Native Language Identification Using Recurring N-grams – Investigating Abstraction and Domain Dependence

Serhiy BYKH *Detmar MEURERS*
Seminar für Sprachwissenschaft, Universität Tübingen
{sbykh,dm}@sfs.uni-tuebingen.de

ABSTRACT
Native Language Identification tackles the problem of determining the native language of an author based on a text the author has written in a second language. In this paper, we discuss the systematic use of recurring n-grams of any length as features for training a native language classifier. Starting with surface n-grams, we investigate two degrees of abstraction incorporating parts-of-speech. The approach outperforms previous work employing a comparable data setup, reaching 89.71% accuracy for a task with seven native languages using data from the International Corpus of Learner English (ICLE). We then investigate the claim by Brooke and Hirst (2011) that a content bias in ICLE seems to result in an easy classification by topic instead of by native language characteristics. We show that training our model on ICLE and testing it on three other, independently compiled learner corpora dealing with other topics still results in high accuracy classification.

TITLE AND ABSTRACT IN GERMAN

Muttersprachenerkennung mittels rekurrenter N-Gramme – Untersuchungen zur Abstraktion und Domänenabhängigkeit

Die Muttersprachenerkennung befasst sich mir der Erkennung der Muttersprache eines Autors auf der Basis eines Textes, der von diesem Autor in einer Zweitsprache verfasst worden ist. In der vorliegenden Arbeit diskutieren wir die systematische Verwendung rekurrenter N-Gramme aller Längen als Features für das Trainieren eines Muttersprachen-Klassifizierers. Beginnend mit oberflächenbasierten N-Grammen, untersuchen wir zwei Stufen der Abstraktion unter Verwendung von Wortarten. Unser Ansatz liefert eine Klassifikationsgenauigkeit von 89.71% für Texte aus dem International Corpus of Learner English (ICLE) mit sieben unterschiedlichen Muttersprachen und übertrifft somit die bisherigen Ergebnisse auf vergleichbaren Daten. Ferner untersuchen wir die Behauptung von Brooke und Hirst (2011), dass inhaltliche Aspekte des ICLE zu einer einfacheren Klassifikation der Texte nach dem Thema anstatt nach der Muttersprache führen könnten. Wir zeigen, dass ein auf ICLE Daten trainiertes Modell auch bei Tests auf drei unabhängig erstellten Lernerkorpora eine hohe Klassifikationsgenauigkeit ermöglicht.

KEYWORDS: Native Language Identification, Author Profiling, Text Classification, Second Language Acquisition, Learner Corpora.

KEYWORDS IN GERMAN: Muttersprachenerkennung, Autoren-Profiling, Textklassifkation, Zweitspracherwerb, Lernerkorpora.

Proceedings of COLING 2012: Technical Papers, pages 425–440,
COLING 2012, Mumbai, December 2012.

1 Introduction

Inferring characteristics of an author by automatically analyzing that author's texts is a task that is increasingly drawing attention in recent years. Traits such as gender, age, level of education or native language are some of the properties targeted thus far (e.g., Koppel et al., 2005; Estival et al., 2007; Wong and Dras, 2009).

The work presented in this paper examines one particular characteristic, namely the author's *native language*, with the task being to infer it from a text written in a second language. So we explore the task of *Native Language Identification (NLI)*, which is of interest for a number of reasons. The impact of one's native language on a second language is studied in *Second Language Acquisition* (SLA) research, aimed at understanding how languages are acquired and how language works in general. Of particular relevance here is the notion of Transfer: *"Transfer is the influence resulting from similarities and differences between the target language and any other language that has been previously [...] acquired."* (Odlin, 1989, p. 27). Given the increasing availability of second language corpora with different native languages as well as powerful classification and evaluation techniques, it becomes viable to empirically explore and verify hypotheses regarding the existence and nature of L1 Transfer. Complementing the conceptual relevance for SLA, NLI also is of practical relevance for applications such as systems for profiling phishing emails (Estival et al., 2007) or in the context of learner modeling for intelligent language tutoring systems (Amaral and Meurers, 2008).

NLI started to attract interest less than ten years ago (Koppel et al., 2005), so the area still is quite young, with fundamental questions waiting to be addressed: Is the L1 Transfer effect strong and distinctive enough across domains to support an automatic classification with a reasonable degree of reliability for the typical available document lengths? Which language properties are the most appropriate ones to use as classifier features for the given task and can they reliably be identified? How well can a surface-based approach fare in the task and what is the effect of abstracting away, e.g., to distributional classes such as parts-of-speech (POS)?

In this paper, we consider the NLI task as a text classification problem with the different native languages as the classes. Inspired by the variation n-gram approach to corpus annotation error detection (Dickinson and Meurers, 2003, 2005; Boyd et al., 2008), we will follow a data-driven approach based on *recurring n-grams* as features in a machine learning setup capable of handling large feature spaces. The aim of our work is to contribute a piece to the overall puzzle to solve, starting with a particular take on surface features, recurring word-based n-grams of any size, and exploring the effect of incrementally introducing POS as abstractions. In the second part of the paper, we then explore the generalizability of our results across corpora.

2 Related Work

Koppel et al. (2005) used a subset of the first version of the *International Corpus of Learner English (ICLE)* (Granger et al., 2002) as data set. The ICLE corpus consists of essays written by non-native English speakers at a similar level of English proficiency, namely higher intermediate to advanced. Koppel et al. included texts for five native languages: Bulgarian, Czech, French, Russian and Spanish. Each native language was represented by 258 essays. They used a *Support Vector Machine* (SVM) as classifier and defined features based on the occurrence of function words, character-based n-grams, rare POS bi-grams as well as some error types (e.g., certain spelling errors). Testing was performed using 10-fold cross-validation. The best classification accuracy of 80.2% was obtained using all of the mentioned features combined.

Tsur and Rappoport (2007) replicated Koppel et al. (2005) and investigated the hypothesis that the choice of words in the second language is strongly influenced by the frequency of native language syllables. In support of their hypothesis, the authors report that an approximation using character bi-grams alone allows classification accuracy of about 66%. Since the corpus contains learner essays on several different topics, they also investigated whether the classification with such surface features is influenced by a content bias. Using a variant of the *Term Frequency - Inverted Document Frequency* content analysis metric, they conclude that if a content bias exists in the corpus, it only has a minor effect.

Estival et al. (2007) used a corpus of English emails as data incorporating three native languages, namely English, Arabic, Spanish, and considered a range of different demographic as well as psychometric traits including the native language for author profiling purposes. They used a wide range of features at different levels: character-based features such as frequency of punctuation marks, lexical features such as function words as well as POS, and some features at the structural level such as paragraph breaks. Using *Information Gain* as feature selection technique and *Random Forest* classification, they obtained an accuracy of 84.22%.

Wong and Dras (2009) used the second version of the ICLE corpus (Granger et al., 2009) as data and compiled a data set consisting of seven native languages, namely Bulgarian, Czech, French, Russian, Spanish, Chinese and Japanese, each represented by 70 essays for training and 25 essays for testing (plus 15 additional essays for development). On the one hand, they employed lexical features, such as function words, frequently used character-based uni-, bi-, tri-grams as well as rare and most frequently used POS bi- and tri-grams. On the other hand, they used three syntactic error types as features: misuse of determiners as well as subject-verb and noun-number disagreement. Using an SVM classifier they obtained an accuracy of 73.71%. Extrapolating to a larger training set, they argue that this result is consistent with the findings reported by Koppel et al. (2005). However, the syntactic features used in their study did not improve the results obtained by employing lexical features alone.

Wong and Dras (2011) extended their previous work by investigating more general syntactic features compiled on the basis of parse trees, namely horizontal slices as well as cross-sections of parse trees. These features were used along with the set of lexical ones of Wong and Dras (2009). Using the same data set as Wong and Dras (2009) and a *Maximum Entropy* classifier, they obtained a classification accuracy up to 81.71%, showing that incorporating more sophisticated syntactic features can improve the results.

Brooke and Hirst (2011) conducted several experiments employing two different corpora, namely the ICLE and the Lang-8. The second corpus was compiled by the authors themselves based on the data available on `http://lang-8.com`. This web site contains short personal journal entries of different kinds (personal narratives, requests for translations of particular phrases, etc.), which are posted by English learners in order to obtain feedback from native speakers. Compared to the ICLE corpus, there is a disproportionately high number of contributors from Eastern Asia, the level of English proficiency seems to be significantly lower, and little is known about the context of the writing for Lang-8 (e.g., there is no specification of time or resources used). To obtain texts from Lang-8 which are comparable in size to those in ICLE, Brooke and Hirst (2011) created texts consisting of multiple Lang-8 entries. In their computational approach, they use character, word, and POS-based uni- and bi-grams (excluding proper nouns in case of word-based n-grams) as well as some function words as features. Based on a dataset from ICLE and Lang-8 consisting of seven native languages, namely Chinese, Japanese, French,

Spanish, Italian, Polish and Russian, with each of them represented by 200 texts from each of the two corpora, they conducted experiments using an SVM classifier in a single-corpus evaluation (using 10-fold cross-validation) and a cross-corpus evaluation (training on the one corpus, testing on the other). The single-corpus evaluation on ICLE data yielded an accuracy of 93.8% using all the features together, yet only 25% when training and testing on the Lang-8 data. The results of cross-corpus evaluation were very low, at 15.7% to 22.9%. Based on these results, Brooke and Hirst (2011) argue that a strong content bias is present in ICLE, allowing an easy classification by topic instead of by native language. However, it remains unclear whether the poor Lang-8 results are not due to the properties of this specific corpus, which seems to be highly heterogeneous and incoherent, and whether the poor cross-corpus evaluation results are of general importance or due to the nature of the Lang-8 corpus. Brooke and Hirst then explore the usefulness of artificial learner corpora, which they compiled using machine translation of native language data. The results yield up to 67% in a setup with two native languages. Brooke and Hirst (2012) extend their previous work and show that using automatically translated word bi-grams in combination with a new L1 Transfer metric yields up to 48.3% in a setup with four native languages when tested on ICLE data. The accuracies are far below those reported previously, but the approach promises a low content bias.

3 Data

For our first, core study we use a subset of the *International Corpus of Learner English* (ICLE v. 2; Granger et al., 2009). The overall ICLE corpus consists of 6,085 essays written by English learners of 16 different native language backgrounds. They are at a similar level of English proficiency, namely higher intermediate to advanced and of about the same age. Following the setup of Wong and Dras (2009), we randomly select a set of essays from the same seven native languages – namely, Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese – and we use the same data split with 70 essays for training and 25 essays for testing for each of the languages, resulting in a total of 490 essays for training and 175 for testing. As in Wong and Dras (2009), we only included essays between 500 and 1000 words in length. We tokenized the essays and removed all punctuation marks, special characters and capitalization. Thus each essay is represented as an array of lower-case words.

To get a better sense of how well our approach performs, we conducted ten experiments. We select the data for each of them randomly from the full set of ICLE essays within the mentioned length range. We thus are able to observe the variance of the results based on ten randomly selected samples from the overall corpus subset matching the described criteria. We first describe one of the ten experiments in detail and then turn to the overall ten experiments.

4 Features

Different from previous research, in this study we explore *recurring n-grams of all occurring lengths* as classifier features. By *recurring* we here mean all n-grams that occur in at least *two* different essays of the training set d (the test set is held out, i.e., not considered for determining the features). *Of all occurring lengths* means all recurring n-grams up to the maximum possible n value occurring in d, i.e., all n-grams with $1 \leq n \leq max_n(d)$.

On the one hand, we use recurring *word-based* n-grams directly, i.e., the surface forms. On the other hand, we explore two different classes of recurring *POS-based* n-grams as a generalization, based on a POS tagged version of the corpus using the PennTreebank tagset (Santorini, 1990). In sum, we define our features based on the following three classes of recurring n-grams:

Word-based n-grams (word n-grams): strings of words, i.e., the surface forms

- $n = 1$: *analyzing, attended, ...*
- $n = 2$: *aspect of, could only, ...*
- $n = 3$: *is capable of, the assumption that, ...*
- ...

POS-based n-grams (POS n-grams): *all words* are converted to the corresponding POS tags

- $n = 1$: *nnp, md, nns, vbd, ...*
- $n = 2$: *nns md, nn rbs, nn rbr, cc wdt, vbp jjr, vbp jjs, ...*
- $n = 3$: *cd wdt md, vbp nn md, dt rbr cc, nn jj in, ...*
- ...

Open-Class-POS-based n-grams (OCPOS n-grams)[1]**:** *nouns, verbs, adjectives* and *cardinal numbers* are converted to the corresponding POS tags

- $n = 1$: *far, vbz, much, jj, ...*
- $n = 2$: *nn whenever, jj well, jjs vbd, vbg each, nn always, ...*
- $n = 3$: *vbp currently jj, only to the, cd vbz jj, vb if there, ...*
- ...

We explore the whole range of n values as well as all possible $[1, n]$ intervals. Figure 1 depicts the counts of *different n-grams* for each n (for uni-grams, bi-grams, tri-grams, etc.) and Figure 2 for each $[1, n]$ interval (for uni-grams alone, uni-grams and bi-grams together, uni-grams, bi-grams and tri-grams together, etc.). There are large differences in terms of feature counts, depending on the particular n-gram class and the value of n used. The figures show that increasing the number of different POS tags leads to more possible different features (up to about 160,000 in our setup). The reason for that is the ability of POS to bridge some break points in the word sequences (i.e., places where different words occur, thus ending a recurrent surface n-gram) and hence to lead to more longer n-grams. Thus the n-grams including POS tags may also reach higher n values: For the word-based n-grams $max_n(d) = 29$, whereas POS-based n-grams reach $max_n(d) = 30$ in the training set used.

As expected, the feature counts fall rapidly as the n value passes a certain (n-gram class dependent) threshold (see Figure 1). Longer n-grams may potentially contain some specific information not contained in the shorter ones – they may capture, e.g., transliterations of native idioms (Milton and Chowdhury, 1994). So we do not discard any features a priori. The aim is to investigate up to which value of n the n-grams may be worth considering despite being rare.

We use binary feature vectors as classifier input, i.e., each essay is represented by a vector containing $\{0, 1\}$ values. If an essays contains a particular n-gram, then the corresponding value in the vector is 1, and 0 otherwise. Since the n-gram frequencies (especially in case of the longer ones) are rather low, we consider such a representation to be a reasonable simplification.

5 Tools

To extract all recurring n-grams, we implement a dynamic programming algorithm collecting all n-grams of length n based on the n-grams collected for $n-1$. The algorithm terminates once no n-grams for a given length can be found in the given data. To obtain the n-gram classes incorporating POS tags, we used the *OpenNLP* POS-tagger (`http://opennlp.apache.org`).

[1] Similar representations are also used by Baroni and Bernardini (2006) for the identification of "translationese".

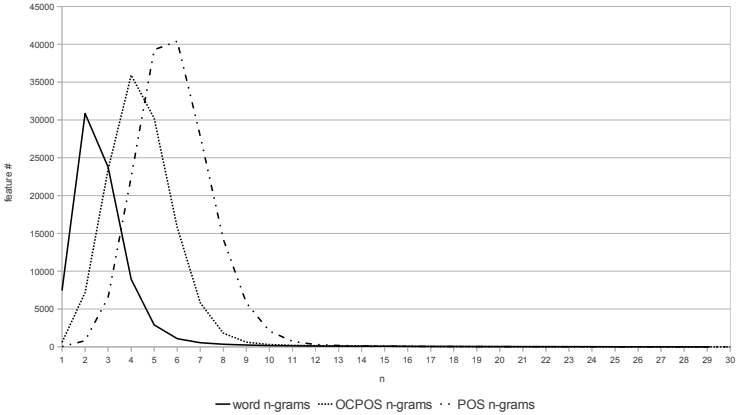

Figure 1: Feature counts for single n n-gram settings for the single ICLE sample

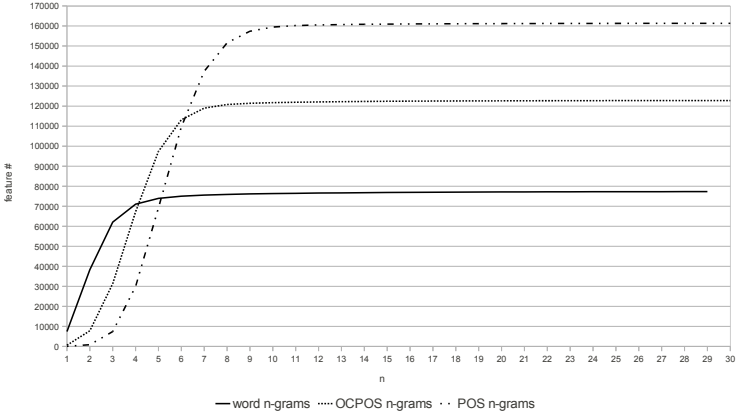

Figure 2: Feature counts for $[1, n]$ n-gram settings for the single ICLE sample

To choose the classifier to use, we conducted several preliminary tests employing different machine learning tools. We explored using *TiMBL* (Daelemans et al., 2007), which provides an implementation of the k-NN algorithm, incorporating a range of distance metrics. We then tested different Support Vector Machines (SVMs), which are well-known for their ability to handle large feature sets: *WEKA SMO* (Platt, 1998; Hall et al., 2009), *LIBSVM* (Chang and Lin, 2011), and *LIBLINEAR* (Fan et al., 2008). In our trials, the *LIBLINEAR* classifier yielded by far the best results and was in addition usually faster than the others as well. Hence, we employ the *LIBLINEAR* classifier in our study.

6 Results

The classification results for all feature settings are presented in Figures 3 and 4.

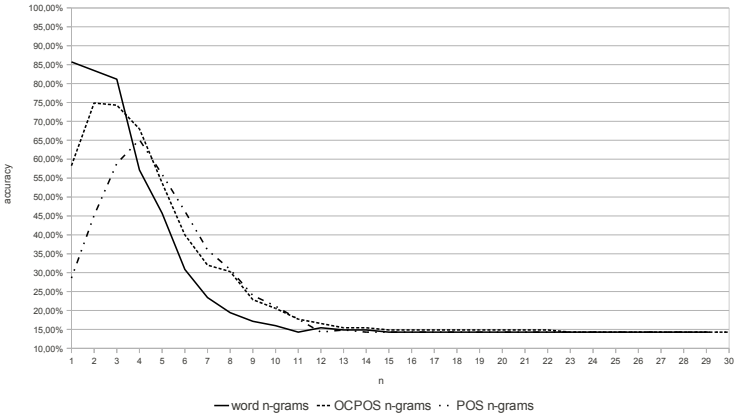

Figure 3: Results for single n n-gram settings for the single ICLE sample

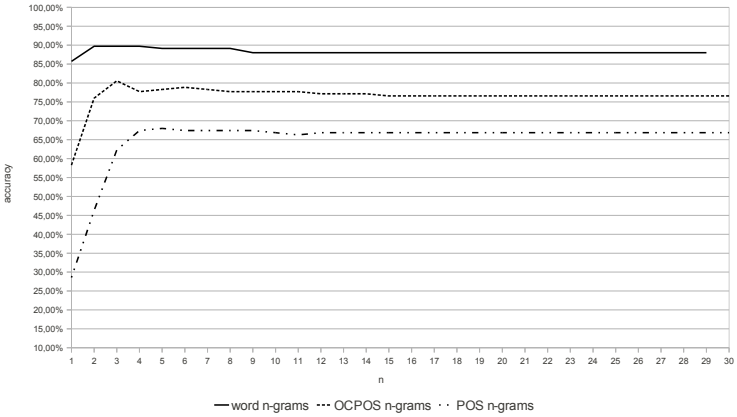

Figure 4: Results for $[1, n]$ n-gram settings for the single ICLE sample

Figure 3 shows the classification accuracy for all n values of the n-grams separately (i.e., for uni-grams, bi-grams, tri-grams, etc.), whereas Figure 4 depicts the classification accuracy for all $[1, n]$ intervals (i.e., for uni-grams alone, uni- and bi-grams together, uni-, bi-, tri-grams together, etc.). There are seven different native languages as classes, each represented by an equal number of essays, so 14.29% is the random baseline against which to interpret the results.

Best Accuracy Range The highest accuracy achieved by our recurring n-gram approach is 89,71% using word-based n-grams with intervals from $[1,2]$ to $[1,4]$. This is 16% higher than the best result reported by Wong and Dras (2009) and about 8% higher than that reported by Wong and Dras (2011) on a comparable data set. Brooke and Hirst (2011) reported a slightly better result, 93.8% for seven native languages, but as discussed in Section 2 they used more data and a different data split.

The confusion matrix in Table 1 shows the distribution of correctly classified as well as misclassified samples for each of the native languages. The performance on the different native languages is generally comparable; only the result for Russian is slightly below the others.

	BG	**CN**	**CZ**	**FR**	**JP**	**RU**	**SP**
BG	23	0	0	0	0	2	0
CN	0	24	0	0	1	0	0
CZ	0	0	23	1	0	1	0
FR	1	0	0	22	0	0	2
JP	0	0	1	0	24	0	0
RU	1	0	3	1	1	19	0
SP	1	1	0	1	0	0	22

Table 1: Confusion matrix for the best result for the single ICLE sample: 89,71%, word-based n-grams, $[1,2]$; BG: Bulgarian, CN: Chinese, CZ: Czech, FR: French, JP: Japanese, RU: Russian, SP: Spanish

However, there are clear differences in terms of accuracy between the n-gram classes utilized in this study. As mentioned above, the best result is obtained using pure surface forms, the word-based n-grams. The more different POS tags are incorporated, i.e., the bigger the step from the surface to the more general forms, the lower the accuracy. The information loss involved in the abstraction thus outweighs the broader applicability. The best results are presented in detail in Table 2.[2]

	n Intervals			**Single n**		
Features	$[1,n]$	**Accuracy**	**Feature #**	n	**Accuracy**	**Feature #**
word n-grams	2	89.71%	38,300	1	85.71%	7,446
OCPOS n-grams	3	80.57%	31,263	2	74.86%	7,176
POS n-grams	5	68.00%	69,139	4	65.14%	22,462

Table 2: Best results for the single ICLE sample

Table 2 shows that POS-based n-grams, i.e., features at the highest generalization level, yield about 13% lower accuracy than the Open-Class-POS-based n-grams, and the latter are performing about 9% worse than word-based n-grams. There is a gap of about 22% between the surface-based and the most generalized n-gram representation used in our study. However, even the most general POS-based n-grams still yield a result of 68%, which is reasonably high considering the baseline of 14.29%. The accuracy of 80.57% obtained using Open-Class-POS-based n-grams is in line with the best results published for a comparable data set.

[2] If more than one setting per feature class yields the same best accuracy, only the lowest n or $[1,n]$ interval is listed.

Different n Values Using intervals of n always leads to better results than using n-grams of a particular single n value alone (see Figures 3 and 4). One can also see that the more POS generalization is incorporated, the longer n-grams are needed to obtain the best results. In this study, the accuracy benefited from n-grams up to $n = 5$. Thus n-grams with $n > 3$, which are generally not considered in the related research, are not a priori useless.

The longer n-grams in the range of $6 \leq n \leq 10$ seem to be too sparse to improve on the accuracy obtained by intervals of shorter n-grams, at least in the data used in this study. There are a lot of different n-grams in that range, especially for n-grams with POS incorporated (see Figure 1), but the impact of lots of different features, with each occurring only in a few essays, seems to be very limited. Moreover, using them in intervals with n-grams of lower n values usually decreases the accuracy (see Figure 4). Thus they seem to introduce some noise into the feature set. However, increasing the size of the data set or incorporating more sophisticated generalizations may still allow such n-grams to become useful.

Finally, "very long" n-grams, i.e., n-grams with $n > 10$, usually encode a few, predetermined phrases, such as the wording of the topic the essay is about, or consist of some other copied passages. Hence, they are unlikely to be relevant for the given NLI task.

Reliability of the Findings Since the results described above are based on a single experiment, one may wonder, how generalizable those findings are. As mentioned in Section 3, we thus conducted nine further experiments. Summing up the results of the ten experiments, we computed the *mean accuracy* values along with the *Sample Standard Deviations (SSD)*. Given that the $max_n(d)$ value varies for the ten training sets, one cannot average over all n for all of the experiments. But as discussed in the previous paragraph, n-grams with $n > 10$ are unlikely to be useful for the purposes of the given task. Hence, we report the accuracy results for the $1 \leq n \leq 10$ range. Figure 5 shows the results for $[1, n]$. Overall, the means curves are very similar to the curves we presented in Figure 4.

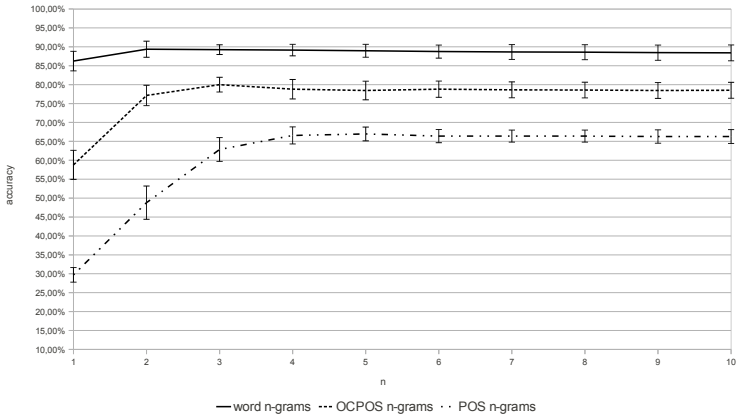

Figure 5: *Mean accuracy* and *SSD* for $[1, n]$ n-gram settings for the ten ICLE samples

The overall best outcomes are shown in Table 3. The best mean accuracy result of 89.37% is yielded by the same setting, namely by the word-based n-grams using the $[1, 2]$ interval.

	n **Intervals**			**Single** n		
Features	$[1, n]$	**Mean accuracy**	**SSD**	n	**Mean accuracy**	**SSD**
word n-grams	2	89.37%	2.12%	1	86.23%	2.59%
OCPOS n-grams	3	80.00%	1.94%	2	73.71%	2.68%
POS n-grams	5	66.97%	1.82%	4	60.91%	3.38%

Table 3: Best mean accuracy results for ten ICLE samples

This best mean accuracy over ten experiments is only 0.34% lower than the corresponding best result from the single experiment described in the *Best Accuracy Range* paragraph of the current section. The SSD with values around 2% for the best performing settings indicates that there is little variance among the experiments.

Discussion The ICLE contains essays from a range of topics, so one may wonder about the impact of the contents on the native language identification. Using only essays of the same topic would in principle be preferable, but it would significantly reduce the amount of data available. As mentioned in Section 2, Tsur and Rappoport (2007) argued that such a content bias is rather marginal for the subset of the ICLE they used. In contrast, the findings of Brooke and Hirst (2011) suggested a high topic bias in the ICLE data. In order to obtain more independence from the content of an essay, there is a clear need for some abstraction away from the surface encoding form and meaning together. Yet, the features in our study with the highest level of generalization and thus probably the lowest topic bias, recurring POS-based n-grams, provide results about 22% below those purely based on surface forms. A combination of surface and generalized forms may be a reasonable middle ground. In that light, the *Open-Class-POS-based n-grams* appear attractive since they replace many of the topic-specific meaning distinctions with POS-tags. They are less tied to the meaning than word-based n-grams, but still yield high accuracy with relatively low feature counts in the best performing n range. At the same time, Brooke and Hirst (2011) observe a comparable drop for word and POS-based features in cross-corpus evaluation with the Lang-8 corpus, and Golcher and Reznicek (2011, p. 31) show that POS n-grams still contain information relevant to topic classification for the German learner corpus FALKO. More research thus is needed to verify which features are sufficiently general and applicable across corpora. We address this issue in the next section.

7 Investigating the cross-corpus generalizability of the results

To address the question whether the models trained and evaluated on the ICLE corpus generalize to other learner corpora, we conducted a second set of experiments.

Data In this second study, we use four different learner corpora. Complementing the ICLE introduced above, we use the NOCE, USE and HKUST corpora compiled by independent research teams.

The *Non-Native Corpus of English / NOCE* (Díaz Negrillo, 2007, 2009) is an English learner corpus consisting of mainly argumentative essays on several topics written by Spanish native speakers. The data was collected at the University of Granada and the University of Jaén using texts by undergraduates pursuing an English degree. The corpus contains 1,022 essays.

The *Uppsala Student English Corpus / USE* (Axelsson, 2000, 2003) is a corpus of learner English consisting of texts written by Swedish students at the Department of English at Uppsala University. The texts contained in the corpus are essays written as part of the regular curriculum and cover several topics of different genres, e.g., argumentation, reflection, literature course assignment, etc. The corpus contains 1,489 essays. Since the essays from the other corpora used in this study are mostly argumentative, to obtain comparable data in terms of the text properties we use only the argumentative subset of the corpus (from the first term). This USE subcorpus consists of 344 essays.

The *Hong Kong University of Science and Technology English Examination Corpus / HKUST* (Milton and Chowdhury, 1994) is an English learner corpus containing texts written by Chinese native speakers. The version of the corpus we are using consists of 1,100 argumentative essays on different topics collected 1992 during the public matriculation examination, which is taken each year by students leaving secondary school. For the present work, we took a 8% random sample of the whole corpus, consisting of manually tagged 77 essays as described in Milton and Chowdhury (1994, p. 128).

As preprocessing, we removed all types of meta-information and annotation contained in the learner corpora (personal information about the author of the text such as the age or the native language, topic tags, error annotation, etc.) as well as all punctuation marks, special characters and capitalization, and we tokenized the essays. Hence, as in the first study each text is represented as an array of lower-case words.

Based on the data described above, we explore the NLI task using a setup with three native languages: Spanish, Swedish and Chinese. First, we compile *two separate test sets*. The first test set consists of randomly selected 70 essays per native language from ICLE. To compile the second test set, we randomly select 70 essays per native language correspondingly from HKUST and USE and 140 essays from the NOCE corpus. Since the NOCE essays tend to be shorter than the other ones, we merge the 140 essays pairwise to obtain 70 texts of a size comparable to the essay size from the other corpora. The texts on average contain 620 words. Second, we compile *ten separate training sets*. Each training set consists of randomly selected 140 essays per native language from the overall ICLE corpus (without the essays selected for the ICLE test). Thus we obtain ten separate training sets with 420 essays each, randomly selected from the ICLE corpus, and two separate test sets with 210 texts each, one compiled using ICLE alone and another compiled using NOCE, USE and HKUST.

This setup allows us to perform ten *single-corpus* evaluations (i.e., training and testing on the same corpus) on the ICLE data alone as well as ten *cross-corpus* evaluations (i.e., training on the one corpus and testing on another corpus) using ICLE data for training and NOCE, USE, HKUST data for testing. With ten separate ICLE training sets, we are able to build ten different classifier models and to observe the variance in the generalizability of the patterns learned on different ICLE subsets. We thus are able to observe the generalizability of the ICLE patterns to other corpora in direct comparison to ICLE itself.

Results Based on the ten different training sets, we conducted tests for each $[1, n]$ n-gram interval with $1 \leq n \leq 10$ using the two best performing n-gram classes (i.e., word- and OCPOS-based n-grams as features), and performed both a single-corpus evaluation and a cross-corpus evaluation. We thus obtained 400 separate accuracy values overall (10 training sets $\cdot$ 2 n-gram classes $\cdot$ 10 n-gram intervals $\cdot$ 2 evaluation types).

Figure 6 sums up the results by depicting the *mean accuracy* values on the two test sets obtained using ten different training sets for both n-gram classes and each of the ten n-gram intervals along with the random baseline. Since in this set of experiments there are three different native language classes, each represented by an equal number of essays, we obtain 33.33% as a random baseline against which to interpret the results.

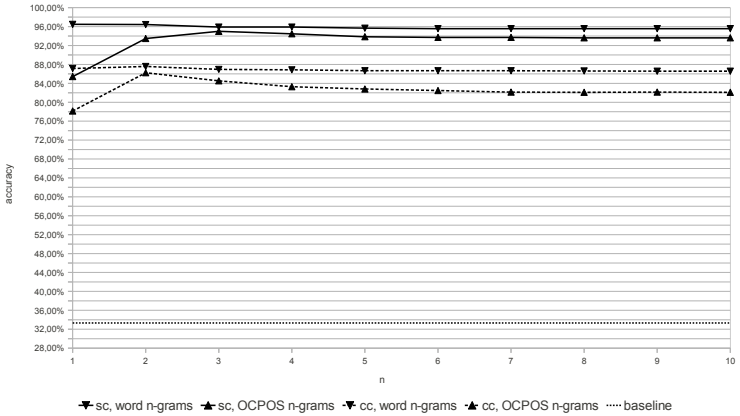

Figure 6: *Mean accuracy* for $[1, n]$ n-gram settings for the ten ICLE training sets (sc = single-corpus, cc = cross-corpus evaluation)

We left the SSD bars out of Figure 6 to keep it readable, but it naturally is interesting to consider the variance. Figure 7 shows the single- and cross-corpus accuracies for the word-based n-grams from Figure 6 together with the corresponding SSD. Figure 8 presents the same for the OCPOS-based n-grams. We see that in both figures the variance is low, with the cross-corpus evaluation showing slightly higher SSD values as expected.

Table 4 shows the best accuracies for both feature classes along with the corresponding SSD values obtained on the two different evaluation types as well as the corresponding n intervals. Though the best performing n-gram intervals differ for both feature classes on single-corpus evaluation, in the cross-corpus evaluation *recurrent bi-grams* perform best for both.

At the end of Section 6, we hypothesized that the more abstract OCPOS-based n-grams may perform better than the surface-near word-based ones in cross-corpus evaluation. However, the accuracies obtained using word-based n-grams are on average as good or better than the ones obtained using OCPOS-based n-grams (see Figure 6 and Table 4). Apparently people with different native language backgrounds make lexical choices which are indicative across a range of topics. A first qualitative analysis points to the use of predicates such as *get*, *take*, *choose*, *make use of*, *consider*, *be able to*, *understand*, or *suggest*. A precise characterization of the nature of this lexical material seems relevant to investigate in future work.

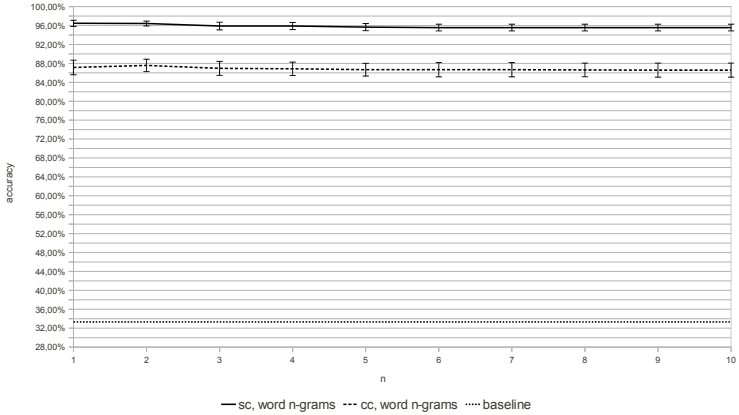

Figure 7: *Mean accuracy* and *SSD* for $[1, n]$ n-gram settings for the ten ICLE training sets, recurring *word-based n-grams* as features (sc = single-corpus, cc = cross-corpus evaluation)

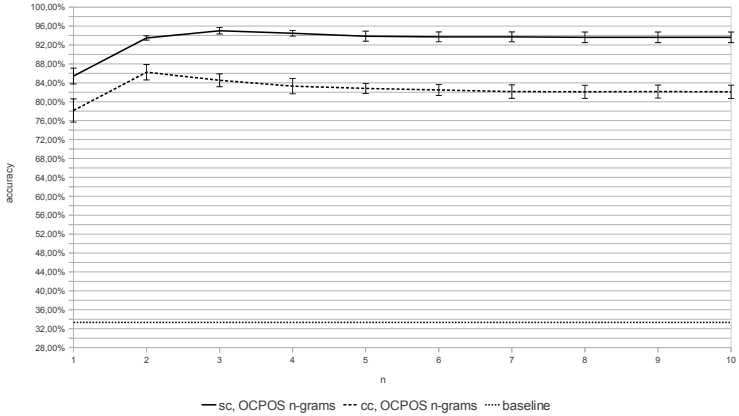

Figure 8: *Mean accuracy* and *SSD* for $[1, n]$ n-gram settings for the ten ICLE training sets, recurring *OCPOS-based n-grams* as features (sc = single-corpus, cc = cross-corpus evaluation)

Features	**Evaluation**	$[1, n]$	**Mean accuracy**	**SSD**
word n-grams	single-corpus	1	96.48%	0.64%
	cross-corpus	2	87.57%	1.32%
OCPOS n-grams	single-corpus	3	95.00%	0.68%
	cross-corpus	2	86.24%	1.63%

Table 4: Best results for ten ICLE training sets

Domain Dependence The experiments we ran with the NOCE, USE and HKUST corpora show far higher accuracies for the cross-corpus evaluation than what is reported by Brooke and Hirst (2011) for the Lang-8 corpus. In a setup with a random baseline of 14.2%, Brooke and Hirst (2011) report 70.1% – 93.8% (depending on the employed feature set) on single-corpus evaluation using ICLE, but only 15.7% – 17.0% for cross-corpus evaluation, training on ICLE and testing on Lang-8. In contrast, in a setup with a random baseline of 33.33% we obtained a best result of 95% – 96.48% (depending on the employed n-gram class) on single-corpus evaluation using ICLE, and 86.24% – 87.57% in a cross-corpus evaluation setup with training on ICLE and testing on NOCE/USE/HKUST (see Table 4 and Figure 6). Thus when using ICLE for training and another corpus instead of ICLE for testing, there is a drop of about 54% – 77% in Brooke and Hirst (2011) but only around 9% in our work. The dramatic drop Brooke and Hirst observed thus seems to be caused by some characteristic of the Lang-8 corpus and not by a general failure of the models learned on the ICLE corpus to generalize to other learner corpora.

The corpora we used for the cross-corpus evaluation were compiled by different research teams using their own essay topic lists. To investigate whether there still may be some topic overlap, we extracted the topics from our NOCE/USE/HKUST test set as well as from the ICLE training set yielding the best cross-corpus evaluation results. In both cases there were more than 100 different topics, and none of them matched between ICLE used for training and NOCE/USE/HKUST used for testing in the cross-corpus setup. Thus topic overlaps seem very unlikely to have notably skewed the results in our cross-corpus evaluation.

Conclusion

In this paper, we explored the task of *Native Language Identification (NLI)*. We derive three different classes of *recurring n-grams* as features, namely *word-*, *POS-* and *Open-Class-POS-based n-grams*. We use these features in a machine learning setup employing a Support Vector Machine (SVM) classifier on randomly selected data from the ICLE corpus incorporating seven different native languages. The best performing class are the word-based n-grams with an accuracy of 89.71%, which compares well to the 81.71% reported by Wong and Dras (2011) as the highest accuracy achieved thus far for a comparable data setup. To investigate the variance, we conducted nine further experiments based on random samples from ICLE. The mean accuracy values obtained from the overall ten experiments are very similar to those from the first experiment. The variance of the outcomes is moderate, with SSD being about 2% for the best performing settings. The bigger the step from the surface-based to more generalized features, the lower the accuracy. The recurring n-gram approach employing Open-Class-POS-based n-grams yields an accuracy of 80.57% and using POS-based n-grams we obtained 68%, which still is reasonably high considering the random baseline of 14.29% for this task.

We then investigated the claim in Brooke and Hirst (2011) that surface-based NLI classification models trained on the ICLE corpus do not generalize to other learner corpora. For this purpose we conducted a second set of experiments comparing *single-corpus* and *cross-corpus* results. In contrast to their cross-corpus findings using the Lang-8 corpus, our results show that the patterns learned on ICLE do generalize well to other learner corpora. More specifically, we showed that training on ICLE and testing on three independently collected corpora, NOCE, USE and HKUST, still yields reasonably high accuracy values of about 88% for a NLI classification task with three native languages. The low results for the Lang-8 corpus reported in Brooke and Hirst (2011) thus must have other reasons, possibly a lack of consistency in the Lang-8 pieces combined into documents, or the very different nature of the ICLE and the Lang-8 data.

References

Amaral, L. and Meurers, D. (2008). From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context: Extending the Conceptualization of Student Models for Intelligent Computer-Assisted Language Learning. *Computer-Assisted Language Learning*, 21(4):323–338.

Axelsson, M. W. (2000). USE – The Uppsala Student English corpus: An instrument for needs analysis. *ICAME Journal*, 24:155–157.

Axelsson, M. W. (2003). *Manual: The Uppsala Student English Corpus (USE)*. Uppsala University, Department of English, Sweden. Available at `http://www.engelska.uu.se/Research/English_Language/Research_Areas/Electronic_Resource_Projects/USE-Corpus`.

Baroni, M. and Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

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

Brooke, J. and Hirst, G. (2011). Native language detection with 'cheap' learner corpora. In *Learner Corpus Research 2011 (LCR 2011)*, Louvain-la-Neuve.

Brooke, J. and Hirst, G. (2012). Measuring interlanguage: Native language identification with l1-influence metrics. In *Proceedings of the 8th ELRA Conference on Language Resources and Evaluation (LREC 2012)*, pages 779–784, Istanbul.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A. (2007). *TiMBL: Tilburg Memory-Based Learner Reference Guide, ILK Technical Report ILK 07-03*. Induction of Linguistic Knowledge Research Group Department of Communication and Information Sciences, Tilburg University, Tilburg, The Netherlands. Version 6.0.

Dickinson, M. and Meurers, W. D. (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, M. and Meurers, W. D. (2005). Detecting errors in discontinuous structural annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 322–329.

Díaz Negrillo, A. (2007). *A Fine-Grained Error Tagger for Learner Corpora*. PhD thesis, University of Jaén, Spain.

Díaz Negrillo, A. (2009). *EARS: A User's Manual*. LINCOM Academic Reference Books, Munich, Germany.

Estival, D., Gaustad, T., Pham, S., Radford, W., and Hutchinson, B. (2007). Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272.

Fan, R., Chang, K., Hsieh, C., Wang, X., and Lin, C. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874. Software available at `http://www.csie.ntu.edu.tw/~cjlin/liblinear`.

Gass, S. and Selinker, L., editors (1983). *Language Transfer in Language Learning*. Newbury House, Rowley, MA.

Golcher, F. and Reznicek, M. (2011). Stylometry and the interplay of topic and L1 in the different annotation layers in the falko corpus. In *Proceedings of Quantitative Investigations in Theoretical Linguistics 4*, pages 29–34, Berlin.

Granger, S., Dagneaux, E., and Meunier, F. (2002). *International Corpus of Learner English*. Presses Universitaires de Louvain, Louvain-la-Neuve.

Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. (2009). *International Corpus of Learner English, Version 2*. Presses Universitaires de Louvain, Louvain-la-Neuve.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Koppel, M., Schler, J., and Zigdon, K. (2005). 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.

Milton, J. C. P. and Chowdhury, N. (1994). Tagging the interlanguage of Chinese learners of English. In *Proceedings joint seminar on corpus linguistics and lexicology, Guangzhou and Hong Kong, 19-22 June, 1993, Language Centre, HKUST*, pages 127–143, Hong Kong.

Odlin, T. (1989). *Language Transfer: Cross-linguistic influence in language learning*. Cambridge University Press, New York.

Platt, J. C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank, 3rd revision, 2nd printing. Technical report, Department of Computer Science, University of Pennsylvania.

Tsur, O. and Rappoport, A. (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.

Wong, S.-M. J. and Dras, M. (2009). Contrastive analysis and native language identification. In *Australasian Language Technology Association Workshop 2009*, pages 53–61.

Wong, S.-M. J. and Dras, M. (2011). Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland, UK.