Automatic Identification of Learners' Language Background based on their Writing in Czech

Katsiaryna Aharodnik[1, 2]
katiaaharodnik@gmail.com

Marco Chang[1]
changreynam1@mail.montclair.edu

Anna Feldman[1]
anna.feldman@montclair.edu

Jirka Hana[3]
jirka.hana@gmail.com

[1]Montclair State University, New Jersey, USA
[2]City University of New York, The Graduate Center, New York, USA
[3] Charles University, MFF, Czech Republic

Abstract

The goal of this study is to investigate whether learners' written data in highly inflectional Czech can suggest a consistent set of clues for automatic identification of the learners' L1 background. For our experiments, we use texts written by learners of Czech, which have been automatically and manually annotated for errors. We define two classes of learners: speakers of Indo-European languages and speakers of non-Indo-European languages. We use an SVM classifier to perform the binary classification. We show that non-content based features perform well on highly inflectional data. In particular, features reflecting errors in orthography are the most useful, yielding about 89% precision and the same recall. A detailed discussion of the best performing features is provided.

1 Introduction

The role of a learner's native language (L1) in second language (L2) acquisition has been widely discussed in the theories of Second Language Acquisition (SLA) (Lado, 1957; Richards, 1971; Corder, 1975). The literature suggests that writers' spelling, grammar and lexicon in second languages are often influenced by patterns in their native language. However, the extent of the importance of L1 for acquiring L2 still cannot be determined exactly and remains a controversial topic of SLA research. Recently, the availability of learner corpora (e.g., Granger, 2003) has provided opportunities for verifying SLA hypotheses. The previous literature suggests that the best performing features for native language identification are largely the features that rely on the content of the data, such as word n-grams, function words and character n-grams (Kochmar, 2011; Koppel et al., 2005; Tsur et al., 2007). This means that future applicability of these features is limited to corpus specific data. The primary goal of our work is to address this problem. We use only non-content based features, part-of-speech tags (POS) and error tags. Exploring these features is useful for corpora independent approaches to native language identification. Our secondary goal is to analyze the features that perform best for highly inflectional data. We approach binary classification as the beginning step in the development of a systematic tool for recognizing a specific L1 from morphologically complex L2 data. We use machine learning techniques to identify features contributing to the classification between Indo-European (IE) and non-Indo-European (NIE) L1 backgrounds of learners of L2 Czech. We employ Support Vector Machines (Joachims, 1999) to perform the classification. The results of the experiments show that the non-content based features, especially error tags, are the strongest indicators of the learners' language background.

2 Related Work

The task of native language identification has branched out from authorship attribution and profiling. For instance, Mosteller and Wallace (1964) have worked with the Federalist Papers to

International Joint Conference on Natural Language Processing, pages 1428–1436,
Nagoya, Japan, 14-18 October 2013.

identify the papers' authors. They looked at function words as their features. There is plenty of work addressing authorship profiling for data in languages other than English, for instance Dutch, Greek and Arabic (Van Halteren, 2004; Stamatos et al., 2001; Estival et al., 2007).

For automatic native language identification researchers have been exploiting learner corpora (Koppel et al., 2005; Wong and Dras, 2009; Tsur et al., 2007; Kochmar, 2011). Several SLA theoretical foundations have been taken as the basis for this task, for instance, the Contrastive Analysis Hypothesis (CAH) (Lado, 1957). CAH posits that difficulties in second language learning are derived from the differences between the source and target languages. It is expected that the more similar L1 and L2 are, the acquisition of L2 is more natural for a learner, and fewer mistakes are made, or positive transfer takes place. At the same time, learners have more difficulties in acquiring L2 if there are more differences between the source and target languages, which results in negative transfer, or errors. Richards (1971) addresses the nature of errors within the Error Analysis approach. He outlines interlingual and developmental types of errors. Developmental errors are common errors for any learner of a given L2, while interlingual errors are specific for each L1 or a group of L1s. Hence, interlingual errors should possess a discriminatory nature (Corder, 1975) and are of primary interest for the purpose of the native language identification.

In the search for empirical evidence, the researchers have looked at learners' errors and other idiosyncrasies in non-native writings as cues to predict a learner's native language and to conform to the above theoretical approaches as well as the phenomenon of language transfer in particular (Jarvis et al., 2012; Tsur et al., 2007).

Koppel et al. (2005) look at 185 error types, including misspellings and syntactic errors as features. Besides errors, function words, character n-grams, and rare POS bigrams of non-standard English extracted from the Brown Corpus are used in the study. Koppel et al. (2005) experiment with essays from the International Corpus of Learner English in five source languages: Bulgarian, Czech, French, Russian and Spanish, and demonstrate that some types of errors are particularly useful for native language identification. Koppel et al. (2005) report slightly above 80% accuracy (with all features combined) compared to 20% baseline for 5-class classification. However, it is unclear from the study if the utilization of error-based features would improve the performance with the same significance if taken on their own. We can only infer from the diagram in the paper that error features perform at slightly higher than 50% on their own and do not contribute significantly to the performance when combined with other features. Koppel et al. (2005) make valuable observations about function words and character n-grams as the most discriminative features.

Wong and Dras (2009) explore the contribution of three syntactic errors to the same task: subject-verb disagreement, noun-number disagreement and misuse of articles. The L1 backgrounds in the experiments are Bulgarian, Czech, French, Russian, Spanish, Chinese and Japanese. The accuracy obtained from classification based on these three features is 24.57% for multi-class classification. This, compared to the baseline of 14.29%, appears to be a significant improvement at the 95% confidence level. To achieve better results, the syntactic features are combined with function words, character n-grams and POS n-grams. The best accuracy is 73.71% using a combination of all the features. The results of this study demonstrate that the three syntactic errors do not contribute noticeably to classification if used without other features.

Tsur et al. (2007) investigate native language identification in the domain of phonology. Tsur et al. (2007) work with essays of Bulgarian, Czech, Russian, Spanish and French L1 backgrounds. The essays are taken from the International Corpus of Learner English. Tsur et al. (2007) suggest that learners' L1 has a strong effect on word choice in their L2 writings. The results of the classification, based only on character bi-grams, yield an accuracy of 66% in a 5-class task. The results demonstrate that the learners' choice of words when writing in a second language is influenced by the phonology of their native language suggesting evidence for language transfer (Tsur et al., 2007).

Kochmar (2011) explores the Cambridge Learner Corpus and provides a systematic error analysis for a number of two-class classification experiments. From her results, we can see that errors contribute to native language identification for learner English data. The highest result of 100% classification accuracy is achieved for misspelled character quad-grams for the Danish-Swedish group of languages. Besides specific L1s, she also looks at binary classification between language families, such as Romance and Germanic. The best result, 84% accuracy, for this group is achieved for the combination of character tri-grams, POS n-grams and corpus derived error rates.

While the previous research widely exploits content based features, in our work we evaluate the usability of non-content based features and show that these features are reliable cues for native language identification. Moreover, all the above studies focus on learners' writings in English. In our work we investigate learner Czech data.

3 Experiments

3.1 Corpus

We use the Czech as a Second Language (CzeSL) (Hana et al., 2010; Jelinek et al., 2012; Rosen et al., 2013) corpus, a newly developed learner corpus of Czech. Czech is a West Slavic language that belongs to the Indo-European language family. It is a morphologically complex language with very rich derivational and inflectional morphology. It has seven noun cases, a complex declension and conjugation system, pronominal clitics and other morpho-syntactic structures, which all make the Czech language difficult for language learners. The CzeSL corpus is unique because it provides opportunities for a researcher to analyze learners' linguistic output of a highly inflectional target language. The corpus consists of several sub-corpora, with a total of 2 million words.

Out of this, about 200K words are corrected and error annotated using a two level annotation scheme. The first layer corrects individual words disregarding their context, for example spelling errors. In addition to manually annotated tags, e.g., error in ending (*incorInfl*) or error in stem (*incorBase*), some tags are added automatically, e.g., missing vowel accent (*formQuant0)* or erroneous character substitution (*formSingCh*). The second layer describes corrections within context that concern mostly morpho-syntactic and stylistic errors, e.g., the valency error (*dep*) includes noun declension and verb-noun agreement errors. For our purpose, we use both layers of annotation. The tagset is described in the annotation manual (Štindlová et al., 2012), in addition to the papers mentioned above.

Each document in the corpus is labeled with metadata information, including the author's proficiency level and native language background. The essays are encoded in the Prague Markup Language format.[1]

We report our findings for a binary classification between IE and NIE language backgrounds. We use 38[2] essays for lower intermediate (A2) and 38 essays for intermediate (B1) levels of proficiency. The essays are equally distributed among the language backgrounds within these levels. The essays are written on several topics, which are consistent throughout the groups. The topics include "My life in Prague", "The best/worst day in my life" and "Holidays", among several others. Every essay is written by a different author.

3.2 Native Speakers' Predictions

Prior to implementing machine learning classification, we decided to conduct an experiment with native speakers of Czech using the same data. The motivation for this experiment is to see whether it is possible for a native speaker to predict the learners' IE vs. NIE language backgrounds based on their essays.

We asked 24 Czech native speakers with NLP and /or linguistic backgrounds to read the essays and make their predictions about the language background of the writers. To avoid content bias, we substituted all proper names of places with a capital X; and personal names with generic names across all essays, e.g., Eva and Pavel.

There were a total of 76 essays to evaluate. An online questionnaire was created,[3] where native speakers read as many randomly assigned essays as they wanted and filled in the keys according to their predictions. The possible answers were "IE", "NIE" and "unclear". As the result of this experiment, an average accuracy of 55% was achieved. This result is only slightly better than the baseline of 50% for two-group classification.

The participants of our experiment all had some training in linguistics. This suggests that if the participants did not have any linguistic background, their performance on the task would probably be even lower. Moreover, the essays could have still contained some contextual cues about the authors' background, which might have triggered a higher result as well. The partic-

[1] http://ufal.mff.cuni.cz/jazz/pml/

[2] The CzeSL corpus is in final stages of development prior to public release. We used only those essays where we had access to all of: data itself, error annotation, morphological annotation and L1 metadata. We are planning to repeat the experiments, once there are more essays with such properties available.

[3] Using the open-source system developed by Jan Štěpánek https://github.com/choroba/inquiry/

ipants expressed their intuitions about the predictions. Specifically, they said they looked at the way the essays were written, the overall amount of errors. If an essay was written reasonably well the participants assumed that the author belonged to the IE group of learners, and vice versa. However, even having these intuitions in mind, our participants' performance was only slightly better than chance.

Overall, the experiment provided interesting observations and guided us towards a machine learning experiment.

3.3 SVM Classification

Data representation

For this experiment, our first goal is to see whether machine learning techniques are able to categorize the same set of data at higher performance rates than human native speakers using non-content based features. Our second goal is to see whether the native speakers' intuitions can be validated, specifically if it is the number of errors or other criteria that help to discriminate between the two classes.

We use the SVM-light classifier (Joachims, 1999). Each feature value is represented as a term weight of the feature, computed as a logarithmic ratio of the token frequency in the file to the total amount of tokens in the file.

$$S_{ij} = \text{round}\ (10\times\ (1+\log\ (tf_{ij}))/(1+\log(l_j)))$$

Equation 1. The formula for computing the term weight of a feature where S_{ij} is a term weight, tf_{ij} is the number of occurrences of term *i* in document *j*, and l_j is the length of the document. (Manning and Schuetze, 1999, p.580).

The feature set includes 264 most frequent POS bi-grams (3 or more occurrences in the data), 305 most frequent POS tri-grams and 35 error types extracted from the corpus. The total of POS n-grams for all essays amounts to 20,000. For error types, the total amount of error type tokens is 2000. After preprocessing, each essay is characterized by a vector with no more than 604 dimensions.

We report the classification results for the best performing parameters (C, γ) of the radial basis function (RBF) kernel SVM on the data set. Classification is performed by running the leave-one-out cross validation technique.

Results

Our best model is trained on a corpus that contains essays of the lower intermediate level of learners and receives 89% precision and the same recall using only orthographic types of errors as features. This is almost 40% higher than the baseline of 50% for the two-class classification. The precision and recall measures for each experiment are described in Tables 1-3.

Features	Precision	Recall
POS bigrams	**78**	74
POS trigrams	70	74
Errors	70	**78**
Errors+POS n-grams	71	75

Table 1. Classifier performance on Level B1 (intermediate) Czech

Features	Precision	Recall
POS bigrams	70	74
POS trigrams	70	78
Errors	**89**	**89**
Errors+POS n-grams	78	**95**

Table 2. Classifier performance on Level A2 (lower intermediate) Czech

Features	Precision	Recall
POS bigrams	74	**89**
POS trigrams	68	79
Errors	84	84
Errors+POS n-grams	**85**	**89**

Table 3. Classifier performance on Level A2 + Level B1 (combined) Czech

The results also demonstrate that the error features of the two levels combined perform distinctively well, at 84%. All features together show 85% precision and 89% recall. From the above experiments, we can see that non-content features such as POS tags and error tags perform well for highly inflectional language data. Moreover, error tags, on their own, may be considered good indicators of a class for this classification. Using features that do not reflect content makes our method more general and topic- and genre independent.

3.4 Classification experiment using error tags only

Following the native speakers' intuitions from the experiment described in Section 3.2, we can assume that the discriminative power of errors should not be surprising; learners of Czech of a NIE language background are likely to make more errors than the learners of the IE group due to the differences between L1 and L2. However, we need to perform a more detailed error analysis to conform or disagree with these intuitions.

The SVM classifier performs fairly well by using only error tags as features. In this section, we further investigate the results of the previous experiment, considering each feature separately.

To verify what error types are good markers for the two groups, we run additional classification experiments on each error-tag feature using the Weka implementation of the Naïve Bayes classifier (Witten et al., 2011). Naïve Bayes is a probability-based classifier. It implements Bayes' Theorem, the basic idea of which is the independence assumption, i.e. the presence or absence of one feature does not depend on the presence or absence of another feature. Naïve Bayes is simple to implement and interpret. We perform the 10-fold cross-validation technique on each data set for this task. We report the precision and F-measure which are calculated from the feature values normalized by total token amounts.

Results

The results of the Naïve Bayes classification experiments for both levels of proficiency are described in Table 4 and Table 5. The best performing features are shown in bold.

Table 4 displays the results for the intermediate level (B1) with morpho-syntactic and stylistic errors, the second layer in CzeSL. At this level, 5 errors out of 13 perform with precision and F-measure higher than 50%. These errors are the errors in valency (*dep*), errors in incorrect use of bookish, dialectal expressions and hypo-corrections (*stylOther*), misuse of grammatical forms (*use*), and odd constituent error (*odd*). The results suggest that these types of errors mostly contribute to the classification performance at this level.

Table 5 shows the results for the lower intermediate level (A2), the first layer of corrections in CzeSL. This level contains corrections of word-level errors, often of orthographic character. The errors that perform with precision and F-measure higher than 50% (6 out of 22) are missing vowel accent (*formQuant0)*, erroneous character substitution (*formSingCh*), incorrect use of 'i' instead of 'y'(*formY0)* ('i' and 'y' have the same pronunciation in Czech), incorrect use of 'y' instead of 'i' (*form Y1*), errors in inflections (*incorInfl*), and errors in stems (*incorBase*).

We also calculate error scores in order to identify which group (IE or NIE) tends to make more errors. The error scores are the ratios of the total frequency of an error type for all files to the total amount of errors in files.

The results of the Naïve Bayes classification suggest that depending on their nature, some errors contribute significantly to classification performance, but some have low discriminative power. For our purposes, these results are important for further analysis of the variety in the performance of the errors.

4 Discussion

Our results demonstrate that written texts regardless of the level of proficiency can be classified at 85% precision using non-content features, POS n-grams, and error tags combined. The results show that error-based features of two levels combined demonstrate a high performance of 84% suggesting that error annotated written learner Czech data provide reliable cues for distinguishing between the learners' IE and NIE backgrounds. The results show 89% precision and the same recall for a lower intermediate level (A2), which is annotated mostly for orthographic errors. The errors at the intermediate level (B1) with error tags of morpho-syntactic and stylistic character perform lower, at 70% precision and 78% recall.

The significant difference in the precision between the two levels suggests that the errors made by learners of a lower level of proficiency discriminate better than the errors made by higher level, i.e. if we have a fairly advanced learner, it would be harder to predict his or her language background. These results are not surprising, though more evidence is needed. It is more important to point out that the noticeably higher performance of orthographic errors suggests that these errors discriminate well between two language backgrounds within one level of proficiency. Consequently, this means that learners of two language backgrounds make errors specific for their L1 group. These results can be compared with previous observations made by other researchers (Tsur et al., 2007; Kochmar, 2011) that character n-grams extracted from learner data contribute significantly to classification performance. As Tsur et al. (2007) hypothesized, the learners' choice of characters in L2 reflects the phonology of their L1. Although we do not specifically look at character n-grams, the better discriminating power of errors of orthographic nature achieved in our experiments might as well reflect spelling conventions of a specific L1 group. Thus, such errors are more likely to be learners' L1 to L2 transfer errors.

Below, we provide a more detailed error analysis to verify the nature of errors and possible reasons for their contribution to the performance.

4.1 Error Analysis

Table 4 describes the results for Level B1. Lexical error (*lex*) shows 69% precision and 66% F-measure. These are lexicon or phraseology errors which occur, for instance, when learners misuse prepositions, choose false friends or false cognates instead of the correct variants. In the phrase *dopadlo to přírodně* the use of the adverb *přírodně* in this context results in an error. The intention of the author is more likely to say 'it ended naturally', but the word *přírodně* means 'naturally' in a sense of nature/non-artifical. The error scores show that learners of the IE language background tend to make more errors of this type (19.9/11.5). IE learners might use false cognates in Czech more often because of the similarity between L1 and L2 languages, e.g., Russian and Polish.

Stylistic errors (*stylOther*) reflect stylistic discrepancies, such as misuse of bookish, dialectal forms, slang, and hyper-corrections. For instance, in the phrase *pláči nad vejdělkem* 'be unhappy about the result' the correct use will be *pláču nad vydelkem*. There is a hypercorrection in the use of *pláči* instead of *pláču*. These types of errors occur exclusively within the IE group of learners (5.4/0) and perform with the highest precision of 79% and F-measure of 59%. This result together with the observations made for the *lex* type of error suggests that the L1s which are closer related to Czech influence the production of the L2 in a less subtle way than more distant languages. Specifically, the use of cognates in the incorrect context or use of incorrect stylistic variants might result in a transfer error in this case (Kroll et al., 2002).

Error Type	Precision	F-measure
agr	46	42
dep	**61**	**56**
lex	**69**	**66**
miss	50	49
stylOther	**79**	**59**
use	**64**	**64**
odd	53	51
sec	50	49
rflx	19	19
stylCol	50	44
vbx	46	44
cvf	44	39
ref	50	46

Table 4. Results of Naïve Bayes classification, Level B1.

Error Type	Precision	F-measure
formCap1	50	41
formCaron0	43	42
formCaron1	46	42
formQuant0	**76**	**73**
formReduChar	56	45
formSingCh	**74**	**70**
formVoiced	40	37
formY0	**70**	**55**
formY1	**81**	**65**
incorBase	**80**	**79**
incoInfl	**67**	**65**
styCol	39	39
wbdOtherJnt	77	49
flex	64	47
formQuant1	50	47
formVoiced0	50	38
fwNc	36	35
fwFab	50	43
missChar	60	49
wbdPreSplit	41	36
formDiaE	41	36
formMeta	76	44

Table 5. Results of Naïve Bayes classification, Level A2.

The valency error (*dep*), e.g., using *bojí se pes* instead of *bojí se psa* 'he is scared of a dog' (dog is a direct object, thus accusative *psa* instead of nominative *pes* must be used) yields 61% precision and 56% F-measure. Valency errors reflect the differences between the morpho-syntactic structures of L1 and L2. The use of grammar category (*use*) type of error shows 64% precision and 64% F-measure. This type includes errors in tense and aspect, incorrectly formed comparative, and singular instead of plural among others. For instance, using *včera bude sněžit* 'yesterday it will snow' instead of *včera sněžilo* results in a verb tense error, future form *bude sněžit* is used instead of the past *snežilo.*The two errors above largely concern Czech morpho-syntax. Thus, greater differences between L1 and L2 grammatical structures might trigger a higher amount of errors within the NIE group. However, a more detailed analysis of error distribution within each group and probably a larger data set are needed to investigate this claim.

The results for Level A2 are displayed in Table 5. The missing vowel accent (*formQuant0*) error occurs in such cases as incorrect *vzpominám* vs. correct *vzpomínám,* or *doufam* vs. *doufám.* This error type performs at 76% precision and 73% F-score. The erroneous character substitution error (*formSingCh*), e.g., incorrect *otevrila* vs. correct *otevrela* or *vezmíme* vs. *vezmeme*, performs well at 74% precision and 70% F-measure. The error scores show that the NIE learners tend to make more errors of this type (6/11). Errors in inflectional endings (*incoInfl*), e.g., using *plavám* instead of *plavu* 'I swim' (1Sg ending '-ám' of one paradigm is used with a verb of another), perform at 67% precision and 65% F-measure. Errors in stems (*incorBase*) e.g., using *ditem* instead of *dítětem,* discriminate rather well, at 80% precision and 79% F-measure. The two types of errors that describe incorrect use of 'i' and 'y' (*formY0* and *formY1*, respectively) show high precision (70% and 81%) and F-measure (55% and 65%). The NIE learners make more errors of type *Y0* (1.4/7.4), e.g., *pražskích* instead of *pražských, vipije* vs. *vypije,* whereas
the IE group solely makes errors in the other type, *Y1* (1.5/0), e.g., *hlavným* instead of *hlavním*, *líbyl* vs. *líbil*. The above results suggest that learners might make some motivated spelling choices related to their L1 backgrounds (Jarvis et al., 2012; Tsur et al., 2007). For instance, speakers of Russian and Belarusian, IE group, would use the letter 'y' more often in the ending because it corresponds to the phonological equivalent of the letter 'ы' of the Cyrillic alphabet, e.g., *главным* in Russian will be a phonological equivalent of incorrect *hlavným* in Czech.

Our analysis of the best performing features shows that learners of *both* language backgrounds within each level of proficiency produce errors that discriminate well and vary in nature between IE and NIE learners. Thus, we cannot strictly follow the intuition that the NIE learners make more errors, although these results might change with a larger number of learner essays. The error analysis at Level B1shows that the best performing morpho-syntactic errors occur more within the NIE group of learners, whereas errors of stylistic and lexical character discriminate better for the IE group compared to NIE within the same level of proficiency. From the analysis of Level A2, we can conclude that the learners' L1 can be traced from some errors which provides evidence for L1 to L2 transfer. At the same time, for other errors, it is impossible to identify their nature and group prevalence based on the data available.

Our results also suggest that a wide range of manually and automatically annotated error tags is a valuable venue to explore in the context of native language identification. Error-based features have been approached by other researchers as it is mentioned in Section 2. Koppel et al. (2005) use 185 error types, which do not appear to contribute significantly to the performance. Wong and Dras (2009) use only three features which do not improve the overall results, and do not perform as high as our error-based features on their own. Kochmar (2011) provides a very systematic error analysis and conducts a number of two-group classifications. Her results demonstrate that using character quad-grams achieves the highest precision of 100% for the Danish - Swedish group. However, the author points out that character quad-grams are likely to create content bias. In our case, tags are used for errors and a high result is achieved.

At the same time, our results cannot be directly compared with the studies described above for several reasons. First, we formulate our task differently – we only identify the learners' L1 language family rather than a specific language. Second, we use a language with a different and more complex morphological structure which might have caused a large amount of learner errors and thus, provided with the discriminative power of these features. Third, we use essays of an intermediate level of proficiency which might have contained more errors than the intermediate to advanced levels discussed previously (Argamon et al., 2009; Tsur et al., 2007).

As we emphasized above, we intentionally do not use lexical features such as function words, because some function words might reflect the content of the essays to a higher degree than other, e.g., pronouns or prepositions. For instance, if learners write about their daily routines they tend to use more prepositions and adverbs of time. If learners write about themselves, they tend to use personal pronouns. Argamon et al. (2009) describe function words as the most effective features for the task. However, when interpreting their results, we should keep in mind that the results might be artifacts of topical bias, since the topics were not strictly controlled in this study.

4.2 SLA Implications

We believe that our results are important for SLA. In particular, the results provide empirical evidence for the different types of errors discussed within the Error Analysis approach. We observe that some of the best performing errors occur when a learner's L1 interferes and affects the production of L2. We suggest that some of the highly discriminative spelling errors at the lower intermediate level are likely to be transfer errors for both groups, in support of the observations made by other researchers in regards to character n-grams. We also suggest that some stylistic errors are highly discriminative at the intermediate level within the IE group. At the same time, some of the errors that occur often within both IE and NIE groups might be developmental, and at this point these observations are not completely evident. Further experiments with more fine-grained error annotation and linguistic analysis might provide better insights on whether the best performing errors are of interlingual or developmental character. Overall, our results suggest that native speakers of Indo-European and non-Indo-European languages approach Czech differently, in their specific L1 background ways and make consistent types of errors across different linguistic levels, in particular lexicon and orthography, based on our data.

5 Conclusions and Future Work

We have described several experiments in which we explore various features to distinguish between two large language groups of learners, Indo-European and non-Indo-European. We have addressed non-content based features, and have shown that they work well for highly inflectional data. Exploring non-content based features is important because it provides opportunities for corpus independent approaches for native language identification. We have also discussed the best performing features and their contribution to the task.

Section 4.2 discussed the implications of our work to the field of SLA, but this work has further applications. By knowing what typical errors L1 learners make, language instructors can concentrate on helping their students to erase their "non-native" footprints. Other applications include marketing research, automatic error-correction and grading applications.

Our results go along with similar observations made for learner English data, that data-driven machine learning approaches are valuable for verifying SLA hypotheses (Jarvis et al., 2012). In addition, we look at Czech as the target language, which has not been discussed in the context of language background identification thus far, to the best of our knowledge. Also, our data shed light on the acquisition of target languages with complex morphology.

As for the future directions of our work, we would like to develop methods to derive best performing error tags automatically. Further, we would like to perform experiments with larger sets of data and to compare the performance of features for other levels of proficiency. Ultimately, we would like to develop a method that will be able to make more fine-grained distinctions between learners' language backgrounds using non-content based features and pin down the actual native language of the learner based on this type of data.

Acknowledgements

We would like to thank the native speakers of Czech for their participation in our experiment and to Jan Štěpánek for tailoring his questionnaire system to our needs. We would also like to thank Jing Peng and the anonymous reviewers for their comments.

This material is based in part upon work supported by the Grant Agency of the Czech Republic P406/10/P328 and National Science Foundation under Grant Numbers 0916280 and 1048406. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Shlomo Argamon, Moshe Koppel, James W.Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52: 119-123.

Stephen P. Corder. 1975. Error analysis, interlanguage and second language acquisition. *Language Teaching*, 8:201-218.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford. 2007. TAT: An author profiling tool with application to Arabic emails. In *Proceedings of the 5th Australasian Language Technology Workshop*, pages 21-30.

Sylviane Granger. 2003. The International Corpus of Learner English: A new resource for foreign language learning and teaching second language acquisition research. *TESOL Quarterly*, 37:538-546

Jirka Hana, Alexandr Rosen, Svatava Škodová, and Barbora Štindlová. 2010. Error-tagged Learner Corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV),* Uppsala, pages 11-20.

Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*, page 199-207.

Scott Jarvis and Scott A. Crossley, editors. 2012. *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach.* Multilingual Matters, UK, US, Canada.

Tomáš Jelínek, Barbora Štindlová, Alexandr Rosen, and Jirka Hana. 2012. Combining manual and automatic annotation of a learner corpus. *Text, Speech and Dialogue Lecture Notes in Computer Science*, 7499:127-134.

Thorsten Joachims. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In *Machine Learning: ECML - 98*, pages 137-142.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In Bernhard Schoelkopf and Christopher Burges and Alexander Smola, editors. *Advances in Kernel Methods – Support Vector Learning*, pages 169-185. MIT Press.

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

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 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 and Data Mining*, Chicago, IL, pages 624-628.

Judith F. Kroll, Erica Michael, Natasha Tokowicz, and Robert Dufour. 2002. The development of lexical fluency in a second language. *Second Language Research*, 18:137-171.

Robert Lado. 1957. *Linguistics across Cultures: Applied Linguistics for Language Teachers*. University of Michigan Press, Ann Arbor, MI, US.

Christopher D. Manning and Hinrich Schuetze. 1999. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, US.

F. Mosteller and D.L. Wallace. 1984. *Applied Bayesian and Classical Inference in the Case of the Federalist Papers* (2nd edition). Springer Verlag, New York.

Jack C. Richards. 1971. A non-contrastive approach to error analysis. *ELT Journal*, 25:204-219.

Alexandr Rosen, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2013. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, 47:1-28.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35: 193-214.

Štindlová Barbora and Alexandr Rosen. 2012. Návod k anotaci chybového korpusu [Learner Corpus Annotation Manual]. Unpublished. http://utkl.ff.cuni.cz/~rosen/public/anotace.pdf

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*, pages 9-16.

Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Tools and Techniques*. 3d Edition, Morgan Kaufmann, San Francisco, 2011.

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