From Language to Family and Back: Native Language and Language Family Identification from English Text

Ariel Stolerman **Aylin Caliskan Islam** **Rachel Greenstadt**
Dept. of Computer Science
Drexel University
Philadelphia, PA
{ams573,ac993,greenie}@cs.drexel.edu

Abstract

Revealing an anonymous author's traits from text is a well-researched area. In this paper we aim to identify the native language and language family of a non-native English author, given his/her English writings. We extract features from the text based on prior work, and extend or modify it to construct different feature sets, and use support vector machines for classification. We show that native language identification accuracy can be improved by up to 6.43% for a 9-class task, depending on the feature set, by introducing a novel method to incorporate language family information. In addition we show that introducing grammar-based features improves accuracy of both native language and language family identification.

1 Introduction

Mining text for features to infer characteristics on its author is an important research field. One author property that has been researched is native language, extracted from the author's writing in a non-native language. Learning the native language of an anonymous author can assist in profiling criminals or terrorists, and may also undermine the privacy of legitimate anonymous authors by helping to unveil their identity.

Influences of native language (L1) on second language (L2), referred as the L1-L2 transfer effect, is seen in writing and can be utilized to identify native language. In this paper we examine aspects of a broader class – the language family to which the native language of an author belongs. In the rest of the paper native language and native language family will be referred as L1 and LF, respectively.

First, we examine the correct classification rates of LF compared to L1. As L1 is a subset of LF, the number of L1 classes is greater than or equal to the number of corresponding LF classes. Therefore, higher LF classification accuracy can be achieved trivially by taking the family of the attributed L1 in a L1 classification task. This can be helpful in cases where high accuracy is preferred over resolution. We introduce a novel, improved method that achieves higher correct classification rate for LF identification, compared to the trivial method.

Our main contribution is showing that L1 identification accuracy can be increased by incorporating family information via LF identification.

We use stylometric analysis and machine learning techniques to identify L1 and LF. We conduct a series of experiments by mining English text written by non-native English authors for linguistic features. We use 4 different feature sets detailed in section 3. We evaluate the accuracy of our results by examining the true-positive rate.

The novelty of our work is in exploring the LF-L2 transfer effect using stylometric methods, and expanding L1 identification methods accordingly. Increasing the state-of-the-art correct classification rate for L1 detection is not our main goal. Instead, we introduce concepts to increase achieved accuracy by incorporating LF knowledge into the classification process.

The next section (2) provides background and prior work. Section 3 describes the experimental setup. In section 4 we describe the different experiments that were performed, followed by results and evaluation. We finalize with discussion on the given results (section 5), followed by conclusions and directions for future research (section 6).

2 Related Work

Literature includes work on extracting demographic and psychological traits from different data formats, such as speech and text samples. Native language and accent identification from speech can be found in (Choueiter et al., 2008; Tomokiyo and Jones, 2001). Identifying an author's native language from L2 text, which is English in most cases, is the closest problem to our work.

Introductory studies in the area identified the written or spoken language itself, focusing on telephone dialogue corpora (Ahmed et al., 2004; Zissman, 1993). Further studies focused on extracting specific information from text or speech after identifying the language being used. Wanneroy et al. (1999) investigated how non-native speech deteriorated language identification and used acoustic adaptation to improve it. Choueiter et al. (2008) classified different foreign accented English speech samples by using a combination of heteroscedastic LDA and maximum mutual information training. Tomokiyo and Jones (2001) characterized part-of-speech sequences and showed that Naïve Bayes classification can be used to identify non-native utterances of English.

The first work that utilized stylometric methods for native language attribution is introduced by Koppel et al. (2005a; 2005b). They explored frequencies of sets of features, and used them with multi-linear support vector machines to classify text by author's native language. They used a set of features consisted of function words, letter n-grams, errors and idiosyncrasies, and experimented on a dataset of authors of five different native languages taken from ICLEv1 (Granger et al., 2002), reaching to 80.2% accuracy. Tsur and Rappoport (2007) revisited Koppel's work using only the 200 most frequent character bigrams, and achieved 65.6% accuracy, with only a small degradation when removing dominant words or function words.

Brooke and Hirst (2012) presented a method of utilizing native language corpora for identifying native language in non-native texts. They used word-by-word translation of large native language corpora to create sets of second language forms that are possible results of language transfer, later used in unsupervised classification. They achieved results above random chance for L1 identification, however insufficiently accurate.

More related work can be found in (Estival et al., 2007; van Halteren, 2008; Carrio-Pastor, 2009; Golcher and Reznicek, 2009; Wong and Dras, 2009; Wong et al., 2011; Brooke and Hirst, 2011; Ahn, 2011). The work mentioned above and our approach both utilize the L1-L2 transfer effect to gain information about an author's native language. Gibbons (2009) proved the impact of native language family's typological properties on L2. As far as we know, our work is the first to combine stylometry and native language family's effect on L2, utilized for L1 identification.

3 Experimental Setting

3.1 Corpus

We use the ICLEv2 (Granger et al., 2009) corpus that contains English documents written by intermediate to advanced international learners of English, with language backgrounds of 16 mother-tongues. The first version of the corpus was used in significant previous work (Koppel et al., 2005a; Koppel et al., 2005b; Tsur and Rappoport, 2007). They reported that they were able to use 258 documents of sizes 500-1000 words for each language they used. We use version 2 of the corpus and restrict all documents in our experiments to those with 500-1000 words as well. However, we found that constraining our documents to these lengths allows us to use only 133-146 documents per language. We conduct a series of experiments with different sub-corpora constructed of documents representing 11 native languages out of the 16 available in the corpus. The native languages we used are: Bulgarian, Czech, Dutch, French, German, Italian, Norwegian, Polish, Russian, Spanish and Swedish, all Indo-European languages. These languages represent 3 language-families in a coarse partition: Germanic, Slavic and Romance, which are used as the LF class in the experiments to follow. All sub-corpora configurations are detailed in section 4.

Since we are looking at a set of languages from both L1 and LF aspects, we maintained only the sub-corpora that allowed a sufficient amount of languages in each represented family, i.e. 3 languages in each of the Germanic, Slavic and Romance families. Therefore we removed 5 of the 16 available languages in the corpus.

3.2 Feature Selection

Koppel et al. represented each document in their experiment as a 1,035-dimensional feature vector: 400 function words, 200 most frequent letter n-grams, 185 misspellings and syntactic errors and 250 rare POS bigrams. The 250 rare POS bigrams are the least common bigrams extracted from the Brown Corpus (Francis and Kucera, 1983), and their appearances are considered to be erroneous or non-standard.

In our experiments we used 4 different feature sets, partially based on that used by Koppel et al. We used the authorship attribution tool JStylo (McDonald et al., 2012) for feature extraction. The feature sets are the following:

Basic: includes the 400 most frequent function words, 200 most frequent letter bigrams, 250 rare POS bigrams and 300 most frequent spelling errors.

The 400 most frequent function words were taken from a list of 512 function words used in the original experiments by Koppel et al. For the 200 letter n-grams, we chose bigrams, as they are shown to be effective for the task in previous research. The 250 rare POS bigrams were extracted from the Brown Corpus using the POS tagger in (Toutanova et al., 2003). Finally, we simplified the error types by considering only misspelled words, based on a list of 5,753 common misspellings, constructed from Wikipedia common misspellings and those used in (Abbasi and Chen, 2008). We ignored any misspellings with 0-1 appearances across the entire sub-corpus. Since many of the rare POS bigrams and misspellings had no appearances, the effective vector lengths vary between 653-870 features.

Extended: identical to the former, with the addition of the 200 most frequent POS bigrams across the entire sub-corpus used for each experiment. These syntactic features were selected as an additional representation of grammatical structures in the text.

There are several methods for natural language classification, including genetic, typological and areal (Campbell and Poser, 2008). We consider the typological classification that uses structural features to compare similarities between languages and classify them into families. Therefore we chose grammatical evidence in L2 as features that may represent similar transfer effects among languages in the same family.

Grammatical: constructed only from the 200 most frequent POS bigrams, representing the grammatical level of the text.

InfoGain: We used the 200 features with the highest information gain extracted from the extended feature set using Weka (Hall et al., 2009), calculated for any given feature by measuring the expected reduction in entropy caused by partitioning the test instances according to that feature.

3.3 Classification

We trained a SMO (Platt, 1998) SVM classifier with polynomial kernel, chosen as SVMs are used extensively in prior work and ours outperformed other methods tested, including decision trees, nearest-neighbors, Bayesian and logistic regression classifiers.

4 Experimental Variations and Evaluation

We conducted 3 different experiments using various sub-corpora and the 4 feature sets described in the previous sections, with L1 and LF classification tasks. We evaluated the results by using the true-positive rate to capture accuracy. Following is a detailed description of the different variations and results.

4.1 9-Class Languages, 3-Class Families

Setup: We compared 9-L1 identification with the corresponding 3-LF identification, using datasets constructed of the sub-corpus containing all 11 languages mentioned before. For the 9-L1 task we randomly sampled documents of 9 languages, 3 for each of the Germanic, Slavic and Romance language families, in order to maintain the same number of languages per family in every experiment. We constructed 16 different 9-L1 sets, choosing 3 out of 4 Germanic languages, 3 out of 4 Slavic languages and the only 3 Romance languages available. In each of the 16 experiments we used the same number of documents per language, varying between 133-146.

In order to compare results with LF identification, we conducted 3 sets of experiments, each containing 16 3-LF experiments, corresponding to the 16 that were performed for L1 identification.

First, we ran the trivial experiment of attributing the family of the predicted language resulted from the L1 identification experiments. This method is denoted as the *trivial* method.

Next, we ran the same experiments conducted for L1, with the only difference of using LF as the class rather than L1. As a result of that configuration, each experiment also contained the same number of documents per language family, varying between 399-438. This method is denoted as the *standalone* method (as it is a standalone experiment, independent of L1 classification results).

Lastly, we ran experiments combining the standalone and trivial approaches. We hypothesize that if L1 is attributed with high confidence, so is the LF of that attributed L1, however if the confidence level decreases, a standalone LF experiment achieves better results. We ran the L1 identification experiments and set a threshold as the averaged probability of the predicted class across the entire test set, based on the class probability distribution outputted by the SVM classifier. To obtain proper probability estimates, we fit logistic regression models to the outputs of the SVM. Every instance classified with probability above the threshold was attributed the family using the trivial method, and every instance below – using the standalone method. This method is denoted as the *combined* method.

Results: We averaged the results of all 16 L1 identification experiments, and those of the 3 sets of 16 LF identification experiments. See figure 1.

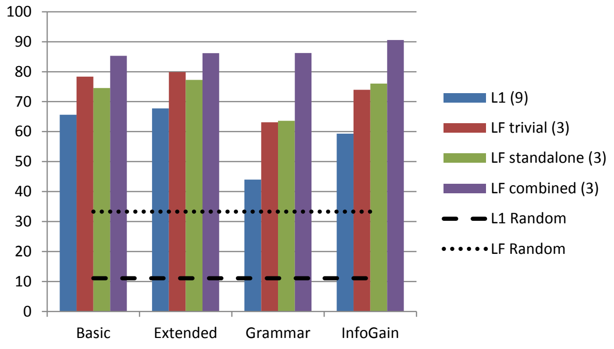

Figure 1: Accuracy for 9-class L1 and 3-class LF identification. The combined method for LF outperforms the other two.

The accuracy for L1 identification was 67.78%, 65.64%, 59.34% and 44.02% for the extended, basic, InfoGain and grammatical feature sets, respectively.

Out of the 3 LF identification experiment sets, the combined method achieved the best accuracy: 90.57%, 86.24%, 86.2% and 85.29% for the InfoGain, grammatical, extended and basic feature sets, respectively. These results support our hypothesis.

The trivial method achieved better results than the standalone method for the basic and extended feature sets: 78.33% and 79.87% for the first, 74.53% and 77.24% for the latter. For the grammatical and InfoGain feature sets, the standalone performed better than the trivial: 63.61% and 76.02% for the first, 63.1% and 73.94% for the latter.

Since the L1 identification experiments have more classes than the LF experiments, the random chance varies between them: 11.11% for L1 and 33.33% for LF. Although the absolute accuracy for LF is consistently higher than for L1, if we subtract the corresponding random chance values to obtain *"effective"* accuracy, in most cases L1 is more accurate than LF. The LF combined method is the only one out of the 3 LF methods that exceeds the effective accuracy of L1, for the grammatical and InfoGain feature sets. Combined with the standard (non-effective) results, it appears that the InfoGain feature set with the LF combined method achieves the highest accuracy with the most added knowledge over random classification, across all tasks and feature sets. It is also notable that the smallest difference between L1 and LF identification accuracy is seen for the grammatical feature set. See figure 2.

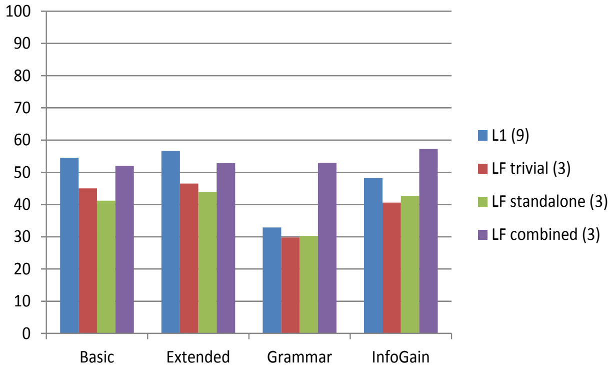

Figure 2: Effective accuracy for 9-L1 and 3-LF identification. Accuracy for L1 exceeds most accuracy results for LF, except for the combined method on the grammatical and InfoGain feature sets.

4.2 3-Class Languages, 3-Class Families

Setup: In order to have the same random-chance baseline for both L1 and LF tasks, we compared 3-L1 with 3-LF identification, using the same subcorpus as before.

For L1 we constructed 9 experiments, in each randomly sampling 3 languages from 1, 2 and 3 different language families (3 experiments each). The reason for this choice is that as more families are used, the farther the chosen languages are from one another. Therefore the choice above is intended to balance the effect of LF in those experiments. We used 133 documents per language for all experiments.

For LF we constructed 2 sets of 9 experiments, in order to examine the notion that languages in the same family have more family-distinguishable commonalities as opposed to random sets of languages. In the first, for each of the experiments we randomly created 3 sets of languages to be considered as families. We randomly sampled documents from all 11 languages to construct sets for the 3 randomly-generated families used as classes. Here we also maintained 133 documents per language family. In the second we ran a similar configuration, only using the actual language families.

Results: The averaged accuracy for L1 was 84.23%, 82.29%, 81.67% and 66.97% for the extended, InfoGain, basic and grammatical feature sets, respectively. These results consistently outperformed the results of both sets of LF experiments. See figure 3.

The accuracy attained for actual language families

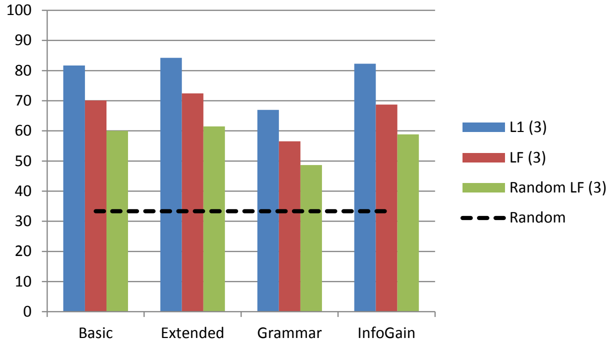

Figure 3: Accuracy for 3-L1, 3-LF and 3-randomly-generated families identification. Using the original families achieves the highest accuracy for LF identification.

was 72.43%, 70.09%, 68.72% and 56.55% for the extended, basic, InfoGain and grammatical feature sets, respectively, which consistently outperformed that of the randomly-generated families: 61.46%, 60.01%, 58.81% and 48.67%. This shows that partitioning the languages into sets by their actual family achieves the highest accuracy for LF identification. As in the previous experiment, the difference in accuracy between L1 and LF identification was the smallest with the grammatical feature set.

4.3 9-Class Languages, Reclassify by Family

Setup: We wanted to examine whether LF classification can improve L1 classification. In this experiment we conducted the same 16 9-L1 experiments from section 4.1. We then set a threshold as in the *combined* method in section 4.1, such that each classified instance with predicted probability less than that threshold is treated as misclassified. For all allegedly-misclassified instances we attributed the family they belong to, using various methods detailed later. As last step we reclassified those instances using a training set constructed only of the 3 languages in the family they were classified as, and considered these results as L1 classification-correction for those instances. We measured the overall change in accuracy.

The entire 16 10-fold cross-validation experiments were conducted 3 times, each with a different method for LF attribution for the instances below the threshold: 1) The standalone method – running LF identification task over all those instances, using the same training set (with families as classes rather than languages), 2) The trivial method – using the family of the predicted language of those instances, and 3) Random – randomly selecting the family.

Results: We averaged the results of all 16 L1 experiments for each of the 3 LF attribution methods and each of the 4 feature sets used.

We measured the net fix in accuracy (added number of correctly classified instances, taking into account corrected classifications and new misclassifications). For all feature sets, LF attribution using the standalone method yielded the highest fix rate, followed by LF attribution using the trivial method. The randomly attributed family method consistently yielded negative fix rate (i.e. reduced overall accuracy). See figure 4.

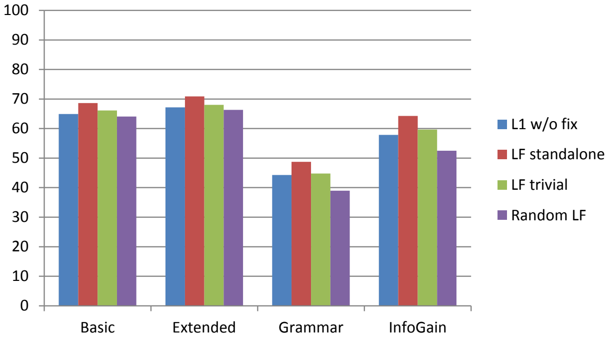

Figure 4: Accuracy for L1 identification without fix and with fixing using LF attribution by the standalone method, trivial method and random selection of family. The standalone method yields the highest net fix in L1 classification accuracy.

The extended feature set yielded the best results. Starting at a baseline of 67.17% for L1 identification without any fix, the true-positive rates obtained for this feature set were 70.9% and 68.05% for attributing LF by the standalone and the trivial methods, respectively. The increase in accuracy is statistically significant ($p < 0.01$). The random family attribution method yielded a decrease in accuracy to 66.35%.

It is notable that although yielding best results for the extended feature set, the standalone method achieved higher increase in accuracy in some of the other feature sets. The increase rates for this method were: 6.43%, 4.48%, 3.73% and 3.67% for the InfoGain, grammatical, extended and basic feature sets, respectively.

5 Discussion

The first notable result is seen in experiment 4.1, where using the combined method for LF identification derives higher accuracy than both the trivial and the standalone methods. This may suggest that when L1 is predicted with high confidence, LF is predicted well, but when the confidence level is low, it is better to run standalone LF classification. Since the combined method uses the best of the two others, it outperforms both.

The most important result is seen in experiment 4.3, where L1 identification is improved by up to 6.43% in accuracy for 9-L1 classification by introducing information about the language family, thus providing a smaller set of language classes in which the actual language is more likely to be found. Attributing LF by standalone experiments yielded higher L1 classification accuracy than attributing it by the family of the predicted language. This outcome seemingly contradicts the results seen in section 4.1, where the latter LF attribution method outperformed the first. However, this only supports the idea suggested above regarding the threshold, that the family of the attributed L1 is the actual family with higher probability than LF attributed by a standalone experiment, only when L1 is attributed with high confidence (i.e. above the selected threshold).

The results in sections 4.1 and 4.2 suggest that all 4 feature sets achieve better accuracy for L1 than for LF (standalone) classification. We believe this is since for L1 we try to distinguish between individual languages as they transfer to English. However, LF identification necessitates finding features that intersect between languages in a particular family, and distinguish well between different families as they are transferred to English. This makes LF identification a more difficult task.

The results obtained for randomly generated families in sections 4.2 and 4.3, which are consistently lower than using the actual families, suggest that the contribution of using the latter yields the best performance. That is, languages in the same family have more commonalities distinguishing them from other families, than random sets of languages have.

Looking at the results using the different feature sets, in most cases the extended feature set outperformed the rest. This shows that adding grammatical features increases accuracy for both L1 and LF. Furthermore, in all experiments using *only* the grammatical features achieved a rather good accuracy (significantly higher than random chance), considering that we used only 200 of these features. This supports the notion that grammatical features are useful for both L1 and LF identification.

Another interesting notion regarding the grammatical feature set is seen in the portion these features consist of the InfoGain feature set for the experiments of section 4.2: 33.05% for L1 and 57.16% for LF. This suggests that the grammatical level of the text has greater significance for identifying LF compared to L1. When analyzing the portion lexical features consist of the InfoGain feature set, an opposite trend is seen: function words and letter bigrams consist 29.94% and 33.94% of the features for L1, as opposed to 17.44% and 23.55% for LF, respectively. This suggests that the lexical level of the text is better for L1 detection than for LF detection. Although less significant, the same trend is seen with spelling errors: 3% for L1 and 1.83% for LF.

6 Conclusion

The main conclusion is that when trying to gain information about the native language of an English text author, integrating family identification can increase the total accuracy, using the method introduced in section 4.3, where all low-confidence classifications are reapplied within a smaller set of candidates – languages within the family attributed to those instances using a standalone experiment.

Furthermore, when dealing with a large number of L1 classes, higher accuracy can be attained by reducing the level of specification to language families, which can be obtained with high accuracy using the combined method presented in this paper that integrates both the trivial LF by predicted L1 and LF by standalone experiment methods using the average confidence level as threshold.

In addition, using the most frequent POS bigrams, which represent the grammatical level of the text, is shown to increase accuracy in both L1 and LF identification tasks, especially for the latter. Using lexical features as function words and character bigrams is helpful especially for L1 identification.

We suggest several directions for future work. First, trying new feature sets that may capture other similarities between languages in the same family. For instance, since languages in the same family tend to share basic vocabulary, it may have some level of transfer to L2 that could be captured by a synonym-based classifier. For instance, “verde” in Spanish and “vert” in French may be translated to “verdant”, whereas “grün” in German and “groen” in Dutch may be translated to “green”.

In addition, we can further explore the notion of increasing accuracy by applying knowledge of a broader class on the task applied in other stylometry-based information extraction tasks. For instance, using wide age ranges as the broader class for classifying age of anonymous authors, or personality prototypes for personality type identification.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2):7:1–7:29, April.

Bashir Ahmed, Sung-Hyuk Cha, and Charles Tappert. 2004. Language identification from text using n-gram based cumulative frequency addition. Proc. CSIS Research Day, May.

Charles S. Ahn. 2011. Automatically detecting authors' native language. Thesis, Naval Postgraduate School, March.

Julian Brooke and Graeme Hirst. 2011. Native language detection with 'cheap' learner corpora. In *The 2011 Conference of Learner Corpus Research (LCR2011)*.

Julian Brooke and Graeme Hirst. 2012. Measuring interlanguage: Native language identification with l1-influence metrics. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Lyle Campbell and William J. Poser. 2008. *Language Classification: History and Method*. Cambridge University Press.

Maria Luisa Carrio-Pastor. 2009. Contrasting specific english corpora: Language variation. *International Journal of English Studies, Special Issue*, pages 221–233.

Ghinwa F. Choueiter, Geoffrey Zweig, and Patrick Nguyen. 2008. An empirical study of automatic accent classification. In *ICASSP*, pages 4265–4268.

Dominique Estival, Tanja Gaustad, Son B. Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for english emails. In *10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*, pages 262–272.

Winthrop Nelson Francis and Henry Kucera. 1983. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin.

Erin Elizabeth Gibbons. 2009. The effects of second language experience on typologically similar and dissimilar third language. Thesis, Brigham Young University, Center for Language Studies.

Felix Golcher and Marc Reznicek. 2009. Stylometry and the interplay of topic and l1 in the different annotation layers in the falko corpus. In *Humboldt-Universitat zu Berlin*, QITL-4. [Online: Stand 2012-03-22T16:09:09Z].

Sylvaine Granger, Estelle Dagneaux, and Fanny Meunier. 2002. *International Corpus of Learner English : Version 1 ; Handbook and CD-ROM*. Pr. Univ. de Louvain, Louvain-la-Neuve.

Sylvaine Granger, Estelle Dagneaux, Magali Paquot, and Fanny Meunier. 2009. *The International Corpus of Learner English, Version 2: Handbook and CD-Rom*. Pr. Univ. de Louvain, Louvain-la-Neuve.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005a. Automatically determining an anonymous author's native language. In *Proceedings of the 2005 IEEE international conference on Intelligence and Security Informatics*, ISI'05, pages 209–217, Berlin, Heidelberg. Springer-Verlag.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005b. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 624–628, New York, NY, USA. ACM.

Andrew McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. 2012. Use fewer instances of the letter "i": Toward writing style anonymization. July.

J. Platt. 1998. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Laura Mayfield Tomokiyo and Rosie Jones. 2001. You're not from 'round here, are you?: naive bayes detection of non-native utterance text. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Human Language Technology Conference (HLT-NAACL 2003)*.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hans van Halteren. 2008. Source language markers in europarl translations. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 937–944, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. Wanneroy, E. Bilinski, C. Barras, M. Adda-Decker, and E. Geoffrois. 1999. Acoustic-phonetic modeling of non-native speech for language identification. In *Proceedings of the ESCA-NATO Workshop on Multi-Lingual Interoperability in Speech Technology (MIST)*, The Netherlands.

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

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2011. Topic modeling for native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 115–124, Canberra, Australia, December.

Marc A. Zissman. 1993. Automatic language identification using gaussian mixture and hidden markov models. In *Proceedings of the 1993 IEEE international conference on Acoustics, speech, and signal processing: speech processing - Volume II*, ICASSP'93, pages 399–402, Washington, DC, USA. IEEE Computer Society.