Measuring Interlanguage: Native Language Identification with L1-influence Metrics

Julian Brooke and Graeme Hirst

Department of Computer Science, University of Toronto
Toronto, ON, Canada M5S 3G4
jbrooke, gh@cs.toronto.edu

Abstract

The task of native language (L1) identification suffers from a relative paucity of useful training corpora, and standard within-corpus evaluation is often problematic due to topic bias. In this paper, we introduce a method for L1 identification in second language (L2) texts that relies only on much more plentiful L1 data, rather than the L2 texts that are traditionally used for training. In particular, we do word-by-word translation of large L1 blog corpora to create a mapping to L2 forms that are a possible result of language transfer, and then use that information for unsupervised classification. We show this method is effective in several different learner corpora, with bigram features being particularly useful.

Keywords: Native Language Identification, Unsupervised Text Classification, Interlanguage

1. Introduction

In Second Language Acquisition (SLA) research, an *interlanguage*[1] is an emerging second language (L2) system (Selinker, 1992). One of the defining qualities of an interlanguage is the use of native language (L1) features, a phenomenon which is known more generally as *language transfer* (Odlin, 1989). Though in related languages this may provide a early boost to learning, *language interference* is often the result where the two systems differ significantly, with learners continuing to use L1 features that are not appropriate to the L2, even after years of exposure.

In computational linguistics, native language identification (Koppel et al., 2005) is a task in which features of the L2 texts written by learners of various different language backgrounds are used to identify those language backgrounds. One potential application of this is in author profiling, which can be used to identify those who misrepresent themselves online (Fette et al., 2007). Another important use is as a preprocessing step to ESL error correction (Leacock et al., 2010): for example, Rozovskaya and Roth (2011) use L1-specific information to improve their preposition-correction system, while recent work in collocation correction relies on the specific forms present in the native language (Chang et al., 2008b; Dahlmeier and Ng, 2011).

Most previous work in L1 identification has avoided standard lexical features (e.g. word *n*-grams); the reason for this is not that these features would not be useful, but rather that there is significant topic variation across the languages in the corpora used for this task. Our recent work (Brooke and Hirst, 2011) suggests that this problem in fact extends even to non-lexical features, leading us to reject traditional within-corpus evaluation (i.e. crossvalidation). In this paper, we explore a novel approach to L1 identification which relies only on externally-derived lexical information. It involves deriving metrics from large weblog corpora for four L1s (Chinese, Japanese, Spanish, French), lessening our reliance on scarce learner corpora. More specifically, we use the average ratios of (translated) word counts in different languages as indicators of interlanguage. If we see the unlikely English bigram *take coffee* in a learner text, our classification of that text will then depend on whether there are patterns of language in some L1 that could be the source of this L2 feature: among French, Spanish, Chinese, or Japanese, is there one language where we see a word that means *take* together with a word that means *coffee*? We show that this method is superior to L2-only cross-corpus classification results using standard features.

[1] Not to be confused with the idea of *interlingua* in machine translation.

2. Related Work

Early native language detection includes that of Koppel et al. (2005). They classified texts from the International Corpus of Learner English (ICLE) into one of five (European) native language backgrounds using support vector machines (SVMs). They described their feature set as stylistic; features included the frequency of function words, rare POS bigrams, letter *n*-grams, and spelling errors. They reported a performance of just over 80% on the task using the full feature set.

Other work on the ICLE includes that of Tsur and Rappoport (2007), who are concerned with identifying phonological language transfer; they focus on the construction of character *n*-gram models, reporting 66% accuracy with just these sub-word features, with only a small drop in performance when the dominant topic words in each subcorpus (as identified using *tf-idf*) are removed. Wong and Dras (2009) investigated particular types of syntactic error: subject-verb disagreement, noun-number disagreement, and determiner problems, relating the appearance of these errors to the features of relevant L1s. However, they reported that these features do not help with classification, and they also note that character *n*-grams, though effective on their own, are not particularly useful in combination with other features. In their most recent work, Wong and Dras (2011) test the usefulness of syntactic production rules and other features derived from parse trees. Only the

former are effective relative to previous feature sets; using just production rules, however, results in a 30% error reduction, or 80% performance on a 7-langauge task. In contrast with other work, Wong and Dras (2011) use binary rather than frequency features and a maximum entropy (MaxEnt) classifer rather than an SVM.

The work of Kochmar (2011) is distinct from those above in a number of ways: she uses a different corpus of essays, derived from the Cambridge Learner Corpus[2], and concentrates on pairwise (SVM) classification within two European language sub-families. An exhaustive feature analysis indicates that character *n*-gram frequency is the most useful feature type for her task; unlike Wong and Dras (2011), syntactic production rules provided little benefit. With respect to lexical features, she presents some results using word *n*-grams, but regards them as attributable to topic bias in the corpus. Error-type features (e.g. spelling, missing determiner) as provided by the corpus annotation offered little improvement over the high performance offered by the distributional features (e.g. POS/character *n*-grams).

Golcher and Reznicek (2011) use a string distance metric to identify the native language of German learners in the Falko corpus (Lüdeling et al., 2008), and contrast this with a topic classification task in the same corpus. Even after taking steps to mitigate topic bias (removing the influence of the words in the title), the usefulness of the three feature types that they investigate (word token, word lemma, and POS) is remarkably similar across the two tasks, with the word features dominating in both cases. Surprisingly, the effect of POS is higher in topic classification than it was on L1-classification. Our recent work (Brooke and Hirst, 2011) also tests the confounding effect of topic in the context of native language identification. In order to motivate the use of new corpora for future research, we segregate a portion of the ICLE by topic and shows that the core set of commonly used features for L1-identification all show significant drops in performance when topic-segregated 2-fold cross-validation is compared to standard (randomized) 2-fold cross-validation. This is particularly true of character *n*-grams, which actually dropped more drastically (32%) than word *n*-grams (14%).

Finally, we note that native language identification has also been included as an element of larger author profiling studies (Estival et al., 2007; Garera and Yarowsky, 2009). A closely related task is the identification of translated texts and/or their language of origin (Baroni and Bernardini, 2006; van Halteren, 2008; Koppel and Ordan, 2011), though the tasks are distinct because the learners included in native language identification studies are usually at a level of linguistic proficiency below that of a professional translator (who in any case may be writing in his or her L1, rather than an L2), and are not operating under the requirement of faithfulness to some original text.

3. Method

The core of our method is the derivation of L1-transfer metrics. Given an L2 text, we derive an L1-transfer metric by an averaging, across all relevant elements of a given type in the text, the ratio of the *potential prevalence* in contrasting L1 corpora (our training corpus). We will use the term *potential prevalence* to refer to counts that are filtered through some mapping; we cannot directly count L2 elements in L1 corpora, but we can count patterns that might produce them. More formally, let $L_1,\ldots,L_p$ be the set of native languages we are interested in identifying, with corresponding corpora $C_1,\ldots,C_p$, and a small finite set of general feature types $T_1,\ldots,T_q$. As we will discuss in more detail later, our feature types include unigrams and bigrams. Our initial set of L1-ratios is then of size $p\times(p-1)$ i.e. one for each feature type for each pair of non-identical languages. For the moment, we assume a function P that provides a *potential prevalence value* for any given textual element e_{ij}, of type T_i, in some L1 corpus C_k, i.e. $P(e_{ij},C_k)\rightarrow\mathbb{N}$. For a given element e_{ij}, we calculate its potential prevalence ratio $R_{lm}(e_{ij})$ for languages L_l, L_m as

$$R_{lm}(e_{ij}) = \log\frac{P(e_{ij},C_l)}{P(e_{ij},C_m)}$$

Note that the use of logarithms ensures that the two potential prevalence ratios derived from any languages are symmetric, $R_{lm}(e_{ij}) = -R_{ml}(e_{ij})$. Next, for all elements of type T_i in a given source text (the set E_i), we calculate the value of a feature $f_{lmi}\in F$ (corresponding to L_l,L_m,T_i) as the average of all the prevalence ratios for all relevant elements in the text:

$$f_{lmi}(E_i) = \frac{\sum_{e_{ij}\in E_i} R_{lm}(e_{ij})}{|E_i|}$$

Then, we define our set of L1-influence metrics V based on a combination of these basic features by language. A particular L1-influence metric v_{li}, l and i as above, is given by:

$$v_{li}(E_i) = \sum_{m=1}^{p} f_{lmi}(E_i)$$

Intuitively, each basic ratio in F provides an indication of whether a text is patterning more like one of two languages, while the set of L1-influence metrics V provides an indication of how much a text is patterning like a particular language in contrast with all other languages. Finally, we normalize these metrics in the context of the test corpus, so they all have the same standard deviation. For some text with textual elements E_i:

$$v'_{li}(E_i) = \frac{v_{li}(E_i) - \overline{v_{li}}}{\sigma_{v_{li}}}$$

A text is classified as the language L_c with the highest normalized influence metric, i.e.

$$c = \arg\max_l v'_{li}$$

The above provides an abstract basis for our classification using L1-influence metrics. However, we need to define the potential prevalence function, which depends directly on the type of feature T being extracted. Our main feature is what we call *boundary bigrams* (or just bigrams),

[2] http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus

which correspond to the L2 (translated) bigram associated with two consecutive words in an L1 corpus. Let us consider some L1 corpus C, with tokens $w_1 \dots w_n$, each of which has some (possibly empty) set of translations $t_i = t_{i1}, \dots, t_{ij}$, with each t consisting of one or more words in the target (L2) language, say $t_{ij1} \dots t_{ijm}$. Then the potential prevalence function for the boundary bigram feature T_1 for an element e_{i1} corresponding to a ordered pair of target words ($t't''$) is the count, across all adjacent words $w_i \dots w_{i+1} \in C$ and across all their potential translations $t_{i1}, \dots, t_{ij}, \dots, t_{im}$, $t_{(i+1)1}, \dots, t_{(i+1)k}, \dots, t_{(i+1)l}$, of the number of instances where, $t' = t_{ijq}$ and $t'' = t_{(i+1)k1}$, given $|t_{ij}| = q$. That is, a count of all the instances where the last word of one of the translations of some $w_i \in C$ is equal to the first word of the bigram, and the first word of one of the translations of $w_{i+1} \in C$ is equal to the second word of the bigram.[3] For instance, consider the French phrase *prends un café*, which a (partial) list of translations for each word as below:

w_i	*prends*	*un*	*café*
t_{i1}	take	a	coffee
t_{i2}	hold	an	java
t_{i3}	go by	one	cafe

An appearance of this phrase in a corpus would generate a boundary bigram count for *take-a*, *a-coffee*, *take-an*,..., *by-a*, *by-an*,..., *one-cafe*. They are boundary bigrams because we only consider the bigrams that straddle word boundaries (not *go-by*, for instance); assuming a reliable bilingual lexicon, within-word bigrams (when they occur) will involve only correct usage of the L2, but we intend boundary bigrams to find lexical patterns that reflect transfer from the L1.

A related way of using L1 corpora is to derive information via the use of *k-window collocational pairs*. These k-window collocational pairs differ from boundary bigrams in three key ways: first, they do not require strict adjacency, which is to say that for an integer k, w_i, $w_j \in C$ are considered k-window collocations if $|i - j| \leq k$. Second, we consider only those translations of length 1, i.e. only t_{ij} s.t. $|t_{ij}| = 1$. Third, collocational pairs are unordered, i.e. the sequence $w'w''$ will result in the same collocation counts as the sequence $w''w'$. Otherwise the potential prevalence function for collocational pairs is similar to boundary bigrams, a count of target language word pairs over all the words and all the translations of these words in the L1 corpus. In our *prends un café* example, the 2-window collocational pairs include all combinations of all the single word translations, e.g. (*coffee*, *take*), (*hold*, *java*), but not anything with *by* or *go* since these are part of a multiword translation. Here, we only test 2-window collocational pairs.

For the unigram feature type, we simply count all target words (t_{ijk}) in all translations for all tokens in the corpus. For POS unigrams, bigrams, and trigrams, each word w_i is given a corresponding POS tag p_i, and we count sequences of these POS tags, and then map them to a single, coarse-grained tag set consisting of nouns, verbs, adjectives, adverbs, conjunctions, pre/postpositions, pronouns, numbers, punctuation, and the catch-all category of (other) function words, so that these counts can be compared across L1s. For combined unigram and bigram counts, we sum the potential prevalence ratios derived for each feature.

Not all elements of the text are equally useful for L1 classification. We posited that classification would be better if commonly occurring features of English were filtered, since these may vary randomly across L1 and produce noise. We implement this by fixing a maximum n-gram count, as derived from an independent corpus, for the elements used to calculate the L1-influence metrics. Appropriate thresholds were selected by optimizing in the held-out development set. We exclude proper nouns, which can of course be useful for L1-identification but should not be attributed to language transfer, which is our main interest here.

4. Data and Resources

The data and resources used in this work can be divided into four categories: the (L1) corpora for deriving potential prevalence, resources for analysis of these corpora (e.g. segmenters, taggers), bilingual lexicons, and evaluation resources. For the L1 data, we choose to draw primarily from a single web corpus, the ICWSM Spinn3r dataset (Burton et al., 2009), which, although primarily an English corpus, also contains a large number of blog posts in other languages. There is a great deal of variation in the amount of data available for each language; for consistency, we choose a fixed length sample (100 million tokens, after segmentation) for each of the five languages. Chinese, however, was underrepresented with only 19 million tokens, and so we extracted additional blogs from a popular Chinese site.[4] In additional to ICWSM English data, we used the Google 1T 5-gram Corpus (Brants and Franz, 2006), which includes counts based on one trillion tokens from the web, for our count thresholds.[5]

For the European languages it was possible to use simple heuristics for tokenization, while for Chinese we needed a special segmenter: we employed the Stanford Chinese segmenter (Chang et al., 2008a), in the Chinese Treebank tagset mode. For Japanese, the MeCab morphological analyzer[6] served as our segmenter as well as part-of-speech tagger. For the other languages, POS tagging was carried out using the Tree Tagger (Schmid, 1995) and the associated parameter files for each language.

We did not have immediate access to sufficiently large machine-readable bilingual dictionaries for any of the (non-English) L1s, so we took advantage of the various websites which offer free online bilingual translations. Over the course of several months, we slowly and politely queried these websites for English translations of words that ap-

[3] We also tested using pointwise mutual information as a potential prevalence indicator for boundary bigrams, but it was not as effective as raw counts. More generally, a probabilistic interpretation of potential prevalence assigns far too much probability mass to nonsense bigrams we will never see in actual texts.

[4] http://www.sina.com

[5] We summed relevant trigram counts to get our thresholds for the 2-window collocations.

[6] http://mecab.sourceforge.net/

Table 1: Native language classification results

Configuration	Accuracy (%)					
	ICLE texts		Lang-8 texts		FCE texts	
	No Filter	**w/Filter**	**No Filter**	**w/Filter**	**No Filter**	**w/Filter**
Guessing baseline	25.0	25.0	25.0	25.0	25.0	25.0
Unigrams	43.5	44.6	26.0	26.9	22.0	22.5
Bigrams	42.9	**48.3**	**36.4**	**39.2**	28.5	29.0
2-window collocations	32.1	46.9	31.9	38.3	29.5	**32.0**
POS unigrams	25.0	30.1	26.4	30.0	**32.5**	26.0
POS bigrams	17.0	25.3	26.9	29.1	25.5	27.0
POS trigrams	16.5	28.8	27.8	28.4	23.0	26.5
Unigram + Bigrams	**44.2**	46.2	27.9	30.1	24.3	23.7

peared often (at least 5 times)[7] in the corresponding subcorpus. For Chinese, we used *iciba.com*, for French *larousse.fr*, for Spanish *spanishdict.com*, and for Japanese *jisho.org*; our choice of websites was based on dictionary quality, ease of extraction, and, in particular for the European languages, the ability to deal with inflected forms, i.e. to find their corresponding lemma without need for additional lemmatization on our part. Although we attempted to keep the size of the dictionaries comparable, in terms of lemmas the Chinese and Japanese lexicons are markedly larger than the French and Spanish ones;[8] if inflected forms are considered, however, the European-language lexicons are larger.

For all languages, we ignored translations longer than three English words, as we found that many of these were explanations rather than translations. Some very common words in some lexicons had only explanatory entries; for these (fewer than 10 in each lexicon) we manually inserted a direct translation based on examples or, in the case of certain particles, left them with an empty translation. The translations of verbs and nouns were generally in base form, which would have resulted in only uninflected bigrams; instead, we used the part-of-speech tagging to create simple correspondences between forms in the L1 and inflected forms in English. For instance, plurals in French are translated into plural forms in English; for Chinese, however, which does not mark number on most nouns, both English forms are included as potential translations. All of the dictionaries categorized their translations by part of speech, and in general we used the translations for only the part of speech as given by the tagger, though all translations were used if that strategy failed.

Our first evaluation corpus is the International Corpus of Learner English (ICLE), version 2 (Granger et al., 2009), which has English-learner essays (primarily argumentative) for 16 languages. For each of the 4 languages investigated here, we used the first 50 texts in each subcorpus for development, and the next 200 for testing. Our second evaluation corpus is our new Lang-8 corpus (Brooke and Hirst, 2011), which consists of 154,702 journal entries (mostly from Asian languages) from the Lang-8 learner website. Since the average entry in the Lang-8 is significantly shorter than those in the ICLE (about 150 tokens), we concatenate multiple entries together to form our 'texts' of roughly the same length as those in the ICLE, also 200 for each language.[9] Our third corpus is a small sample of texts from the Cambridge Learner Corpus that has recently been made available (Yannakoudakis et al., 2011); these texts consist of short answers from the First Certificate in English (FCE) exams. The set we use here consists of only 50 texts per language, and the average length of the texts is roughly half of the other two corpora; thus we expect classification to be harder.

5. Evaluation

Table 1 contains the L1 classification results for the various feature types and evaluation corpora. The boundary bigram and k-window collocations are obviously the most useful feature types; their performance is consistently well above chance, even without filtering. By comparison, the POS features do not appear to transfer properly and perform often near or even below chance, perhaps because the sequences in which POSs appear are just simply too language dependent. The effectiveness of unigram features vary widely: in the ICLE, they are roughly as good as bigrams, but in the FCE they are worse than guessing. We suspect that these variations may reflect a fundamental difference in the nature of the two corpora: the short answers in the FCE are constrained to a very restricted topic and genre—letters expressing gratitude at winning a prize—which may limit the extent to which vocabulary choice can distinguish among L1s. It is therefore the choice of which words are put together that is particularly telling, reflecting transfer from the L1.

One very clear result is the effect of filtering: in nearly every case, filtering out elements that were common in English improved classification accuracy. This effect is most

[7] All of our query-derived lexicons in fact may have more than just those words appearing 5 times in the corpus, but this is the last cutoff point that all dictionaries reached. We do not, however, believe there is much benefit to be gained from further extraction, since such rare words rarely have definitions in the online dictionaries.

[8] Chinese: 109,061, Japanese: 85,867, Spanish: 26,627, French: 26,495.

[9] Although this may appear to be a fairly small sample of this corpus, in fact the 200 texts nearly exhausts the data available for the two European languages.

Table 2: Confusion matrix for best ICLE result

Native	**Classified as**			
Language	Chinese	Japanese	French	Spanish
Chinese	103	42	27	28
Japanese	41	111	18	30
French	15	30	86	69
Spanish	13	33	68	86

pronounced in the ICLE (from which we also took our development set), but it is visible in the other two corpora as well. The bigram threshold was 10^6 appearances in a corpus with roughly 10^{12} bigram tokens. One negative result is that the features do not appear to combine well; in general, summing unigram and bigram metrics did not improve performance. Table 2 contains the confusion matrix for the bigram L1-influence metric in the ICLE, our best result. The Asian languages are the easiest to distinguish, while the two closely related European languages are distinct from the Asian but often misclassified as each other. This is exactly what we should expect given our knowledge about how the languages are related to each other. We suspect performance would be much higher if we had not included languages that are so closely related to each other as well as English (that is, French and Spanish), though even these two languages are distinguished better then chance.

We also looked at the individual bigrams that contributed to the metrics, in particular those with very high or low potential prevalence ratios. Among the most telling features for Chinese, we noticed a number of Chinese-influenced adjective-noun collocations (e.g. *main income*, *medium industry*), but there were also syntactic errors of number (e.g. *they depends*). The patterns were less clear for European languages like French, though we noted certain verb-preposition combinations (e.g. *tolerated to*, *witnessing in*) that seemed to be cases of language transfer. There was also a great deal of noise, which might be eliminated by further filtering, for instance focusing only on specific POS patterns.

6. Conclusion

In this paper, we have presented a method for using native language corpora as a source of information for native language identification in non-native texts. In particular, our approach relies on the phenomenon of language transfer, where patterns of the L1 intrude into the L2. The results offered here are well above chance, though they are not good enough for us to conclude that this method alone is sufficient. However, there are aspects of our method that make it distinct from traditional machine-learning approaches: in particular, our metric can provide a small set of features that may represent an huge number of rare (but telling) events that might otherwise be filtered out by feature selection. Our method also offers an explicit connection between L2 forms and the L1 forms that created them; this information could be used to improve automated error correction.

7. References

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21:259–274.

Thorsten Brants and Alex Franz. 2006. *Web 1T 5-gram Corpus Version 1.1*. Google Inc.

Julian Brooke and Graeme Hirst. 2011. Native language detection with 'cheap' learner corpora. Presented at the 2011 Conference of Learner Corpus Research (LCR2011).

Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA.

Pi-Chuan Chang, Michel Gally, and Christopher Manning. 2008a. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the ACL '08 Third Workshop on Statistical Machine Translation*.

Yu-Chia Chang, Jason S. Chang, Hao-Jan Chen, and Hsien-Chin Liou. 2008b. An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3):283–299.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Correcting semantic collocation errors with L1-induced paraphrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 107–117.

Dominique Estival, Tanja Gaustad, Son B. Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING '07)*, pages 263–272.

Ian Fette, Norman Sadeh, and Anthony Tomasic. 2007. Learning to detect phishing emails. In *Proceedings of the 16th International World Wide Web Conference (WWW '07)*, pages 649–656.

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP '09)*, pages 710–718.

Felix Golcher and Marc Reznicek. 2011. Stylometry and the interplay of title and L1 in the different annotation layers in the Falko corpus. In *Proceedings of Quantitative Investigations in Theoretical Linguistics 4*.

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

Ekaterina Kochmar. 2011. Identification of a writer's native language by error analysis. Master's thesis, University of Cambridge.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text

for errors. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*, pages 624–628.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool.

Anke Lüdeling, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt, and Maik Walter. 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, 42:67–73.

Terence Odlin. 1989. *Language Transfer*. University of Cambridge Press.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*, pages 47–50.

Larry Selinker. 1992. *Rediscovering Interlanguage*. Longman, London.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition (CACLA '07)*, pages 9–16.

Hans van Halteren. 2008. Source language markers in EUROPARL translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08)*, pages 937–944.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 53–61.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 1600–1610.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 180–189.