nonetheless sufficient regularity in the syntactic and semantic characteristics of preposition contexts such that preposition choice can be predicted with an acceptable degree of accuracy. This is done by a classifier trained on a large number of correct contexts to acquire associations between prepositions and particular contexts, so that, given a novel instance, the preposition most likely to occur in that context can be selected with a certain degree of confidence. This can be easily adapted for error correction: given the occurrence of a preposition, Dapper can analyze the context of the preposition and determine what preposition is most likely to occur in that context; if it does not match the preposition used by the learner, an error is likely.

The preposition's context is described by a number of syntactic and semantic features that include the POS and stem of the lexical item modified by the preposition, the POS and stem of its object, what kind of named entities are involved (if any), the POS of the words in a ±3 word window around the preposition, and the WordNet lexicographer classes of any verbs and nouns involved. All this information is extracted automatically by using the C&C tools pipeline (Clark & Curran, 2007), which includes a stemmer (Minnen, Carroll, & Pearce, 2001), POS tagger, CCG parser, and named entity recognizer; the WordNet information is taken from the WordNet lexicographer files (http://wordnet.princeton.edu/man/lexnames.5WN).

Each occurrence of a preposition is turned into a feature vector which records its context in terms of the features selected; these vectors are then used for training and testing. We use the British National Corpus (BNC, Burnard, 2000) as our source of both training and test data. At the moment we focus our efforts on nine high-frequency prepositions to ensure sufficient data for training: ***at***, ***by***, ***for***, ***from***, ***in***, ***of***, ***on***, ***to***, and ***with***. Since these are the most frequent prepositions in English, we expect them to occur with high frequency in learner writing, too. The training data consist of nearly 9 million feature vectors representing prepositions and their context.

In testing, Dapper is presented with a set of over 530,000 feature vectors belonging to one of the nine preposition classes taken from a section of the BNC not used in training. The system is required to assign the correct label to the feature vector. Our best accuracy score obtained on L1 data on this task—where accuracy measures the number of times the class label was correctly assigned—is 70.06%.[3] To estimate an upper bound for the task, we conducted a small experiment with two native British English speakers, both graduate students. They were asked to carry out an analogous task, namely selecting one of the nine prepositions to complete a set of 841 contexts; their choice was marked as accurate only if it agreed with the original text, as well as being grammatical in the context. Accuracy for this task averaged 88%.[4]

TESTING ON L2 DATA

The Cambridge Learner Corpus

To test the model's performance on real examples of learner data, we use a 2 million word subset of the Cambridge Learner Corpus (CLC).[5] The CLC is a corpus of written learner English currently standing at over 30 million words (see http://www.cambridge.org/elt/corpus/learner_corpus2.htm). It is developed jointly by Cambridge ESOL and Cambridge University Press. The essays are error tagged according to a learner error coding system devised by Cambridge University Press which has been manually applied to the data. Details of the scheme can be found in Nicholls (2003). In brief, for errors involving single lexical items, the coding scheme records both the POS of the item and the type of error (missing, unnecessary, incorrect items); so, for example, 'RT' denotes a wrong preposition error, 'UD' an unnecessary determiner, and so on.

Our subcorpus presents a great deal of variation among the characteristics of the learners. Around 60 L1s are present, as are several different exam types, such as the Business English Certificate, the Certificate of Proficiency in English, and the Key English Test. We therefore expect to find a range of skills and topics in our data and a corresponding range of vocabulary and syntactic sophistication. Common exam questions require the writing of a business letter or report, essays which recount personal experiences or argue in favor of a position, or more informal letters to friends.

The aim of Dapper is to reliably detect preposition errors in L2 writing, so of course the best test of its success lies in assessing its performance on sentences containing such errors, noting how well it does in recognizing them and suggesting an appropriate, more idiomatic alternative. However, this is not sufficient: we must also ensure that it does not raise false alarms, that is, it does not flag the presence of an error where there is in fact none. This possibility, not unlikely given its imperfect performance on L1 data, is made even more likely by the factors liable to impair NLP tools, as described in the next section. Avoiding such false alarms is of even greater importance here than the L1 task because it would harm a learner's confidence and progress to be notified of nonexistent errors. Therefore, in creating a test set we take care to include both correct and incorrect preposition instances.

At this stage DAPPER is trained only to recognize appropriate usage of a set of nine prepositions and only where a preposition is actually required. So, we consider just those errors where a preposition is needed, but the one chosen by the student is incorrect. Additionally, we must ensure that for all these instances, both the incorrect preposition and the suggested correction are part of our set of nine of which the system has knowledge, otherwise it would be impossible for it to process them successfully. Text from all levels of proficiency is extracted.

The extracted sentences undergo some preprocessing steps in addition to those used to treat the L1 data. They are stripped of any residual XML markup relating to the text's structure (e.g., paragraph and script breaks), as well as both the XML tags for the error codes and the actual corrections inserted by the annotators. This is because DAPPER must be presented with text of the same kind it would receive if a student were inputting sentences directly, free of any markup introduced by the corpus annotators.

NLP Tools and Learner Language

In using NLP tools and techniques developed with and for L1 text on L2 data, a loss of performance can occur.[6] Looking specifically at those tools underpinning the workflow of DAPPER, some issues in particular stand out as being more likely; in the next sections, we will discuss whether these are indeed affecting performance. The tools in the C&C pipeline have been trained on correct English text from a particular domain (*Wall Street Journal*), and applying them to a different type of text and domain may have an adverse effect on their performance. This is an important consideration given that the output of the pipeline lies at the heart of DAPPER's feature vector construction. On the other hand, this problem was not encountered in parsing the BNC; additionally, the syntactic structure of the learners' writing will very likely be simpler than that of newspapers and so should not prove a challenge for the parser in any case.[7]

Errors in word order could also pose a problem for the tagger because they may lead to incorrect parses and relation assignment. In the sentence *I can you also send a map of London*, for example, *you* is given as the direct object of *can* rather than *send*. Parsing can also be affected by agreement errors, which are frequent in learner writing: subjects and verbs which disagree in number may not be recognized by the parser as belonging together. Overall, though, a quick overview of a subset of parsed learner sentences reveals that, in our case, the L2 nature of the data does not cause too many problems for the parser, and POS tags, syntactic structure, and grammatical relations are mostly correct.

More significant issues are likely to arise at the word level because there are several aspects of the analysis of individual lexical items that are susceptible to problems. This is of particular concern because our analysis of the contribution of various features to L1 results has shown that lexical items often play a very important role in enabling DAPPER to assign instances to the correct class. One of the most evident issues is that L2 writing contains a large number of spelling mistakes. Spelling mistakes which lead to nonexistent words negatively affect various components of the model: the lexical item may be incorrectly stemmed or not stemmed at all, and any information associated with the (correctly spelled) item will not be retrieved, including the use of the item itself as a feature. An example sentence with a preposition error illustrates this point, with the relevant word in bold.

(1) John understood straightaway the **reson of** her visit.

The misspelled word in this sentence, *reason*, is one which is not only frequent in English, but also has a very strong collocational tie with the preposition *for*. However, the link to this

crucial information is lost because the misspelled word cannot be matched to its correctly spelled equivalent. In analyzing the performance of Dapper, this is an important factor to take into account.

A related issue is the presence of misspelled words which actually result in another English word, some examples of which are given below, with the correct item included.

(2) I would like all the members to notify **stuff** [**staff**] in their section.

(3) Excepting the **brakes** [**breaks**] between each course, the seminar was well organised.

Here, the problem is that the information accessed by Dapper will be that which pertains to the incorrectly spelled but legitimate word, which can increase the risk of errors going undetected or false alarms being raised if the system identifies an error where there is none. Furthermore, while the previous issue could be addressed by running a spell check on the text before submitting it to the system, this type of misspelling would not be picked up in this way. The error coding is only of partial assistance: these errors are tagged sometimes as 'spelling errors resulting in a legitimate word' ('SX') and sometimes as 'replace noun' ('RN'), analogous to those cases in which a completely different lexical item has been used inappropriately (e.g., *coming* for *arrival*, *firm* for *office*) and cannot always be identified reliably.

Having introduced the CLC and pointed to some error types which we anticipate could impair Dapper's performance, in the rest of this article we provide a detailed analysis of the results obtained and examine whether these characteristics of learner writing do actually prove problematic for our system or whether, instead, other unanticipated problems arise.

RESULTS

Overall Performance

Our investigations of Dapper's performance on L2 data proceeded in two stages. We first tested it on correct preposition instances only because this allows a comparison with the L1 task and texts in order to understand the extent to which the system can be applied to types of texts and styles different from the ones used in training. We then tested it on a small number of instances of incorrect preposition use, its intended target. The relatively small size of this test set is designed to facilitate manual inspection of the results with the aim of detecting particular error patterns on the classifier's part which may impede better performance.

A full discussion of our findings cannot be given for reasons of space limitations; a detailed treatment can be found in De Felice (2008). Here we give the results obtained and focus on the main trends which emerge from our analysis. For the correct data, 5,753 instances are submitted to the system, of which 69% are accurately identified (see Table 1).

Table 1
Accuracy on L2 prepositions

Instance type	Accuracy
Correct	69%
Incorrect	39%
Average	54%

Note: Accuracy on incorrect instances refers to the classifier successfully identifying the presence of an error and assigning the correct preposition to the instance.

This is a very encouraging result because it is only 1% lower than that achieved on the BNC data. A loss of accuracy in moving to a different domain can be expected, but the loss here is quite small, pointing to a robust model which is not too tied to a specific type of data. The L2 nature of the texts does not, at least at first glance, create too many problems for the NLP tools.

To test Dapper's ability to identify and correct preposition errors, we submitted to it a set of 1,116 instances of erroneously used prepositions. Of these, the system flagged the presence of an error, that is, found a disagreement between the preposition it believed to be most appropriate for that context and the one used by the writer, in 76.43% of cases. Such a high accuracy score is initially very encouraging because it suggests that over three fourths of the errors are detected. However, it is not enough to be able to point to a perceived error in the text; the system can only be considered truly successful if error detection is accompanied by an appropriate suggestion for an alternative preposition choice—a much harder task. Of the instances flagged as incorrect, just over half (51.70%) are also corrected appropriately. This means that of 1,116 errors, the proportion actually identified and corrected is only 39.5%.

In trying to understand the possible reasons for the gap between performance on correct (whether L1 or L2) and incorrect text, the incorrectly labeled instances are inspected with particular regard to three factors: interrelation with another type of error, disagreement with the annotators, and grammatical acceptability. Interrelation with other types of errors refers to those cases in which the preposition error is due to the presence of another error in its immediate context. This includes not only spelling mistakes and incorrect POS, but also more complex cases where the error actually lies in the choice of lexical item as head or object of the preposition. The sentence below is an example, shown with and without its corrections.

(4) He greeted me for a lunch there and I greeted him for a drink.
He **greeted** [**treated**] me **for** [**to**] **#UD a /#UD** lunch there and I **greeted** [**treated**] him **for** [**to**] a drink.

For these preposition contexts, the classifier chooses the class ***with***, presumably on the basis of the high frequency of phrases such as *greet with a smile*, *a kiss*, and so on. However, the problem in this sentence clearly lies in the choice of verb rather than preposition, and indeed the annotators mark this by suggesting a different verb. The correct verb requires a different preposition, which leads to the preposition in the text also being marked as erroneous. However, Dapper of course does not have access to the corrected version of the text because it sees the sentence before any annotations have been marked. It is therefore very unlikely in these cases that the preposition it suggests is the one noted as appropriate by the annotators. Arguably, this is an issue related mainly to the annotation scheme: it might be more appropriate to tag these cases as a single error rather than two separate ones since the error effectively consists of a wrong lexical item rather than a wrong lexical item and a wrong preposition. However, despite the perceived high frequency of spelling errors and misused lexical items—and the expectation that they would prove a significant obstacle to good performance—our analysis reveals that these factors account for only 3% of cases in which Dapper did not assign the correct label to an instance.

The other two factors mentioned above, disagreement with the annotators and grammatical acceptability, are closely related. Both refer to cases where the classifier selects a preposition which is correct in the given context but not the correct one in that particular case either because it is not a pragmatically appropriate choice or because, despite being prag-

matically and grammatically appropriate, it is not the preposition selected by the annotators for that particular context. The former case includes examples such as the following:

(5) The view in Interlaken is wonderful.
The view **#RT in** [**from**] **/#RT** Interlaken is wonderful.
Classifier choice: **OF**

We can see in this sentence that the preposition suggested by Dapper is correct and yields a grammatical sentence, albeit one with a rather different meaning (an issue which also arises in the L1 data). These errors are evidence of the system using the linguistic knowledge it has acquired to inform its choices and, as such, constitute a different type of error, one which could be addressed for example by allowing the classifier to suggest more than one option, assuming such an output would also contain the more appropriate preposition. Classifier errors stemming from a grammatically correct but pragmatically incorrect choice account for around 7% of all errors. Despite being a significant amount, it is impossible to prevent these errors altogether without knowledge of the wider discourse; the best approach is to find a solution that deals with their occurrence, such as the one mentioned above.

Let us now consider instances where the classifier's suggested correction for an error it has identified does not correspond to the one proposed by the annotators. Since the annotators' corrections are used as the benchmark against which Dapper's performance is evaluated, these instances are counted as the classifier being wrong. The following sentence provides an example of this:

(6) I've known him since we were on primary school.
I've known him since we were **#RT on** [**in**] **/#RT** primary school.
Classifier choice: **AT**

As in the previous example, the classifier's choice here is grammatically correct; additionally, it also yields essentially the same meaning as the preposition selected by the annotators. The difference in choice lies in the individual preferences of the annotators, which are by no means consistent: for example, we find instances of sentences such as *I live **at** Green Street* corrected as both ***on*** and ***in*** *Green Street*. Examples of this sort should not be considered mistakes on Dapper's part since the system fulfils its function by correctly signaling the presence of an error and offering a correction which is well suited to the context, regardless of its agreement with the annotators. If we exclude these cases from the error count, then, we will obtain a more realistic indication of our model's performance, one that is free from the annotators' bias. We find that these 'nonerrors' account for 11% of the classifier's reported errors, making disagreement with the text the most prominent among the factors involved in affecting its performance.

On the basis of this analysis, the classifier's accuracy score can be recalculated to include such correctly identified errors. We assume that of the instances initially marked as mistakes made by the classifier, either because the error was not spotted or because the suggested correction differs from the one chosen by the annotators, 11% should actually be considered as correct instead. This brings accuracy to around 46%. We could further claim that the true measure of Dapper's error rate on this task should also exclude the two other types of errors discussed above, since those effectively relate to issues other than those of preposition error identification and correction. In doing so, accuracy rises further to 51.5%. While still lower than what is achieved on L1 or L2 correct data, it is close to the figure obtained by Gamon et al. (2008). Further research is necessary to better understand the causes of the remaining errors.

Individual Prepositions

To gain a clearer picture of the model's performance, highlight areas which may need improving, and possibly raise important points relevant to the wider discussion on the application of NLP tools to learner language, we also look at precision and recall scores for individual prepositions in the two tasks. This type of analysis was also performed on the L1 data, described in more detail in De Felice (2008) and De Felice and Pulman (2008). Table 2 below reports the results for the L1 data to facilitate comparisons, Table 3 refers to the correct instances, and Table 4 to the incorrect ones. The tables in the Appendix present confusion matrices for the L1 and L2 correct data. While insights arising from their analysis cannot be discussed here, more details are available in De Felice (2008).

By comparing results for the various tasks, we can establish whether the model performs in the same way on the different types of data or whether other kinds of issues arise. Recall is the measure of the system's success at recognizing that the target (correct) preposition is required in a given context. In the incorrect prepositions task, the system measures its success at detecting misused prepositions: low recall means that too many errors are going undetected. Recall is calculated as

$$\text{Recall} = \frac{\text{\# of times need for target preposition identified}}{\text{\# of times target preposition needed}}$$

Precision is the measure of the system's ability to return correct answers without also misclassifying other instances: it shows what proportion of instances assigned to a particular class actually do belong to that class. Low precision means that the class labels being assigned are often inappropriate. In our task, this would imply that learners are receiving false alarms (correct instances being mistakenly labeled as incorrect) and that errors are not being corrected appropriately, both of which are to be avoided. Therefore it is important, in a learner setting, to privilege precision over recall. The formula is the following:

$$\text{Precision} = \frac{\text{\# of times class label assigned correctly}}{\text{\# of times class label assigned}}$$

Table 2
Individual Prepositions Results: L1 Test Data

	Proportion of training data	Precision	Recall
of	27.83%	74.28%	90.47%
to	20.64%	85.99%	81.73%
in	17.68%	60.15%	67.60%
for	8.01%	55.47%	43.78%
on	6.54%	58.52%	45.81%
with	6.03%	58.13%	46.33%
at	4.72%	57.44%	52.12%
by	4.69%	63.83%	56.51%
from	3.86%	59.20%	32.07%
macroaverage		63.67%	57.38%

Table 3
Individual Prepositions Results: Correct L2 Data

	Proportion of test data	Precision	Recall
of	17.17%	67.66%	89.57%
to	20.35%	88.86%	78.31%
in	28.72%	69.85%	75.30%
for	10.95%	66.67%	57.46%
on	6.12%	57.29%	46.88%
with	4.64%	58.75%	35.21%
at	5.65%	45.28%	57.54%
by	3.06%	45.86%	40.91%
from	3.34%	57.75%	21.35%
macroaverage		62.00%	55.84%

Table 4
Individual Prepositions Results: Incorrect L2 Data

	Precision	Recall
of	21.29%	63.10%
to	49.22%	33.87%
in	34.72%	59.69%
for	48.46%	39.87%
on	70.09%	35.05%
with	18.75%	11.54%
at	48.96%	35.34%
by	33.33%	22.22%
from	52.94%	16.67%
macroaverage	41.97%	35.26%

Note: Frequency figures cannot be given due to intellectual property restrictions.

An examination of the precision and recall scores raises several interesting points, not all of which can be discussed in detail because of space limitations. For example, although averages for the correct L2 data are within less than 2% of their L1 counterparts, they are not derived from similar sets of figures. A full explanation for these discrepancies can only be had by looking more closely at the data in order to detect whether there are any particular contexts which cause the incorrect class assignations. Another issue to address is whether the distribution of the erroneous prepositions in the text mirrors that of their correct counterparts. This not only gives useful insights into the error patterns of L2 writers, but can also help explain some problems in Dapper's performance, should the texts be found to present a different distribution than that observed in training. Indeed, we find there are several significant divergences, and it is not always the case that prepositions which are most frequent in L1 English are also the ones most frequently misused by learners. Notable examples include ***at***, which occurs in erroneous constructions at a much higher frequency than its observed frequency in English would lead one to expect, and ***to***, which displays the opposite behavior, being tagged as incorrect much less frequently. In the correct data, we note that overall the distribution and relative frequencies of the data are very similar, except that in the L2 text the most frequent preposition is ***in*** rather than ***of***, which is in fact also less frequent than ***to***.

DISCUSSION

While it is encouraging that in principle the system can indeed be used to recognize and cor-

rect preposition errors, such a low recall score is not ideal because it would still leave far too many errors undetected, potentially misleading students about their actual level of achievement. However, results reported on similar tasks (see section on related work above) suggest that this is a challenging undertaking. A full analysis of all error tendencies and pairs of prepositions often confused with each other is not possible here. We restrict our analysis to some error patterns which are particularly representative of the kinds of problems encountered.

Content Issues

Several of the misclassifications are probably due to peculiarities of the type of content found in learner data. A typical assignment is to write a piece of (usually autobiographical) narrative regarding an event. In these essays, temporal expressions (e.g., *at first*, *at present*, *at 18*) serve as convenient structural and chronological markers. Unfortunately, they are something that Dapper does not handle very well, so among the misclassified PPs we find many temporal ones. Examples include instances of ***at*** being assigned to ***in*** (*at first*, *at present*, *at 12pm/5am*/etc., *at 65/18*/other ages) and instances of ***by*** and ***from*** being assigned to ***at*** (these regard almost exclusively a small set of phrases: *by the time that …*, *by the end*, *by the age of …*, *from the moment that …*, and *from time to time*). This is an issue not just because it leads to the misclassification of correct instances, but also because erroneous instances are not corrected in the appropriate way, as in the following example:

(7) The training programme will start **#RT at** [**on**] **/#RT** the 1st August 1999.
Classifier decision: **IN**

It is important to note that this type of confusion is not peculiar to the L2 data, as analogous misclassifications are found in the L1 results, too. However, these phrases have a higher frequency here than in the BNC texts, which in turn leads to this particular type of confusion occurring more often. This problem may point to something which requires more attention within the model itself, for example, in its treatment of numerals.

A different group of mistakes bears a relation to the prevalence of another type of L2 essay, the 'opinion piece,' in which students are required to state their opinion on an issue or argue for or against a position. Many PPs which are misclassified are typical of this style: *from my point of view* (assigned to ***at***) and *on the one/other hand* (assigned to ***in***) are two which particularly stand out. Their very high frequency suggests that learners may be overly reliant on a small set of 'prefabricated' chunks (Granger, 1998) and should be encouraged to explore other means of expressing opinions in their writing. Their lower frequency in the BNC data does not mean that argumentation and statements of opinion do not figure there at all but that they are more likely to be phrased with more variation and less reliance on these two particular lexical chunks.

Stylistic Issues

The style and sentence structure used by L2 writers also seems to be at the origin of many of Dapper's problems. An informal style is often found, either because the students have yet to master the complexities of more formal writing or because the assignment itself requires it, for example, if the text to be produced is a letter to a friend. The phrase *by the way* is one of the clearest examples of this, being both very frequent and regularly misclassified as ***in***. The expression *of course* is also very frequent and, again, misclassified as ***in***. Although it might

not appear to be a structural or chronological marker like the other phrases discussed, it does belong to a style that is more informal and typical of an insecure writer than what one would find in the BNC and, as such, figures very often in student essays where writers may not be fully confident of the force of their argument. Presumably these, and other informal phrases, do not occur very often in the training data, so the model is unable to recognize a strong relationship between the separate items in them. This is further evidence of that fact that the presence of a relatively small number of semifixed phrases is at the root of the classifier's errors; this finding also lends further support to theories that show that L2 writers tend to rely on a small set of lexical chunks and overuse them, displaying less stylistic variation than L1 writers.

There are also several peculiarities in sentence structure which mislead Dapper into assigning an instance to the wrong class. For example, learners, perhaps as the result of L1 interference, often begin sentences with *with*-PPs of a kind which can be compared to the Latin ablative absolute in function, as in the following examples:

(8) *With her mind panicking*, she called the police.

(9) *With his voice strong and confident*, he began to speak.

These phrases are not usually marked by the CLC annotators as incorrect, although they sound stilted to many fluent speakers of English. The preposition suggested by Dapper for these cases is ***in***; while this would not improve the readability of the sentence, it points to the underlying fact that English sentences rarely begin with the preposition *with*. On the other hand, adjuncts headed by *in* are often found in this position, so it is possible that the classifier draws on this knowledge to make its (incorrect) class assignation. Nor is this problem restricted to sentence-initial clauses, as the following examples of wrongly corrected erroneous instances show:

(10) It's all made **#RT with** [**of**] **/#RT** metal.
Classifier decision: **IN**

(11) The problem was that the software **#RT in** [**of**] **/#RT** our company didn't work.
Classifier decision: **FOR**

Here, we observe that the structure of the sentence does not follow an order which sounds natural in L1 English, consistent with a general tendency found in L2 writing to underuse the possessive marker in favor of the *of*-construction (which may be closer to the structure of their L1). Such differences in sentence organization and choice of phrases often have a negative effect on the system's performance.[8]

We can therefore see that differences in text type may have an impact on performance after all because the classifier appears less attuned to the more informal kind of language used by students and to the narrative/expository texts they produce. In this respect, then, the Tetreault and Chodorow (2008b) approach of including school textbook data in training may be beneficial since it will be similar to the types of texts seen in testing.

Classifier Issues

Finally, we also observe that problems often arise because of the way Dapper is currently set up, namely, the fact that it outputs only one possible class for each instance submitted.

This means that even where more than one preposition is found normally associated with a particular lexical item, only the strongest of these associations will be given. Certain lexical items are particularly susceptible to this, as is evident from the following examples, where the words appear with an incorrect preposition, but the correction suggested by Dapper, ***of***, is equally wrong.

(12) There is a nice view **#RT in** [**from**] **/#RT** my window.
Classifier decision: **OF**, influenced *by a view of … a city*, *castle*, and so on

(13) This can be a #RJ bonus /#RJ quality **#RT to** [**in**] **/#RT** a person.
Classifier decision: **OF**, influenced by *quality of life*, *quality of this product*, and so on

The verb *look* is also regularly involved in these mistakes and is always associated with the preposition ***at***. This is problematic where the phrase is actually *looking forward **to***, a frequent phrase in the data which is regularly misused by students as *looking forward **for*** since the error goes uncorrected.

This problem lies in the way the classifier is currently set up rather than the data. One simple way of overcoming it is to allow Dapper to output and rank more than one choice showing learners what alternatives are available; however, this solution would only be of assistance to more advanced or confident students.

In conclusion, in light of the analysis in this section, the trends in precision and recall are largely explained as being driven by the peculiarity of the texts used to assess Dapper. Generally, there is much overlap among the issues encountered in the correct and incorrect L2 data tasks and, to an extent, even with the L1 data. Differences stand out in text type and in the distribution of vocabulary, and certain confusion pairs involving particular lexical items exert a very strong influence on the system's decisions.

From a pedagogical perspective, the claim that L2 writers tend to rely on a small set of fixed expressions is reinforced. It appears that the texts produced by these students, though generally not ungrammatical, are nevertheless rather different from standard English texts: students should be encouraged to try a variety of formulations and syntactic structures, if their final aim is to sound as native-like and fluent as possible.

Problems in Using L2 Data

In the section on the testing of L2 data above, a number of features of L2 writing were introduced which were believed to cause problems for the classifier. In fact, the issues affecting performance have proved to be partly different: misspellings and other ill-formed input are not the main source of difficulty for Dapper.[9]

However, such errors do occur and are often found in instances in which the classifier's choice is incorrect. Some examples of misspelled words include: *by all meance* (*means*), *viewus of* (*views*), *with enthousiasm*. In all these cases, the prepositions chosen by the student are actually correct and indeed have a fairly strong collocational link to the lexical item, but Dapper does not recognize them as what they are intended to be and chooses a different, incorrect preposition instead. This kind of problem could potentially be solved by introducing a spell check filter at the preprocessing stage, although of course spelling errors resulting in real words would not be detected. Further experiments are needed to assess the feasibility of such a filter.

Grammatical errors, too, sometimes lead to incorrect class assignment, as in the following example:

(14) This amount will be **increase** [**increased**] **with** premiums for special wins.

Here, the verb *increase* is in the wrong form. This leads the parser to tag it as a noun rather than a verb, which in turn triggers the choice of the preposition ***in*** for that particular context (cf. *an increase in* …), which is of course a mistake on Dapper's part. This is an example of a well-formed sentence where a single grammatical error is sufficient to lead the classifier astray. Unfortunately, these errors are also hard to detect or filter out without relying on the annotators' tagging, which would not be present in instances not taken from the corpus.

The main factor affecting Dapper's performance is the different syntactic structure used by learners. This is represented both by a higher-than-average use of particular phrases and discourse markers and by a tendency to place any kind of adjunct clauses (not just temporal and locative ones) in sentence-initial position. While not ungrammatical, these rhetorical strategies are perhaps not always ones that L1 English writers would use and are especially typical of text types which are not very frequent in the BNC (e.g., student essays).

CONCLUSIONS AND FUTURE OUTLOOK

In this article, we have presented a model for assigning prepositions to specific contexts, which performs with 70% accuracy on L1 data and up to 69% accuracy on L2 data. On incorrect preposition instances, our system yields an average precision of 42% and average recall of 35%; we have discussed in detail several of the factors which impair the system's performance.

Some of the problems identified can be addressed by further improving the usage model acquired for L1 prepositions, as well as better treatment of cases in which more than one preposition may be appropriate. Issues relating to the use of NLP tools with learner language must also be taken into account. Foremost among these are the differences between the texts used in training and those produced by L2 writers. By including different types of text at the training stage, such as a corrected version of the CLC itself, we could make the model more familiar with characteristics of learner writing (e.g., the use of adjuncts in initial position and particular fixed phrases). Another possible solution to the issue of fixed phrases being mislabeled would be to enhance the model so that it recognizes a particular set of semi-fixed expressions and always defaults to the correct preposition for those expressions. This could be done fairly easily since, as we have seen, the set of such expressions is actually not very large.

We did not expect human learners and the NLP model to run into the same kinds of problems. However, there are certainly aspects of the language which prove problematic for both, for example, confusion about temporal PPs. This means that the incorrect L2 data are likely to have a high proportion of instances which the classifier has been shown to be poor at resolving—in either L1 or L2 tasks—which cannot but bring down its performance. We expect the impact of such problems to be reduced if further improvements are made to the underlying L1 model.

This contextual feature-based approach may find applicability beyond the domain of frequent prepositions. For example, we have also developed a component designed to perform an analogous task on determiners (see De Felice, 2008; De Felice & Pulman, 2008)

which achieves 92% accuracy on L1 data; work on L2 data is still underway. Less frequent prepositions have not yet been included in the study, but it would be interesting to investigate the effect of smaller amounts of training data on performance.

NOTES

[1] This is in addition to potential confusion caused by lack of correspondence in the preposition systems of the learner's L1 and L2.

[2] The Web 1T corpus, also known as the Google n-gram corpus (Brants & Franz, 2006), is a collection of 1 trillion words collected from the web, arranged in n-grams from 1 to 5.

[3] This compares favorably to Lee and Knutsson (2008) and to the score of 69% reported on L1 text by Chodorow et al. (2007).

[4] Cohen's kappa for Subject 1 = 0.867 and Subject 2 = 0.860; intersubject agreement is kappa = 0.842. All three scores are considered 'very good.'

[5] This has been made available to us by Cambridge University Press, whose assistance is gratefully acknowledged.

[6] For a good introduction to these issues, see among others Meunier (1998).

[7] Assuming, of course, that the utterance is essentially well formed; there are cases where a combination of errors makes it impossible for the parser to assign a correct parse or indeed any parse.

[8] Arguably the sentences shown here would sound much more 'correct' if they were rewritten entirely to eliminate the use of the preposition *of* in this way, but that is not in line with the error annotation guidelines.

[9] It could still be argued that a text in which prepositions are used correctly is also more likely to be generally well formed.

REFERENCES

Brants, T., & Franz, A. (2006). Web 1t 5-gram version 1. Philadelphia: Linguistic Data Consortium.

Burnard, L. (Ed.). (2000). *The British national corpus users reference guide*. Oxford: British National Corpus Consortium, Oxford University Computing Services.

Chodorow, M., Tetreault, J., & Han, N. (2007). Detection of grammatical errors involving prepositions. In F. Costello, J. Kelleher, & M. Volk (Eds.), *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions* (pp. 25-30). Prague, Czech Republic: Association for Computational Linguistics.

Clark, S., & Curran, J. (2007). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics, 33*, 493-552.

De Felice, R. (2008). *Automatic error detection in non-native English*. Unpublished doctoral dissertation, Oxford University Computing Laboratory.

De Felice, R., & Pulman, S. (2008). A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of COLING* (pp. 169-176). Manchester, UK: Coling 2008 Organizing Committee. Retrieved April 20, 2009, from http://www.aclweb.org/anthology-new/C/C08

Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W., Belenko, D., et al. (2008). Using contextual speller techniques and language modeling for ESL error correction. In Y. Matsumoto & A. Copestake (Eds.), *Proceedings of the International Joint Conference on Natural Language Processing* (pp. 449-456). Hyderabad, India: Asian Federation of Natural Language Processing.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 145-160). Oxford: Oxford University Press.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). SST speech corpus of Japanese learners' English and automatic detection of learners' errors. *ICAME, 28*, 31-48.

Lee, J., & Knutsson, O. (2008). The role of PP attachment in preposition generation. In A. Gelbukh (Ed.), *Proceedings of CICling* (pp. 643-658). Berlin/Heidelberg: Springer.

Meunier, F. (1998). Computer tools for the analysis of learner corpora. In S. Granger (Ed.), *Learner English on computer* (pp. 19-37). London: Longman.

Minnen, G., Carroll, J., & Pearce, D. (2001). Applied morphological processing of English. *Natural Language Engineering, 7*, 207-223.

Nicholls, D. (2003). The Cambridge learner corpus—Error coding and analysis for lexicography. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 572-581). Lancaster, UK: UCREL, Lancaster University.

Parrott, M. (2000). *Grammar for English language teachers*. Cambridge: Cambridge University Press.

Tetreault, J., & Chodorow, M. (2008a). Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of the Coling 2008 Workshop on Human Judgments in Computational Linguistics* (pp. 24-32). Manchester, UK: Coling 2008 Organizing Committee. Retrieved April 20, 2009, from http://www.aclweb.org/anthology-new/C/C08

Tetreault, J., & Chodorow, M. (2008b). The ups and downs of preposition error detection in ESL writing. In *Proceedings of Coling* (pp. 865-872). Manchester, UK: Coling 2008 Organizing Committee. Retrieved April 20, 2009, from http://www.aclweb.org/anthology-new/C/C08

APPENDIX

Table 5
Confusion Matrix for Prepositions: L1 Data

Target preposition	Confused with								
	at	by	for	from	in	of	on	to	with
at	–	4.65%	10.82%	2.95%	36.83%	19.46%	9.17%	10.28%	5.85%
by	6.54%	–	8.50%	2.58%	41.38%	19.44%	5.41%	10.04%	6.10%
for	8.19%	3.93%	–	1.91%	25.67%	36.12%	5.60%	11.29%	7.28%
from	6.19%	4.14%	6.72%	–	26.98%	26.74%	7.70%	16.45%	5.07%
in	7.16%	9.28%	10.68%	3.01%	–	43.40%	10.92%	8.96%	6.59%
of	3.95%	2.00%	18.81%	3.36%	40.21%	–	9.46%	14.77%	7.43%
on	5.49%	3.85%	8.66%	2.29%	32.88%	27.92%	–	12.20%	6.71%
to	9.77%	3.82%	11.49%	3.71%	24.86%	27.95%	9.43%	–	8.95%
with	3.66%	4.43%	12.06%	2.24%	28.08%	26.63%	6.81%	16.10%	–

Table 6
Confusion Matrix for Prepositions: Correct L2 Data

Target preposition	Confused with								
	at	by	for	from	in	of	on	to	with
at	–	3.62%	10.87%	2.17%	47.83%	15.94%	7.25%	5.80%	6.52%
by	16.35%	–	6.73%	0.00%	49.04%	8.65%	8.65%	10.58%	0.00%
for	9.33%	5.22%	–	1.49%	29.85%	35.07%	6.34%	7.09%	5.60%
from	15.89%	4.64%	8.61%	–	33.77%	17.22%	7.95%	9.27%	2.65%
in	16.91%	7.35%	19.85%	2.21%	–	35.54%	9.07%	5.64%	3.43%
of	3.88%	2.91%	13.59%	0.97%	54.37%	–	6.80%	12.62%	4.85%
on	9.63%	2.67%	4.81%	1.07%	54.55%	13.90%	–	7.49%	5.88%
to	25.20%	3.94%	8.27%	3.15%	27.56%	22.44%	6.30%	–	3.15%
with	2.89%	6.36%	12.14%	1.73%	35.26%	25.43%	8.67%	7.51%	–

ACKNOWLEDGMENTS

We wish to thank Stephen Clark and Laura Rimell for stimulating discussions and helpful observations, and the three anonymous reviewers for their insightful comments. We acknowledge Cambridge University Press's assistance in accessing the Cambridge Learner Corpus data. Rachele De Felice was supported by an AHRC scholarship for the duration of her studies.

AUTHORS' BIODATA

Rachele De Felice completed her Ph.D. at Oxford University in 2008 on the automatic detection of preposition and determiner errors in learner writing. She is currently working as a Postdoctoral Research Fellow at Educational Testing Service where she continues to address issues in L2 English with a particular focus on speech acts and L2 pragmatics.

Stephen Pulman is Professor of Computational Linguistics at Oxford University Computing Laboratory. His interests include formal and computational semantics for natural language, automated reasoning and language, and combining statistical and symbolic models of language.

AUTHOR'S ADDRESS

Rachele De Felice
Educational Testing Service
Rosedale Road MS R-11
Princeton, NJ 08541
U.S.A.
Phone: 609 734 5039
Email: rdefelice@ets.org