Exploring the Data-Driven Prediction of Prepositions in English

Anas Elghafari **Detmar Meurers** **Holger Wunsch**
Seminar für Sprachwissenschaft
Universität Tübingen
{aelgafar,dm,wunsch}@sfs.uni-tuebingen.de

Abstract

Prepositions in English are a well-known challenge for language learners, and the computational analysis of preposition usage has attracted significant attention. Such research generally starts out by developing models of preposition usage for native English based on a range of features, from shallow surface evidence to deep linguistically-informed properties.

While we agree that ultimately a combination of shallow and deep features is needed to balance the preciseness of exemplars with the usefulness of generalizations to avoid data sparsity, in this paper we explore the limits of a purely surface-based prediction of prepositions.

Using a web-as-corpus approach, we investigate the classification based solely on the relative number of occurrences for target n-grams varying in preposition usage. We show that such a surface-based approach is competitive with the published state-of-the-art results relying on complex feature sets.

Where enough data is available, in a surprising number of cases it thus is possible to obtain sufficient information from the relatively narrow window of context provided by n-grams which are small enough to frequently occur but large enough to contain enough predictive information about preposition usage.

1 Introduction

The correct use of prepositions is a well-known difficulty for learners of English, and correspondingly the computational analysis of preposition usage has attracted significant attention in recent years (De Felice and Pulman, 2007; De Felice, 2008; Lee and Knutsson, 2008; Gamon et al., 2008; Chodorow et al., 2007; Tetreault and Chodorow, 2008a, 2008b).

As a point of reference for the detection of preposition errors in learner language, most of the research starts out by developing a model of preposition usage for native English. For this purpose, virtually all previous approaches employ a machine learning setup combining a range of features, from surface-based evidence to deep linguistically-informed properties. The overall task is approached as a classification problem where the classes are the prepositions and the instances to be classified are the contexts, i.e., the sentences with the prepositions omitted.

A focus of the previous literature is on the question which linguistic and lexical features are the best predictors for preposition usage. Linguistic features used include the POS tags of the surrounding words, PP attachment sites, WordNet classes of PP object and modified item. Lexical features used include the object of the PP and the lexical item modified by the PP. Those syntactic, semantic and lexical features are then extracted from the training instances and used by the machine learning tool to predict the missing preposition in a test instance.

While we agree that ultimately a combination of shallow and linguistically informed features is needed to balance the preciseness of exemplars

Coling 2008: Poster Volume, pages 267–275,
Beijing, August 2010

with the usefulness of generalizations to avoid data sparsity problems, in this paper we want to explore the limits of a purely surface-based prediction of prepositions. Essentially, our question is how much predictive information can be found in the immediate distributional context of the preposition. Is it possible to obtain n-gram contexts for prepositions which are small enough to occur frequently enough in the available training data but large enough to contain enough predictive information about preposition usage?

This perspective is related to that underlying the variation-n-gram approach for detecting errors in the linguistic annotation of corpora (Dickinson and Meurers, 2003; Dickinson and Meurers, 2005; Boyd et al., 2008). Under that approach, errors in the annotation of linguistic properties (lexical, constituency, or dependency information) are detected by identifying units which recur in the corpus with sufficient identical context so as to make variation in their annotation unlikely to be correct. In a sense, the recurring n-gram contexts are used as exemplar references for the local domains in which the complex linguistic properties are established. The question now is to what extent basic[1] n-gram contexts can also be successfully used to capture the linguistic properties and relations determining preposition usage, exploring the trade-off expressed in the question ending the previous paragraph.

To address this question, in this paper we make use of a web-as-corpus approach in the spirit of Lapata and Keller (2005). We employ the Yahoo search engine to investigate a preposition classification setup based on the relative number of web counts obtained for target n-grams varying in the preposition used. We start the discussion with a brief review of key previous approaches and the results they obtain for the preposition classification task in native English text. In section 2, we then describe the experimental setup we used for our exploration and discuss our results in section 3.

1.1 Previous work and results

The previous work on the preposition prediction task varied in i) the features selected, ii) the number of prepositions tackled, and iii) the training and testing corpora used.

De Felice (2008) presents a system that (among other things) is used to predict the correct preposition for a given context. The system tackles the nine most frequent prepositions in English: *of, to, in, for, on, with, at, by, from*. The approach uses a wide variety of syntactic and semantic features: the lexical item modified by the PP, the lexical item that occurs as the object of the preposition, the POS tags of three words to the left and three words to the right of the preposition, the grammatical relation that the preposition is in with its object, the grammatical relation the preposition is in with the word modified by the PP, and the WordNet classes of the preposition's object and the lexical item modified by the PP. De Felice (2008) also used a named entity recognizer to extract generalizations about which classes of named entities can occur with which prepositions. Further, the verbs' subcategorization frames were taken as features. For features that used lexical sources (WordNet classes, verbs subcategorization frames), only partial coverage of the training and testing instances is available.

The overall accuracy reported by De Felice (2008) for this approach is 70.06%, testing on section J of the *British National Corpus (BNC)* after training on the other sections. As the most extensive discussion of the issue, using an explicit set of prepositions and a precisely specified and publicly accessible test corpus, De Felice (2008) is well-suited as a reference approach. Correspondingly, our study in this paper is based on the same set of prepositions and the same test corpus.

Gamon et al. (2008) introduce a system for the detection of a variety of learner errors in non-native English text, including preposition errors. For the preposition task, the authors combine the outputs of a classifier and a language model. The language model is a 5-gram model trained on the English Gigaword corpus. The classifier is trained

[1] While Dickinson and Meurers (2005) also employ discontinuous n-grams, we here focus only on contiguous n-gram contexts. Using discontinuous n-gram contexts for preposition prediction could be interesting to explore in the future, once, as a prerequisite for the effective generation of discontinuous n-grams, heuristics have been identified for when which kind of discontinuities should be allowed to arise for preposition classification contexts.

on Encarta encyclopedia and Reuters news text. It operates in two stages: The presence/absence classifier predicts first whether a preposition needs to be inserted at a given location. Then, the choice classifier determines which preposition is to be inserted. The features that are extracted for each possible insertion site come from a six-token window around the possible insertion site. Those features are the relative positions, POS tags, and surface forms of the tokens in that window. The choice classifier predicts one of 13 prepositions: *in, for, of, on, to, with, at, by, as, from, since, about, than,* and *other*. The accuracy of the choice classifier, the part of the system to which the work at hand is most similar, is 62.32% when tested on text from Encarta and Reuters news.

Tetreault and Chodorow (2008a) present a system for detecting preposition errors in learner text. Their approach extracts a total of 25 features from the local contexts: the adjacent words, the heads of the nearby phrases, and the POS tags of all those. They combine word-based features with POS tag features to better handle cases where a word from the test instance has not been seen in training. For each test instance, the system predicts one of 34 prepositions. In training and testing performed on the Encarta encyclopedia, Reuters news text and additional training material an accuracy figure of 79% is achieved.

Bergsma et al. (2009) extract contextual features from the Google 5-gram corpus to train an SVM-based classifier for predicting prepositions. They evaluate on 10 000 sentences taken from the New York Times section of the Gigaword corpus, and achieve an accuracy of 75.4%.

Following De Felice (2008, p. 66), we summarize the main results of the mentioned approaches to preposition prediction for native text in Figure 1.[2] Since the test sets and the prepositions targeted differ between the approaches, such a comparison must be interpreted with caution. In terms of the big picture, it is useful to situate the results with respect to the majority baseline reported by De Felice (2008). It is obtained by always choosing *of* as the most common preposition in section J of the BNC. De Felice also reports another interesting figure included in Figure 1, namely the accuracy of the human agreement with the original text, averaged over two English native-speakers.

Approach	Accuracy
Gamon et al. (2008)	62.32%
Tetreault and Chodorow (2008a)	79.00%
Bergsma et al. (2009)	75.50%
De Felice (2008) system	70.06%
Majority baseline *(of)*	26.94%
Human agreement	88.60%

Figure 1: Preposition prediction results

2 Experiments

2.1 Data

As our test corpus, we use section J of the BNC, the same corpus used by De Felice (2008). Based on the tokenization as given in the corpus, we join the tokens with a single space, which also means that punctuation characters end up as separate, white-space separated tokens. We select all sentences that contain one or more prepositions, using the POS annotation in the corpus to identify the prepositions. The BNC is POS-annotated with the CLAWS-5 tagset, which distinguishes the two tags `PRF` for *of* and `PRP` for all other prepositions.[3] We mark every occurrence of these preposition tags in the corpus, yielding one prediction task for each marked preposition. For example, the sentence (1) yields four prediction tasks, one for each of the prepositions *for, of, from,* and *in* in the sentence.

(1) But **for** the young, it is rather a question **of** the scales falling **from** their eyes, and having nothing to believe **in** any more.

In each task, one preposition is masked using the special marker `-*-MASKED-*-`. Figure 2 shows the four marked-up prediction tasks resulting for example (1).

Following De Felice (2008), we focus our experiments on the top nine prepositions in the BNC: *of, to, in, for, on, with, at, by, from*. For

[2] The Gamon et al. (2008) result differs from the one reported in De Felice (2008); we rely on the original paper.

[3] `http://www.natcorp.ox.ac.uk/docs/URG/posguide.html#guidelines`

```
But -*-MASKED-*-for the young , it is
rather a question of the scales falling
from their eyes , and having nothing to
believe in any more .

But for the young , it is rather a
question -*-MASKED-*-of the scales
falling from their eyes , and having
nothing to believe in any more .

But for the young , it is rather
a question of the scales falling
-*-MASKED-*-from their eyes , and having
nothing to believe in any more .

But for the young , it is rather a
question of the scales falling from
their eyes , and having nothing to
believe -*-MASKED-*-in any more .
```

Figure 2: Four prediction tasks for example (1)

each occurrence of these nine prepositions in section J of the BNC, we extract one prediction task, yielding a test set of 522 313 instances.

Evaluating on this full test set would involve a prohibitively large number of queries to the Yahoo search engine. We therefore extract a randomly drawn subset of 10 000 prediction tasks. From this subset, we remove all prediction tasks which are longer than 4000 characters in length, as Yahoo only supports queries up to that length. Finally, in a web-as-corpus setup, the indexing of the web pages performed by the search engine essentially corresponds to the training step in a typical machine learning setup. In order to avoid testing on the training data, we thus need to ensure that the test cases are based on text not indexed by the search engine. To exclude any such cases, we query the search engine with each complete sentence that a prediction task is based on and remove any prediction task for which the search engine returns hits for the complete sentence. The final test set consists of 8060 prediction tasks.[4]

2.2 Experimental Setup

Recall that the general issue we are interested in is whether one can obtain sufficient information from the relatively narrow distributional window of context provided by n-grams which are small enough to occur frequently enough in the training data but large enough to contain enough predictive information about preposition usage for the instances to be classified. By using a web-as-corpus approach we essentially try to maximize the training data size. For the n-gram size, we explore the use of a maximum order of 7, containing the preposition in the middle and three words of context on either side.

For each prediction task, we successively insert one of the nine most frequent prepositions into the marked preposition slot of the 8060 n-grams obtained from the test set. Thus, for each prediction task, we get a *cohort* consisting of nine different individual queries, one query for each potential preposition. For example, the second prediction task of Figure 2 yields the cohort of nine queries in Figure 3 below, where the candidate prepositions replace the location marked by `-*-MASKED-*-of`. The correct preposition *of* is stripped off and kept for later use in the evaluation step.

1. `rather a question` **`of`** `the scales falling`
2. `rather a question` **`to`** `the scales falling`
3. `rather a question` **`in`** `the scales falling`

   ⋮

9. `rather a question` **`from`** `the scales falling`

Figure 3: Cohort of nine queries resulting for the second prediction task of Figure 2

In cases where a preposition is closer than four words to the beginning or the end of the corresponding sentence, a lower-order n-gram results. For example, in the first prediction task in Figure 2, the preposition occurs already as the second word in the sentence, thus not leaving enough context to the left of the preposition for a symmetric 7-gram. Here, the truncated asymmetric 5-gram "`But` **`<prep>`** `the young ,`" including only one word of context on the left would get used.

We issue each query in a cohort to the Yahoo search engine, and determine the number of hits returned for that query. To that end, we use Yahoo's BOSS service, which offers a

[4] For a copy of the test set, just send us an email.

JSON interface supporting straightforward automated queries. As part of its response to a query, the BOSS service includes the `deephits` field, which gives an "approximate count that reflects duplicate documents and all documents from a host".[5] In other words, this number is an approximate measure of how many web pages there are that contain the search pattern.

With the counts for all nine queries in a cohort retrieved from Yahoo, we select the preposition of the query with the highest count. For the cases in which none of the counts in a 7-gram cohort is greater than zero, we use one of two strategies:

In the **baseline** condition, for all n-gram cohorts with zero counts (5160 out of the 8060 cases) we predict the most frequent preposition *of*, i.e., the majority baseline. This results in an overall accuracy of 50%.

In the **full back-off** condition, we explore the trade-off between the predictive power of the n-gram as context and the likelihood of having seen this n-gram in the training material, i.e., finding it on the web. In this paper we never abstract or generalize away from the surface string (e.g., by mapping all proper names to an abstract name tag; but see the outlook discussion at the end of the paper), so the only option for increasing the number of occurrences of an n-gram is to approximate it with multiple shorter n-grams.

Concretely, if no hits could be found for any of the queries in a cohort, we back off to the sum of the hits for the two overlapping 6-grams constructed in the way illustrated in Figure 4.

```
[rather a question of the scales falling]
                    ⇓
[rather a question of the scales]
       [a question of the scales falling]
```

Figure 4: Two overlapping 6-grams approximate a 7-gram for back-off.

If still no hits can be obtained after backing off to 6-grams for any of the queries in a cohort, the system backs off further to overlapping 5-grams, and so on, down to trigrams.[6]

[5]Cited from `http://developer.yahoo.com/search/boss/boss_guide/ch02s02.html`

[6]When backing off, the left-most and the right-most trigram do not include the target preposition of the original 7-gram. However, this only affects 13 cases, cf. Figure 6.

3 Results

Figure 5 shows the results of the **full back-off** approach. Compared to the baseline condition, accuracy goes up significantly to 76.5%. Thus, the back-off strategy is effective in increasing the amount of available data using lower-order n-grams. This increase of data is also reflected in the number of cases with zero counts for a cohort, which goes down to none.

	Full back-off
Correct	6166
Incorrect	1894
Total	8060
Accuracy	76.5%

Figure 5: Overall results of our experiments.

Figure 6 provides a detailed analysis of the back-off experiment. It lists back-off sequences separately for each maximum n-gram order. The prediction tasks for which a full 7-gram can be extracted are displayed in the third column, with back-off orders of 6 down to 3. Prediction tasks for which only asymmetric 6-grams can be extracted follow in column 4, and so on until 4-grams. There are no predictions tasks that are shorter than four words. Therefore, n-grams with a length of less than 4 do not occur.

The "sum" column shows the combined results of the full 7-gram prediction tasks and the prediction tasks involving truncated, asymmetric n-grams of lower orders.

There are 6999 prediction tasks for which full 7-grams can be extracted. The remaining 1061 of the 8060 prediction tasks are the cases where the system extracts only asymmetric lower-order n-grams, for the reasons explained in section 2.2.

For 2195 of the 6999 7-gram prediction tasks, we find full 7-gram contexts on the web, of which 1931 lead to a correct prediction, and 264 to an incorrect one, leaving 4804 prediction tasks still to be solved through the back-off approach. Thus, full 7-gram contexts lead to high-quality predictions at 88% precision, but they are rare and with a recall of 28,7% cover only a fraction of all cases.

	sum	**7-grams** (3 + prep + 3)	**6-grams** (truncated 7-gram)	**5-grams** (truncated 7-gram)	**4-grams** (truncated 7-gram)
Total	8060	6999	656	182	223
Predictions	2900	2195	379	119	207
correct	*2495*	*1931*	*326*	*91*	*147*
incorrect	*405*	*264*	*53*	*28*	*60*
Requiring back-off	5160	4804	277	63	16
Precision	86%	88%	86%	76.5%	71%
Recall	32.6%	28.7%	79.6%	59.1%	90.2%
		Back-off order 6			
Predictions	2028	2028			
correct	*1620*	*1620*			
incorrect	*408*	*408*			
Still requiring back-off	2776	2776			
Predict. orders 7+6	4223	4223			
correct	*3551*	*3551*			
incorrect	*672*	*672*			
Precision	84.1%	84.1%			
Recall	56.1%	56.1%			
		Back-off order 5			
Predictions	2180	2020	160		
correct	*1542*	*1411*	*131*		
incorrect	*638*	*609*	*29*		
Still requiring back-off	873	756	117		
Predict. orders 7 – 5	6782	6243	539		
correct	*5419*	*4962*	*457*		
incorrect	*1363*	*1281*	*82*		
Precision	79.9%	79.5%	84.8%		
Recall	86.1%	86.8%	79.6%		
		Back-off order 4			
Predictions	905	743	106	56	
correct	*488*	*382*	*68*	*38*	
incorrect	*417*	*361*	*38*	*18*	
Still requiring back-off	31	13	11	7	
Predict. orders 7 – 4	7806	6986	645	175	
correct	*5998*	*5344*	*525*	*129*	
incorrect	*1808*	*1642*	*120*	*46*	
Precision	76.8%	76.5%	81.4%	73.7%	
Recall	99.5%	99.8%	97.9%	94.9%	
		Back-off order 3			
Predictions	47	13	11	7	16
correct	*21*	*5*	*7*	*3*	*6*
incorrect	*26*	*8*	*4*	*4*	*10*
Still requiring back-off	0	0	0	0	0
Predict. orders 7 – 3	8060	6999	656	182	223
correct	*6166*	*5349*	*532*	*132*	*153*
incorrect	*1894*	*1650*	*124*	*50*	*70*
Precision	76.5%	76.4%	81.1%	72.5%	68.6%
Recall	100%	100%	100%	100%	100%

Figure 6: The results of our experiments

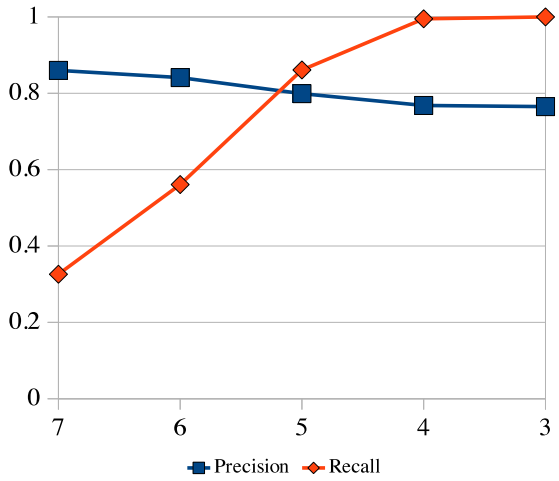

Figure 7: Development of precision and recall in relation to back-off order

Approximating 7-grams with two overlapping 6-grams as the first back-off step provides the evidence needed to correctly predict 1620 additional prepositions, with 408 additional false predictions. The number of correctly solved prediction tasks thus rises to 3551, and the number of incorrect predictions rises to 672. This back-off step almost doubles recall (56.1%). At the same time, precision drops to 84.1%. For 2776 prediction tasks, a further back-off step is necessary since still no evidence can be found for them. This pattern repeats with the back-off steps that follow. To summarize, by adding more data using less restricted contexts, more prediction tasks can be solved. The better coverage however comes at the price of reduced precision: Less specific contexts are worse predictors of the correct preposition than more specific contexts.

Figure 7 visualizes the development of precision and recall with full and truncated 7-grams counted together as in the "sum" column in Figure 6. With each back-off step, more prediction tasks can be solved (as shown by the rising recall curve). At the same time, the overall quality of the predictions drops due to the less specific contexts (as shown by the slightly dropping precision curve). While the curve for recall rises steeply, the curve for precision remains relatively flat. The back-off approach thus succeeds in adding data while preserving prediction quality.

As discussed above, we use the same set of prepositions and test corpus as De Felice (2008), but only make use of 8060 test cases. Figure 8 shows that the accuracy stabilizes quickly after about 1000 predictions, so that the difference in the size of the test set should have no impact on the reported results.

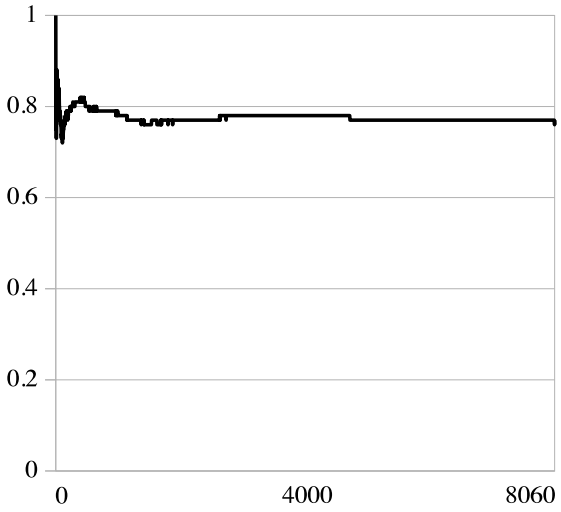

Figure 8: The accuracy of the n-gram prediction stabilizes quickly.

4 Conclusions and Outlook

In this paper, we explored the potential and the limits of a purely surface-based strategy of predicting prepositions in English. The use of surface-based n-grams ensures that fully specific exemplars of a particular size are stored in training, but avoiding abstractions in this way leads to the well-known data sparsity issues. We showed that using a web-as-corpus approach maximizing the size of the "training data", one can work with n-grams which are large enough to predict the occurrence of prepositions with significant precision while at the same time ensuring that these specific n-grams have actually been encountered during "training", i.e., evidence for them can be found on the web.

For the random sample of the BNC section J we tested on, the surface-based approach results in an accuracy of 77% for the 7-gram model with back-off to overlapping shorter n-grams. It thus outperforms De Felice's (2008) machine learning

approach which uses the same set of prepositions and the full BNC section J as test set. In broader terms, the result of our surface-based approach is competitive with the state-of-the art results for preposition prediction in English using machine learning to combine sophisticated sets of lexical and linguistically motivated features.

In this paper, we focused exclusively on the impact of n-gram size on preposition prediction. Limiting ourselves to pure surface-based information made it possible to maximize the "training data" by using a web-as-corpus approach. Returning from this very specific experiment to the general issue, there are two well-known approaches to remedy the data sparseness problem arising from storing large, specific surface forms in training. On the one hand, one can use smaller exemplars, which is the method we used as backoff in our experiments in this paper. This only works if the exemplars contain enough context for the linguistic property or relation that we need to capture the predictive power. On the other hand, one can abstract parts of the surface-based training instances to more general classes. The crucial question this raises is which generalizations preserve the predictive power of the exemplars and can reliably be identified. The linguistically-informed features used in the previous approaches in the literature naturally provide interesting instances of answers to this question. In the future, we intend to compare the results we obtained using the web-as-corpus approach with one based on the Google-5-gram corpus to study using controlled, incremental shallow-to-deep feature development which abstractions or linguistic generalizations best preserve the predictive context while lowering the demands on the size of the training data.

Turning to a linguistic issue, it could be useful to distinguish between lexical and functional prepositions when reporting test results. This is an important distinction because the information needed to predict functional prepositions typically is in the local context, whereas the information needed to predict lexical prepositions is not necessarily present locally. To illustrate, a competent human speaker presented with the sentence *John is dependent ___ his brother* and asked to fill in the missing preposition, would correctly pick *on*. This is a case of a functional preposition where the relevant information is locally present: the adjective *dependent* selects *on*. On the other hand, the sentence *John put his bag ___ the table* is more problematic, even for a human, since both *on* and *under* are reasonable choices; the information needed to predict the omitted preposition in this case is not locally present. In line with the previous research, in the work in this paper we made predictions for all prepositions alike. In the future, it could be useful to annotate the test set so that one can distinguish functional and lexical uses and report separate figures for these two classes in order to empirically confirm their differences with respect to locality.

References

Bergsma, Shane, Dekang Lin, and Randy Goebel. 2009. Web-scale n-gram models for lexical disambiguation. In *IJCAI'09: Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1507–1512, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Boyd, Adriane, Markus Dickinson, and Detmar Meurers. 2008. On detecting errors in dependency treebanks. *Research on Language and Computation*, 6(2):113–137.

Chodorow, Martin, Joel Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30, Prague, Czech Republic, June.

De Felice, Rachele and Stephen Pulman. 2007. Automatically acquiring models of preposition use. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 45–50, Prague, Czech Republic, June. Association for Computational Linguistics.

De Felice, Rachele. 2008. *Automatic Error Detection in Non-native English*. Ph.D. thesis, St Catherine's College, University of Oxford.

Dickinson, Markus and W. Detmar Meurers. 2003. Detecting errors in part-of-speech annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pages 107–114, Budapest, Hungary.

Dickinson, Markus and W. Detmar Meurers. 2005. Detecting errors in discontinuous structural anno-

tation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 322–329.

Gamon, Michael, Jianfeng Gao, Chris Brockett, Alexander Klementiev, William Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using contextual speller techniques and language modeling for esl error correction. In *Proceedings of IJCNLP*, Hyderabad, India.

Lapata, Mirella and Frank Keller. 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2(1):1–30, February.

Lee, John and Ola Knutsson. 2008. The role of pp attachment in preposition generation. In Gelbukh, A., editor, *Proceedings of CICLing 2008*.

Tetreault, Joel and Martin Chodorow. 2008a. Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of COLING-08*, Manchester.

Tetreault, Joel and Martin Chodorow. 2008b. The ups and downs of preposition error detection in esl writing. In *Proceedings of COLING-08*, Manchester.