Identification of a Writer's Native Language by Error Analysis

Ekaterina Kochmar

University of Cambridge
Computer Laboratory
St. John's College

June 2011

This dissertation is submitted for
the degree of Master of Philosophy
in Advanced Computer Science

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation does not exceed the regulation length of 15,000 words, including tables and footnotes.

Word count: 14770

Signed

17 June 2011

Identification of a Writer's Native Language by Error Analysis

Ekaterina Kochmar

Summary

In this project, we investigate the task of native language identification. We study a set of Indo-European languages, and demonstrate how machine learning techniques can be used to identify native language of a text's author.

A number of different features are extracted and applied to this task. Their contribution to overall performance is investigated and reported.

We explore the hypotheses that the choice of words in a free text is influenced by a writer's native language, and that the errors committed by a writer are based on the differences between the writer's native language system and that of English. We identify the error types typical for speakers of different native languages, and show how using different features based on the discriminative error types can improve classification.

Acknowledgments

I would like to thank my supervisor, Prof. Ted Briscoe, for his guidance and constant support. I am grateful for his encouragement and valuable suggestions throughout the course of this work.

I would also like to thank Helen Yannakoudakis and Øistein Andersen for their much appreciated help and their ability to identify the weak spots in my work and offer suggestions for improvement.

Contents

Chapter 1

Introduction

1.1 Project Motivation

English has become the most widely used language in science and business. Given that there are currently about 7000 living languages in the world[1], for the majority of people it is not a native language. Foreign languages are hard to fully master, and communicating in English poses a number of problems for non-native speakers. The way people use English, and the errors and idiosyncrasies of the variety of English they speak are highly influenced by their mother tongue. This is most evident with respect to spoken language: speakers of different native languages speak English with different accents. Speech provides a number of acoustic cues for identifying the speaker's native language, so that we usually can relatively easily tell if somebody is not a native speaker, or even where they come from. The question is whether we can identify a writer's native language in any similar way, i.e. if there are any cues in texts that help identification.

It turns out that written text also provides a number of reliable cues for native language identification such as grammar and spelling idiosyncrasies typical of non-native English. Consider the following phrase extracted from an essay written by a non-native speaker of English: *interpretate a short comedy*[2]. To tell that the author's native language in that case is Italian, it suffices to know that the Italian verb *interpretare* stands for 'play, perform'. Thus, this error can be explained by *transfer* of linguistic knowledge from one's native language to English.

The less proficient in English the learners are, the more they rely on linguistic phenomena of their native languages when speaking and writing in English. Having limited access to the lexical terms and idioms, and lacking intuitions about semantic differences between synonyms, non-native speakers often misuse English words and construct grammatically incorrect phrases. Phonological features and spelling conventions of one's mother tongue

[1] http://www.ethnologue.com/

[2] This example is taken from the Cambridge Learner Corpus.

manifest themselves in systematic spelling errors. Differences in the grammatical systems of the languages result in certain types of grammatical errors committed by the learners. Certain word order characteristics of a native language affect the way learners construct phrases in English. Manifestation of these native language specific characteristics in the use of English depends on two factors. Firstly, the more similar a language system is to English, the less unnatural the non-native use of English is. Secondly, learners at higher levels of language proficiency manifest fewer native language specific properties. Advanced learners of English rely on their native language systems much less than intermediate learners. At higher levels of proficiency all learners, irrespective of their language background, face similar difficulties in learning English related to the complexity of the English system itself (Richards, 1971).

An assumption that the characteristics of a text's author can be inferred from the text itself underlies any research in the field of *authorship profiling*, and, in particular, *native language identification*, which has a number of potential applications.

First of all, there are situations where we get an anonymous text and would like to know more about its authors. An obvious example is phishing and spam e-mails, in which case only textual data is available. Identification of the authors' native language along with other characteristics would help tracking them down. Information of this kind is useful not only for forensics, but also for any user-targeted services.

Another outcome of this type of research is identification of non-native English idiosyncrasies and distribution of errors among speakers of different languages. This information could be used in different natural language processing (NLP) tasks such as speech recognition, part-of-speech tagging and parsing, as suggested by Tomokiyo and Jones (2001). NLP tools are usually trained on native English data, which means that they are less robust when applied to non-native English. Such tools are often used as preprocessing steps in many areas, and the final results depend on their accuracy. Thus, they would benefit from language specific characteristics.

Error correction systems have also been shown to benefit from error models based on the user's native language (Lee, 2009; Gamon et al., 2008). Grammar and spell checkers are among the applications that could be improved with the use of native language profiles: these tools can focus on the types of errors typical for speakers of a particular native language and identify them with higher accuracy.

In the field of second language acquisition (SLA), native language profiling can be used to highlight the most problematic areas for language learners and design interactive learning systems (Lee, 2009; de Felice, 2008). This information could also be taken into account when designing English proficiency tests. These tests differ with respect to the level of proficiency, but not with respect to the native language. Since speakers of different languages face different problems when learning English, tailoring the tests to different language backgrounds would make them more appropriate.

Finally, profiling language learners with respect to their mother tongues could also shed light on the process of second language acquisition, which is of interest to applied linguists and cognitive scientists.

1.2 Previous Work

1.2.1 Authorship Profiling

Identification of an author's native language is a type of authorship attribution problem that has been investigated to a considerable degree. In the case of authorship attribution, researchers aim at finding a set of features that are relatively invariant for an author or group of authors across different topics, but vary from one group or author to another. Such a set of discriminative features can then be applied to identify the likely author of a text. In order to find author-specific characteristics, information of different types has been explored, including complexity-based features such as average type/token ratio, sentence and word length, sentence level features such as part-of-speech (PoS) n-grams, syntactic rules, distribution of function words and punctuation marks. Function words such as conjunctions, prepositions, and auxiliary verbs are considered to be good indicators of an author's writing style, as they are context-independent and unlikely to be biased towards specific topics. Word tokens and token-based n-grams, on the other hand, are more topic-specific and are useful for topic classification rather than author attribution. Therefore, they are rarely used for this task. Character n-grams, however, have been proven to be useful (McCombe, 2002; Stamatatos et al., 2001). Koppel and Schler (2003) stressed the importance of finding instances of author-specific idiosyncratic usage that could serve as a unique fingerprint of the author. These instances might include particular neologisms, foreign or unusual words as well as spelling and grammatical errors.

Tasks closely related to authorship attribution include gender categorisation (Argamon et al., 2003; Corney et al., 2002; Koppel et al., 2002) and demographic and psychometric traits analysis (Estival et al., 2007). The choice of features varies with the problem, but many of the feature types used for authorship attribution are used for related tasks as well. Koppel et al. (2002) showed that automatic text analysis based on function words and PoS n-grams can identify an anonymous author's gender with an accuracy of approximately 80%.

1.2.2 Native Language Identification

The problem of native language identification has been addressed by a number of researchers.

In Estival et al. (2007), a set of English e-mails was collected and native language was considered as one of the 10 demographic and psychometric characteristics of anonymous authors. This study was not focused on native language identification, and the data set contained e-mails by speakers of only 3 languages, English, Arabic and Spanish, with 62.90% of the texts written by English speakers. A number of different machine learning algorithms were applied to find the best classifier for each of the characteristics. An accuracy of 84.22% was obtained on this three-class problem using a random forests classifier with a set of character-based, lexical and e-mail structure-specific features selected using Information Gain.

Tomokiyo and Jones (2001), also performed classification with respect to three native languages: Chinese, Japanese and English. They collected and transcribed texts from 31 Japanese, 8 English and 6 Chinese speakers. To avoid topic influence on the classification, they used token-based n-grams with nouns being replaced with their PoS tags. They obtained an accuracy of 89% to 100% on two- and three-way decisions distinguishing between native and non-native English, Chinese and Japanese non-native English, and between non-native Chinese, Japanese and native English. A Naive Bayes classifier with a multinomial event model was used.

Although these two studies show promising results, direct comparison is not possible due to the unavailability of the data sets used in their experiments. A number of other researchers have used the International Corpus of Learner English (ICLE)[3] which contains essays written by intermediate to advanced learners of English.

The first piece of research in the area to use this data set was by Koppel et al. (2005b). Replicating their earlier work on authorship attribution (Koppel and Schler, 2003), they used a set of function words and character n-grams as features, in addition to 185 error types, including misspellings and syntactic errors. As has been noted in Koppel et al. (2005b), language idiosyncrasies are most discriminative for authorship attribution. Similarly, spelling and syntactic errors reflecting certain orthographic and syntactic conventions from the authors' native languages serve as strong discriminative features for native language identification. Futhermore, 250 rare PoS bigrams were extracted from the Brown corpus as instances of non-standard English. Using this feature set, they applied multi-class classification with support vector machines (SVM) to Bulgarian, Czech, Russian, French and Spanish texts extracted from the ICLE. The best accuracy on this data set is 80.2% obtained using all the features. This is significantly better than the majority baseline of 20%. In their experiments, use of error types in addition to other features always improved the accuracy, in some cases by 5 percentage points[4].

Based on the results of Koppel et al. (2005b), Tsur and Rappoport (2007) formed the

[3]http://www.uclouvain.be/en-cecl-icle.html

[4]Koppel et al. (2005b) did not report the results explicitly; they can only be roughly estimated from their graph.

hypothesis that the choice of words people make when writing in a foreign language is strongly influenced by the phonology of their native language. To test this hypothesis, the researchers performed multi-class classification with SVM on the same data but using only character n-grams. With character unigrams, they obtained an accuracy of 46.78% which is more than twice as high as the baseline of 20%. However, as character bigrams are closer to the language phonology, an accuracy of 65.60% was obtained using a list of the 200 most frequent bigrams. This agreed with their hypothesis of the motivated choice of words.

Finally, Wong and Dras (2009) based their approach on the *contrastive analysis* hypothesis, according to which errors committed by learners can be explained by the differences between English and the learners' native language. The three most common types of such errors in non-native English are subject–verb disagreement (*The information are very detailed*), noun number disagreement (*Both of the company's name were written in Chinese*), and misuse of determiners (*I would like to know how much a membership costs*). In Wong and Dras (2009) these syntactic error types were explored. They ran ANOVA tests and concluded that misuse of determiners is highly statistically significant for all the examined languages, while the other two types of errors are not statistically significant, even though some languages lack such linguistic phenomena as subject–verb or noun number agreement. Wong and Dras (2009) ran experiments on the same set of five languages, as well as on Chinese and Japanese, using an SVM classifier. They obtained an accuracy of 24.57% for the multi-class classification with the majority baseline of 14.29% which is a significant improvement at the 95% confidence level. Finally, they combined the syntactic error features with lexical features such as function words, character n-grams and PoS n-grams. The best accuracy was 73.71% using a combination of all the lexical features, while adding syntactic error features did not improve this result.

The three studies performed on the ICLE data explored different types of features, and their results are directly comparable. However, none of the previous researches has systematically studied contribution of different feature types. Furthermore, none of the studies has explored classification over a set of native languages in any systematic way. Our work is aimed at filling these gaps.

1.2.3 Contrastive Analysis

In Koppel and Schler (2003) and Wong and Dras (2009), promising results were obtained when using error types. Their approach is based on the *contrastive analysis* hypothesis from the field of second language acquisition, first formulated by Lado (1957). According to this hypothesis, difficulties in acquiring a foreign language are caused by the differences between the foreign language and the learner's native language. In the process of foreign language acquisition, there occurs *language transfer*: the characteristics of the native language are carried over into the foreign language. This transfer can be positive if

certain phenomena in two languages coincide. Then it helps acquiring the new language. However, when the two systems differ, the transfer is negative and it causes errors related to these differences. The more dissimilar the two systems are, the more transfer errors could be committed and the more difficult it is to learn English for speakers of such languages. This also means that by analysing the native language it is possible to predict such difficulties in advance.

Richards (1971) suggested that the language transfer or *interlanguage* errors should be distinguished from *intralingual* and *developmental* errors caused by the difficulty of English itself and independent of the language background. Thus, transfer errors are potentially powerful in distinguishing between different native languages, but an important step in any research that makes use of them is to carefully identify such errors and not to confuse them with other error types.

One approach to identification of the interlanguage errors is to consult linguistics books and try to identify the differences between the language systems. A widely used example of such differences is the use of articles: there are no articles in Russian in contrast to Germanic languages like English and German. This means that native speakers of Germanic languages have certain intuitions about how the articles operate. As a result, learning how to correctly use articles in English would not pose a serious problem for native speakers of German. On the other hand, Russian speakers lack a basic knowledge about how articles should be used. This difference between the language systems constitutes a problem for Russian learners, and this is a topic that language instructors should pay special attention to. Furthermore, a high percentage of article-related errors in a text is a reliable cue that the native language of the author does not utilise articles.

However, not all language system differences are easy to identify and not all of them are significant. For example, two languages may have different verbal systems. This difference may seem important. As a result, language instructors may pay much attention to the English verbal system, and test designers may give lots of examples on this topic. However, by just comparing language systems, it is hard to predict whether the difference would cause significant problems. This is a disadvantage of any knowledge-based approach: linguistic evidence may be not sufficient.

We undertake a data-driven approach instead. In this research, we rely on the evidence extracted from a big corpus of learner English (see Section 2.1). The advantage of a data-driven approach is that it allows us to explore a bigger data set, and test a number of hypotheses. The major problematic areas and the typical errors for language learners can be reliably identified.

1.3 Project Goals and Objectives

We study a problem of authorship profiling with respect to the native language. We address it as a classification problem with the native languages being the classes. Our hypothesis is that machine learning techniques can be effectively applied to the problem.

For classification purposes, a set of appropriate features should be found. The source of the features in this case is the textual data. Previous studies (see Section 1.2) show that native language can be identified relatively accurately, and some of the results suggest that there is room for improvement. For example, error analysis is one of the promising directions of study. In Wong and Dras (2009), only a limited number of error types have been considered, and the results suggest that the error types for classification should be selected more carefully. None of the previous studies has explored the feature space systematically, and the different features' contribution has not been reported. Moreover, a systematic study of native languages has not been undertaken in any of these studies.

Therefore, the main goals of this project are:

- On the theoretical side:
  - explore the feature space for this classification task. Measure and report contribution of different features. Evaluate the importance of different features for identifying different native languages.
  - Perform language identification on a wide set of languages.
- On the practical side: build a native language identification tool that would rely on a set of discriminative features, and, using machine learning techniques, would be able to identify a writer's native language reliably.

The data used in this project (see Section 2.1) allows us to experiment on a wide set of closely related languages. Thus, a set of native languages can be explored in a systematic way, which has not been done previously. Moreover, the data availability allows us to test a number of hypotheses in the course of this project:

- Similar language background is shared not only by speakers of the same native language, but also by speakers of languages within one language group. Identification of the native language group or family is a broader class problem in this context. The accuracy of identifying a particular language, its group, or family will be explored.
- A number of closely related languages are considered in this study. One of the goals is to estimate how accurately different pairs of languages can be distinguished, and how this accuracy changes depending on how similar the languages are.

- Errors are considered to be strong indicators of the native language. Classification based on the error types will be opposed to the distributional analysis, which considers linguistic properties in general. These two approaches will be compared and the differences will be evaluated.

- Finally, if errors are indeed strong native language indicators, it would be possible to make native language error profiles that could be used for native language classification, and for other language-specific tasks.

Chapter 2

Data and Resources

In this chapter, we present the data and the resources used in our project.

2.1 Data

The Cambridge Learner Corpus[1] (CLC) is a large corpus of learner English. It has been developed by Cambridge University Press in collaboration with Cambridge Assessment, and contains examination scripts from learners of English from 86 native language backgrounds. The scripts have been produced by language learners taking Cambridge Assessment's English as a Second Language (ESOL) examinations[2]. In this project, we use a set of texts produced by learners sitting the First Certificate in English (FCE) examination[3]. This examination assesses English at an upper-intermediate level, which suggests that the learners sitting this exam still manifest a number of transfer errors.

Each text has been written in response to two tasks asking learners to write a letter, a report, an article or a short story. These texts are 200 to 400 words long. Typical prompts are presented below:

A meeting has been arranged to discuss ways of making your town, or village, more attractive. You know what is wrong with your town or village and you have some ideas about how to improve the look of it. Write what you say.

and:

Describe the most embarrassing moment of your life.

The scripts have been anonymised and annotated using XML. In addition, the linguistic errors committed by the learners have been manually annotated using a taxonomy of 80

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

[2] http://www.cambridgeesol.org/

[3] http://www.cambridgeesol.org/exams/general-english/fce.html

error codes Nicholls (2003). Each error has been manually identified and tagged with an appropriate code, specifying the error type and a suggested correction. The following is an example of an error-coded sentence:

I would like to be your guide <NS type="RT"> *on|during* </NS> *these days.*

where *RT* denotes a "replace preposition" error.

Finally, the scripts are also linked to meta-data about the exam and the learner. This includes the year of examination, the question prompts, the learner's native language, nationality, age, sex as well as grades obtained. This facilitates examination of the data with respect to different characteristics, such as the learners' native language.

2.2 Language Data Sets

In this project, we perform a systematic study on a set of closely related Indo-European languages. For this purpose, texts produced by speakers of 5 *Germanic* (German, Swiss German, Dutch, Swedish and Danish) and 5 *Romance* languages (French, Italian, Catalan, Spanish and Portuguese) are collected. As has been noted earlier (see Section 1.3), classification performed on such a set would show to what extent two or more closely related native languages can be distinguished from each other. Given that English itself belongs to the group of Germanic languages, the task becomes more challenging, as speakers of these languages might find writing in English easier than speakers of other languages and therefore commit fewer errors.

To perform a systematic study, one needs to first identify relations between the Indo-European languages. Computational construction of language taxonomies has been studied in a number of papers.

In Ellison and Kirby (2006), the similarities between the word forms of the language defined in terms of their confusion probabilities were considered. This idea strongly correlates to psycholinguistic models of word cognition. Kullback-Liebler divergence and Rao distance were then used to measure the distance between languages on the basis of their confusion probability matrices. Figure 2.1 shows a tree that corresponds to the results of Ellison and Kirby (2006).

In Pagel (2009), parallels between biological and linguistic evolution were established, and statistical methods from phylogenetics and comparative biology were applied to study language evolution. As opposed to Ellison and Kirby (2006), who studied the inner-language cognate forms, Pagel (2009) worked with the Swadesh list (Swadesh, 1952) of 200 words, known as the fundamental vocabulary of the world's languages, to identify inter-language word cognates, i.e. words derived from a common ancestral word. The rates of lexical evolution for the meanings of the words in this fundamental vocabulary were estimated using a statistical likelihood model of word evolution. For each of the 200

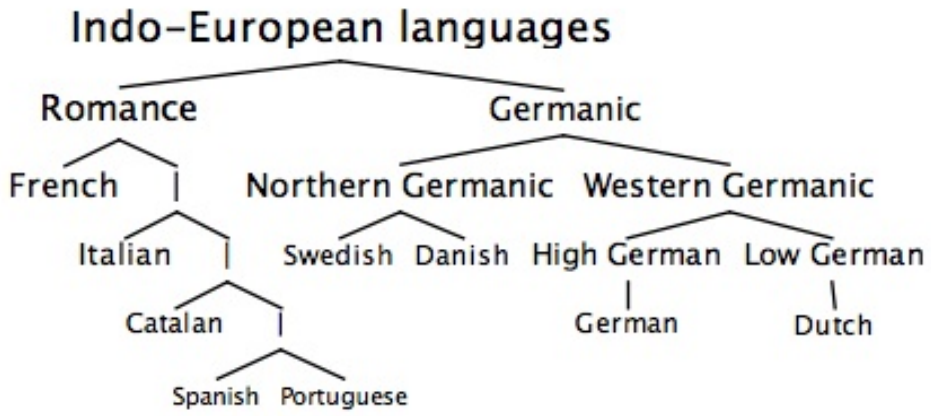

Figure 2.1: Tree representation of the Indo-European languages based on Ellison and Kirby (2006).

meanings, the mean of the posterior distribution of rates derived from a Bayesian Markov chain Monte Carlo (MCMC) model was calculated and scaled to represent the expected number of cognate replacements per 10,000 years. Figure 2.2 represents a part of the resulting tree containing the given set of languages.

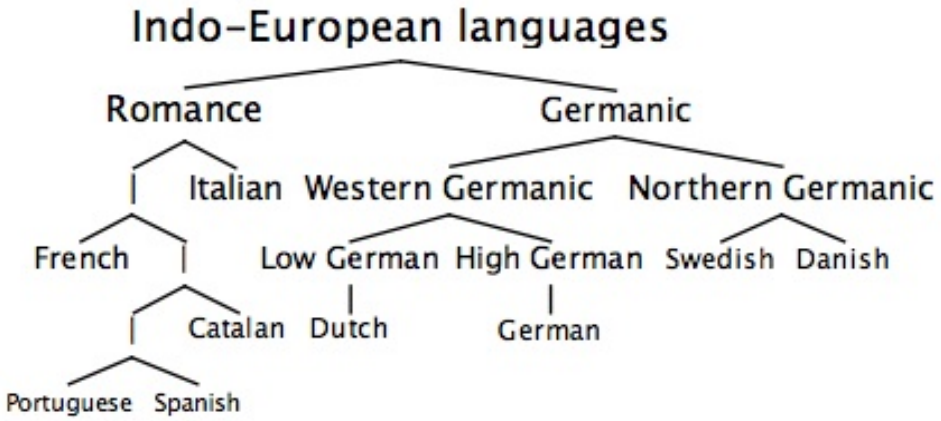

Figure 2.2: Tree representation of the Indo-European languages based on Pagel (2009).

While the results of these studies correlate to a considerable degree, in particular in what relates to the *Germanic* languages, slightly different classifications are provided for the *Romance* languages. To examine the CLC data, we applied an agglomerative clustering algorithm using different sets of features, including word and character n-grams of different lengths, and misspellings. This method showed yet another grouping of the *Romance* languages. A sample clustering is shown in Figure 2.3. Clustering the data using different sets of features shows that *Catalan* and *Spanish* most often fall in one cluster, with *Portuguese* usually being placed very close to them.

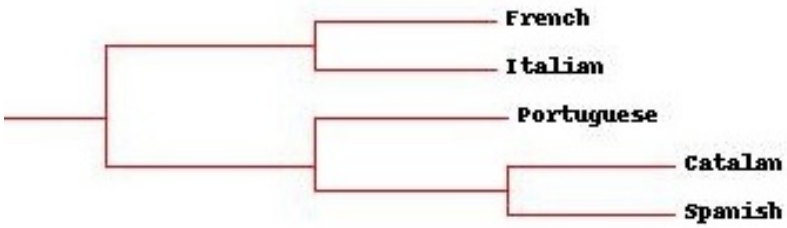

Figure 2.3: Clustering with misspelled English character quadgrams.

This shows, that while classification of *Germanic* languages is mostly unquestioned, that of *Romance* languages is more controversial. Since the classification of *Germanic* languages reveals a number of binary oppositions, we apply binary classification to it. Our data also contains texts written by speakers of *Swiss German* language, and we consider it as being opposed to the standard *German* language. The classification of *Romance* languages does not allow for binary classification. It shows that multi-class classification would be more appropriate for this set of languages. Therefore, in this project we use a flatter representation of the *Romance* data (see Figure 2.4). However, to test how binary classification would work on this data, we also consider a binary opposition of *Spanish* and *Catalan* languages, since they proved to be close when applying clustering algorithm.

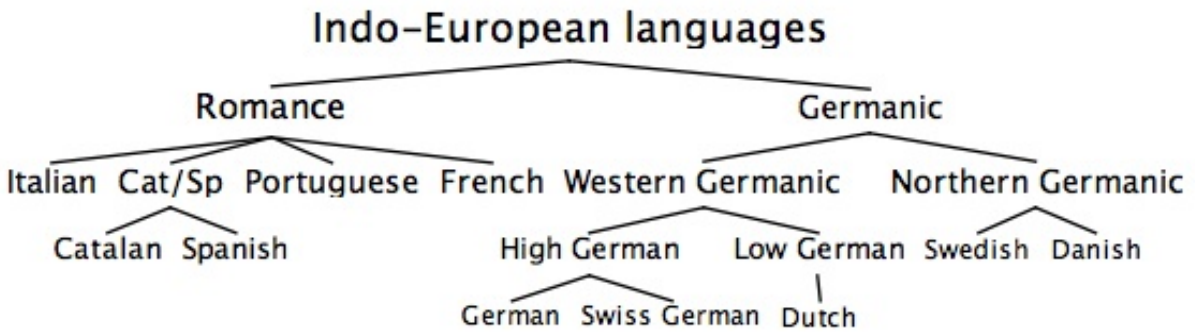

Figure 2.4: Tree representation of the Indo-European languages.

The focus of this study is on binary oppositions between Indo-European languages and language groups, namely *Germanic – Romance*, *Western Germanic – Northern Germanic*, *High German – Low German*, *Swedish – Danish*, *German – Swiss German*, and *Catalan – Spanish*. Multi-class classification of the *Romance* languages is an interesting topic on its own, and it is left for future research.

For every language pair, training and test sets are created with uniform distribution of classes within the sets. To estimate the classifier performance, 5-fold cross-validation is applied. In each run, the data is divided into training and test sets in proportion of 80% to 20%. The data sets are described in more detail below. It is also shown, how diverse with respect to the speakers' nationality the data sets are.

German vs. Swiss German

The *German – Swiss German* data set consists of 51 texts per language, and contains all the available texts for *Swiss German*, and 51 randomly chosen texts for *German*.

`Geographic profile:`

- The *German* subset contains texts written by speakers from Germany (88%), and texts written by speakers from other countries, for example, Austria, Liechtenstein, or Italy.
- In the *Swiss German* subset, all the texts are written by speakers from Switzerland.

High German vs. Low German

The *High German – Low German* data set consists of 102 texts per language group. It includes all the texts used for the *German – Swiss German* classification, and 102 of the *Dutch* examination scripts chosen randomly.

`Geographic profile:`

- The *High German* subset contains texts written by speakers from Switzerland (49%), Germany (44%), and some other countries.
- The *Dutch* subset contains texts written by speakers from the Netherlands (59%), Germany (25%), Belgium (5%), Switzerland (3%), and some other countries, including Italy, Denmark, Turkey.

Danish vs. Swedish

The *Danish – Swedish* data set consists of 62 texts per language, and includes all of the available examination scripts for *Swedish* and 62 scripts for *Danish* chosen randomly.

`Geographic profile:`

- The *Danish* subset includes texts written by speakers from Denmark (63%), France (27%), Iceland (3%), and other countries such as Switzerland, Belgium, and Mauritius.
- Almost all of the texts in the *Swedish* subset are written by speakers from Sweden (94%), and the rest – by speakers from Germany and France.

Western Germanic vs. Northern Germanic

The *Western Germanic – Northern Germanic* data set contains 124 texts from the *Danish – Swedish* group, with 62 texts per language, and 126 texts from the *German – Swiss German – Dutch* group, with 42 texts per language. Hence, a balance in the groups is kept, and no bias towards any of the language groups is introduced.

`Geographic profile:`

- The *Western Germanic* subset contains examination scripts written by speakers from Germany (38%), Switzerland (33%), Netherlands (21%), Belgium (2%), and some other countries.
- The *Northern Germanic* data set contains texts written by speakers from Sweden (47%), Denmark (31%), France (14%), Iceland (2%), and other countries, including Germany, Switzerland, Belgium, and Mauritius.

Spanish vs. Catalan

The *Spanish – Catalan* data set contains 125 randomly chosen examination scripts per language.

`Geographic profile:`

- The *Spanish* subset includes scripts produced by speakers from Spain (55%), Chile (17%), Argentina (16%), Mexico (10%), and some other countries.
- Almost all of the texts in the *Catalan* subset are written by speakers from Spain (96%), and others are produced by speakers from other countries, such as Andorra, or Argentina.

Romance vs. Germanic

The *Romance – Germanic* data set contains 250 randomly chosen texts per group: the previously used 62 texts per language for *Danish – Swedish*, 42 texts per language for *German – Swiss German – Dutch*, and 50 scripts per language for the *Romance* group. We avoid introducing any language group bias by choosing an equal number of texts per language set.

`Geographic profile:`

- The *Romance* subset contains texts written by speakers from Spain (30%), Italy (18%), France (12%), Brazil (12%), Switzerland (9%), Portugal (8%), Chile (4%), Argentina (3%), Mexico (2%), and some other countries.

- The *Germanic* subset includes texts written by speakers from Sweden (23%), Germany (20%), Switzerland (17%), Denmark (16%), Netherlands (10%), France (7%), Belgium (2%), Iceland (1%), and some other countries.

2.3 Tools

We treat native language identification as a machine learning problem. In this study, Support Vector Machines (Vapnik, 1995; Joachims, 1998, 2002) are used through the SVM^{light} package[4].

All the texts that we use in our project have been parsed using the Robust Accurate Statistical Parsing (RASP) system with the standard tokenisation and sentence boundary detection modules (Briscoe et al., 2006). This allows us to extract many relevant and linguistically motivated features such as PoS n-grams or phrase-structure rules (see Section 3.2).

To automatically extract and weight these features, a system used in Yannakoudakis et al. (2011) for a related task of automatically grading 'English as Second Language' (ESOL) examination scripts is reused and extended in this project, since the data and the feature sets used for these tasks coincide to a considerable degree.

In the next chapter, the approach to this learning problem and the features used are explained in detail.

[4]http://svmlight.joachims.org/

Chapter 3

Approach

The methodology of classifying texts with support vector machines is briefly outlined in Section 3.1. We adopt two different approaches to our classification problem. They are presented in Sections 3.2 and 3.3.

3.1 Classification with Support Vector Machines

Machine learning methods have been widely used for text classification, and, in particular, for native language identification (see Section 1.2). The necessity of applying machine learning methods stems from the fact that there is no mathematical model available for text classification: the correspondence between the input data and the output classes is not known in advance and cannot be easily derived. Thus, we rely on the assumption that computers can learn this model from the data. In particular, it is assumed that a computer can learn the input/output functionality from a given set of examples.

This process resembles how people learn to distinguish between different phenomena, for example, between different foreign accents. It is hard to define precise characteristics of a particular accent, and this is hardly the way we learn to identify it. Rather, having heard a number of people speaking with this accent and being told they are all native speakers of some language L – that is, being presented with a set of labelled examples – we are able to identify this accent next time we hear it.

In machine learning, this process is referred to as *supervised learning*, and the examples are called *training data*. A learning machine is given a training set of labelled examples and converts them to *attribute vectors* of dimensionality n.

The function that should be learned from the data is called the *decision function* and it represents the solution for the classification task. The candidate decision functions offering solution to the classification problem are called *hypotheses* and together they form a *hypothesis space*. Once the attribute vectors are available, a number of hypotheses

could be chosen for the problem. Among these hypotheses, *linear functions* are the best understood and the simplest to apply.

Support Vector Machines (SVMs) are learning systems that use a hypothesis space of linear functions in a high-dimensional feature space (Christianini and Shawe-Taylor, 2000). They were originally developed by Vapnik (Vapnik, 1995). This learning algorithm is chosen for the current task for a number of reasons. SVMs are effective in high-dimensional spaces and for large amounts of data, and proved to be highly efficient in text classification tasks. Moreover, SVMs have been successfully applied in a number of previous studies on this topic (see Section 1.2).

Classification is defined as n-ary by the number of classes to be distinguished. In this project, a binary learning problem is considered, i.e. the performed classification is *binary*. In a more general case, the classification problem involves multiple classes. In Section 2.1, we briefly discussed that classification of *Romance* languages could be treated as a multiple-class problem.

Let $X \subseteq \mathbb{R}^n$ denote the input space, and Y be the output domain. In the binary case, Y consists of values $Y = \{-1, 1\}$ for the two classes. The training data contains examples denoted by $S = \Big((\mathbf{x}_1, y_1), ..., (\mathbf{x}_l, y_l)\Big) \subseteq (X \times Y)^l$, where l is the number of examples in the training set, $\mathbf{x}_i \in X$ are the examples and $y_i \in Y$ are their labels. In the binary case, if a real-valued function f: $X \to \mathbb{R}$ gets a value $f(\mathbf{x}) \geq 0$ for the instance $\mathbf{x} = (x_1, ..., x_n)$, it is assigned to the positive class $y = 1$, and otherwise to the negative class $y = -1$. A *linear function* $f(\mathbf{x})$ is defined as follows:

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b = \sum_{j=1}^{n} w_j x_j + b \tag{3.1}$$

where $(\mathbf{w}, b) \in \mathbb{R}^n \times \mathbb{R}$ are the parameters of the function to be learned from the data. The decision rule is defined by the sign of the function, $sgn(f(\mathbf{x}))$.

Figure 3.1 shows linearly separable data for some classification problem. The input space X is split into two parts, each part containing inputs from the two corresponding classes. The two parts are divided by a *hyperplane*, which is defined by the equation $\langle \mathbf{w} \cdot \mathbf{x} \rangle = 0$. The vector $\mathbf{w}$ defines a direction perpendicular to the hyperplane, and the value of b moves the hyperplane along the axis defined by $\mathbf{w}$.

An SVM seeks to find the hyperplane that separates the two classes most cleanly, that is with the greatest possible distance to the nearest data points. These data points are called *support vectors*, and the distance is referred to as the *margin* of the classifier. Figure 3.2 shows a maximal margin hyperplane dividing linearly separable data, and the support vectors are highlighted.

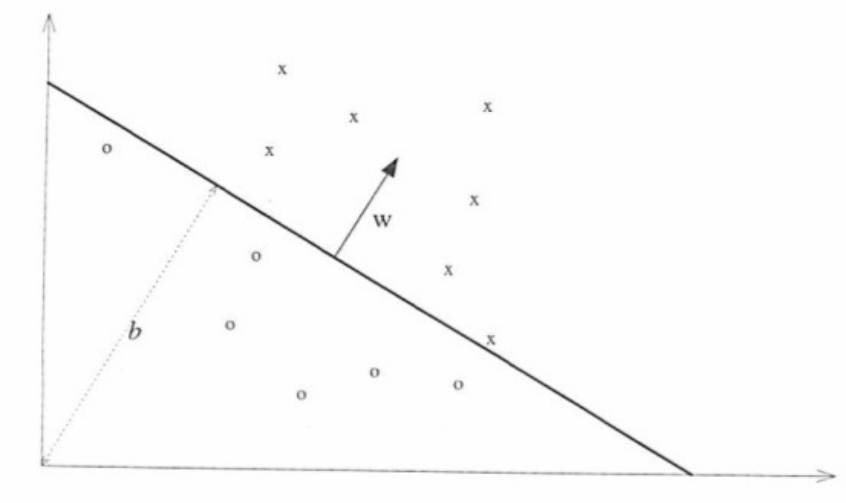

Figure 3.1: A separating hyperplane (**w**, b) for a two dimensional training set (Christianini and Shawe-Taylor, 2000).

Given a linearly separable example $S = \Big((\mathbf{x}_1, y_1), ..., (\mathbf{x}_l, y_l)\Big)$, the hyperplane maximising the margin is realised by the hyperplane $(\mathbf{w}, b)$ that solves the optimisation problem:

$$\begin{aligned} &\text{minimise } _{\mathbf{w},b}: && \langle \mathbf{w} \cdot \mathbf{w} \rangle, \\ &\text{subject to constraints:} && y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1, \\ & && i = 1, ..., l \end{aligned}$$

The performance of the classifier is evaluated by its ability to correctly classify unseen data that is not contained in the training set. This ability, referred to as *generalisation*, is the property of the algorithm that should be optimised.

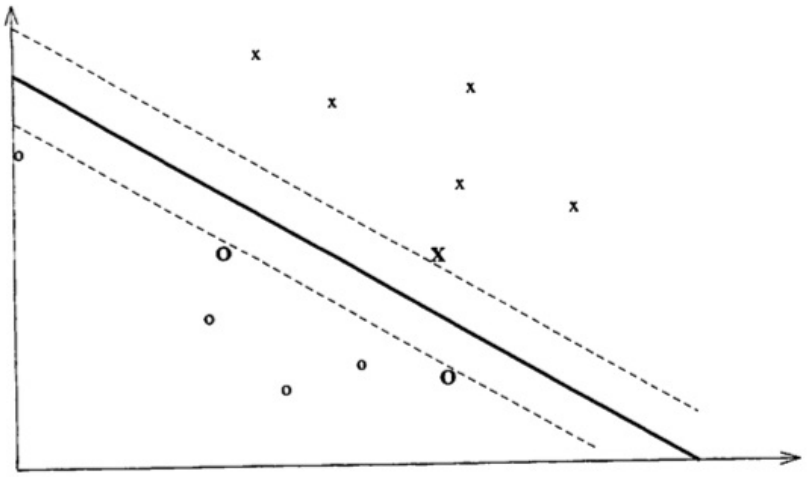

Figure 3.2: A maximal margin hyperplane with its support vectors highlighted (Christianini and Shawe-Taylor, 2000).

In certain cases, the machine learning algorithm may produce complex hypotheses that would provide an accurate fit to the training data, but make incorrect predictions on the new, *test* data. This problem is known as *overfit*, and in order to avoid it, a trade-off between function complexity and algorithm accuracy is set. It is defined in terms of statistical bounds on the generalisation error. These bounds typically depend on certain quantities, in particular, on the margin of the classifier.

To facilitate learning of the decision function by the classifier, the data should be represented appropriately with the most relevant set of *features*. The task of choosing the most suitable data representation is called *feature selection*. There are different approaches to this task.

Typically, one starts by looking for the smallest set of features that conveys the essential information contained in the original attributes. This process is known as *dimensionality reduction*. Since both computational and generalisation performance can degrade as the number of features increases, the learning algorithm can benefit from dimensionality reduction. Next, one seeks to identify and eliminate *irrelevant features*.

For example, in text classification with respect to topics, it is usual not to take function words like articles or prepositions into account, since being roughly equally distributed among different topics, they do not carry any topic-specific information as opposed to content words. The feature space of content words can be further reduced by using word stems or lemmas: if *match* and *matches* relate to *sports*, both can be represented as a single feature *lemma:match*.

In what follows, we present two approaches to the given classification task, and discuss issues related to the selection of features.

3.2 Distributional Analysis

As was stated earlier (see Section 2.1), the texts are produced by English language learners in response to questions eliciting free-text answers. The range of words and linguistic constructions depends upon the speakers, and their choice, as assumed by Tsur and Rappoport (2007), may be influenced by their native languages.

A *distributional* approach is based on an assumption that linguistic properties are distributed differently in texts produced by native speakers of different languages. As opposed to *contrastive analysis* (see Section 1.2.3) discussed in detail in the next section, a distributional approach makes no distinction between correct and incorrect English. The underlying idea is that speakers of different languages not only commit different errors but also use English differently. A distributional analysis seeks to identify specific patterns in the use of English. Selection of relevant features is the first step in finding these patterns.

3.2.1 Words as Features

Words are frequently used as features in many text classification tasks. They are, indeed, highly discriminative for some of these tasks, for example, for topic classification. However, most of the studies on author profiling and native language identification have avoided using words and word n-grams as features (Section 1.2). The reason is that words are more strongly linked to the content and the topic of the text than to any of the author's characteristics. If one author writes children's stories, and another books on computer science, text attribution using only content words would be an easy task because the sets of words used by the authors would be quite different. This, however, would not tell much about the authors' writing styles. The texts, in this case, would be attributed to the topics rather than to the authors.

The same holds for the native language identification task. The scripts used here are produced not only by speakers of different languages, but also in response to different prompts, that is, they are on different topics. For example, there is only a minor overlap of topics between *German* and *Swiss German* scripts: the *German* set contains texts on 25 topics, the *Swiss German* one on 19 topics, and only 3 topics are common for both sets.

An SVM classifier is able to distinguish between these two languages with an accuracy of 87.50% using only word unigrams and bigrams. This is a very good result for such closely related languages. To better understand how this result is obtained, let's look at the features ranked highly by the classifier.

The German set of highly discriminative features contains words like *animals* and *zoos*, *car* and *bicycle*. The prompts that were presented to the German but not the Swiss German speakers contain the following questions: *Is there a need for zoos in the modern world?* and *Which is the best means of getting to work or school – by bicycle or by car?*

The Swiss German set of discriminative features includes the following groups of word unigrams and bigrams: *mobile, mobile phone, a mobile*; *library, people read, recommend.* The list of prompts for Swiss German speakers contains such tasks as *There are both advantages and disadvantages to having a mobile phone. Write your composition*, and *'In a story, the places are often more important than the people.' How true is this of the book or one of the short stories you have read?* Neither of these prompts was presented to German speakers.

Obviously, words are strong indicators for the topics but not for the native languages. They help identifying prompts but they do not tell us anything about native language idiosyncrasies. Therefore, in this study they are not used as features.

3.2.2 Other Types of Features

The following features are used for distributional analysis:

1. `Part of speech (PoS) n-grams` with n ∈ [1, 3] represent sequences of part-of-speech tags assigned to words, and can encode word order and grammatical properties of the writer's variety of English. They might also capture idiosyncratic constructions and sequences of words not typical for English.

   PoS n-grams are extracted from the texts tagged with the RASP tagger (see Section 2.3). The tagger uses the CLAWS tagset[1]. The most probable posterior tag per word is used for the PoS n-gram features.

   The extracted sets of highly discriminative PoS features show that Romance speakers use more punctuation marks like *:*, *...* and *;* than speakers of Germanic languages. They also use modal verbs more frequently. Germanic speakers use extensive sequences of proper nouns (as *NP1_NP1_NP1*), e.g. *Frans van Righoven*.

2. `Character n-grams` with n ∈ [1, 4] can model text content to a certain degree without considering words directly. Tsur and Rappoport (2007) showed that character n-grams reveal phonological characteristics of native languages. They can also encode language-specific misspellings and non-English words. For example, the top discriminative n-gram features for Germanic languages contain such non-typical English bigram as *ko*. It is contained in such misspelled words as **ko*wledge, **ko*nclusion, and named entities like **Ko*penhagen, **Ko*ln, and **Ko*pmansgatan.

   Character n-grams as well as PoS n-grams are weighted using the *tf-idf* scheme, which takes into account both the number of times a unit appears in the text and the frequency of this unit in the whole data set. To scale feature vectors to the same order of magnitude, they are length-normalised by the L2 norm which makes them unit length vectors.

3. `Phrase structure (PS) rules` are extracted from the trees produced for the most likely parse identified by the RASP parser. These rules encode detailed information about the grammatical constructions in the sentences, and, like PoS n-grams, can capture grammatical and syntactic patterns of native language specific use of English. PS rules can also reveal idiosyncratic constructions.

   For example, a rule *T/np_leta-cl* is among the top discriminative features for Romance speakers. It states that a sentence can contain a noun phrase followed by a non-clausal text adjunct with sentence-final punctuation. This rule encodes the following peculiarity of the writing style of speakers of some *Romance* languages:

[1] http://ucrel.lancs.ac.uk/claws/

starting a letter, they delimit a person's name from the rest of the sentence with a colon as in *Dear Sir or Madam: First of all ...* or *Dear Peter: I'm sorry that ...*

This feature is weighted using frequency counts.

4. In a related project on automatically grading ESOL examination scripts (see Section 2.3), error rates were used as a feature. They encode information about language proficiency of a writer. However, the errors committed by writers also depend on how difficult it is for speakers of certain languages to acquire English. If the speakers' native language shares many linguistic properties with English, learning English is easier for them than for speakers of more 'distant' languages. For example, Germanic languages share many common properties with English, which is itself a Germanic language. Presumably, Germanic speakers commit less errors than speakers of Romance languages. On this basis, it is assumed that error rates encode the 'easiness' of learning English and the 'closeness' of a native language to English. Two error rates introduced in Yannakoudakis et al. (2011) are used here:

   - `CLC error rates` are derived from the CLC error annotation (see Section 2.1).
   - `Corpus-based error rates` do not rely on any tagged data. Instead, they are calculated with respect to a trigram language model built from ukWaC (Ferraresi et al., 2008), a large corpus of English of more than 2 billion tokens. To that, frequently occurring trigrams from a subset of highly ranked CLC scripts have been added. A word trigram in a test script is counted as erroneous if it is not contained in the language model.

   Error rate features are scaled to the same order of magnitude as the previous features.

Additionally, frequency thresholding is applied to filter out features occurring less than four times.

3.3 Error-Based Analysis

In Section 1.2.3, the contrastive analysis hypothesis was discussed. According to that, errors committed by non-native speakers of English bear certain characteristics of their native language and are strong indicators of the language background. Within this type of analysis, only features based on the errors committed by the writers are used.

3.3.1 Feature Types

To test the contrastive analysis hypothesis, the following types of features are considered:

- `Error type rates`: Following the contrastive analysis hypothesis, speakers of a language L are expected to make more errors related to the differences in the language system of L and that of English. Then, the number of errors of certain types committed by the speakers of L would be different from the number of errors of same types committed by the speakers of some other language L'. Therefore, error types measured quantitatively can be used as features in the classification task. Let n_t be the number of errors of type t in a text. To convert this count to a feature in a unified feature space, we normalise it by the length of the text.

- `Error type distribution`: Scripts are manually tagged with information about the errors committed. Let N be the total number of errors of different types in a text. Then, the error type ratio is $(\frac{n_t}{N} * 100)\%$.

- `Error content`: Earlier we used an example of an error-annotated sentence:

  I would like to be your guide <NS type="RT"> *on|during* </NS> *these days.*

  Here, code *RT* identifies an error type, while the word incorrectly used is the preposition *on*. We call a combination of the error code and the incorrect piece of text *error content*. Error content in this case is *RT:on*. To check whether speakers of different languages have particular problems with certain words or constructions, error content is used as another feature type. It is weighted using *tf-idf*.

3.3.2 Error Types

The error coding of the CLC data allows us to explore a wide range of hypotheses about the errors committed. This is a significant advantage of our data as compared to previous studies.

- `Error types used in previous studies`: In Wong and Dras (2009) errors of three types were used: those related to misuse of determiners, subject–verb disagreement and noun number disagreement. These phenomena are considered to be among the most difficult for learners of English, since they are absent from other languages, or behave differently. However, Wong and Dras (2009) concluded that only the misuse of determiners is a significant problem for non-native speakers. They also showed that addition of the three error types does not improve classification with lexical features (see Section 1.2). We replicate their experiments and compare the results.

- `Spelling errors`: Tsur and Rappoport (2007) formulated a hypothesis that the choice of words in a foreign language is strongly influenced by the phonology of the native language. Misspelled words and erroneous character n-grams contained in these words may bear characteristics of native language spelling conventions and

may prove to be reliable cues for language identification. In Tsur and Rappoport (2007), character n-grams extracted from all the words were used. However, better results may be obtained using only erroneous, or non-typical English, n-grams as features. In this project, we consider misspellings as well as erroneous character n-grams.

- `Language-specific error types`: The Indo-European languages have such phenomena as determiners, subject–verb and noun number agreement. Presumably, speakers of the Indo-European languages do not have particular problems with any of these phenomena in English. On the other hand, there may be other problematic areas, and the attention should be drawn to them. In this project, we aim at finding such problematic areas, and perform selection of language-specific error types to improve classification.

Typical Error Types

In Table 3.1, the most typical error types for the Indo-European languages are presented. A cell contains '+' if an error type is among the five most typical types for this language, and −' otherwise. The error types are extracted from the data sets using *document frequency*, i.e. selecting errors that occur in the largest number of texts written by the speakers of particular language.

The following notation is used for the languages:

`G` – German	`Sw` – Swiss German
`HiG` – High German	`LoG` – Low German
`Dan` – Danish	`Swe` – Swedish
`W` – Western Germanic	`N` – Northern Germanic
`Sp` – Spanish	`Cat` – Catalan
`Rom` – Romance	`Ger` – Germanic

`Punctuation` appears to be a problematic issue for English language learners. They often miss commas after introductory words and phrases like *in my opinion*, *in principle*, *however*, and between clauses within a sentence. Run-on sentences are also not rare (*Finally I would like to ask if I need any special clothes and Do I need any money?*).

Speakers of *Germanic* languages often start common nouns with a capital letter, which is a transfer error since in many *Germanic* languages all nouns are capitalised. A related error is starting named entities with a lowercase letter, e.g. *december*, *french*, *friday*. This can be explained by rule overgeneralisation. In the CLC, both cases are annotated as

Error	G	Sw	HiG	LoG	Dan	Swe	W	N	Sp	Cat	Rom	Ger
Missing punctuation	+	+	+	+	+	+	+	+	+	−	+	+
Misused punctuation	−	+	+	+	+	+	+	+	+	+	+	+
Unnecessary punctuation	−	+	−	−	−	−	−	−	−	−	−	−
Spelling errors	+	+	+	+	+	+	+	+	+	+	+	+
Misused prepositions	+	−	+	+	+	+	+	+	+	+	+	+
Misused verbs	+	+	+	−	−	+	−	−	+	+	+	−
Verb tense	+	−	−	+	+	−	+	+	−	+	−	+

Table 3.1: Most typical errors for the Indo-European languages.

`misused punctuation`. Another instance of this error type is the misuse of a colon as in *dear Anna : I am writing to...* or *Dear Peter: Im sorry that ...* by speakers of Catalan and Spanish.

`Unnecessary punctuation` marks are often inserted by speakers of Swiss German: they overuse commas as in *Although, I'm going to take warm clothes with me...*

`Spelling` is another problematic issue for all English language learners. Spelling errors usually stem from the spelling conventions and phonology of the native languages. Here are some typical misspellings: *comfortabel* (*komfortabel* 'comfortable' in German), *summerfestival* (*Sommerfest* 'summer festival' in German and Swiss German), *interessts* (*interessiert* 'interested' in German and Swiss German), *parlement* (*parlement* 'parliament' in Dutch), *eksam* (*eksamen* 'examination' in Danish), *attencion* (*atención* 'attention' in Spanish), *advantatge* (*avantatge* 'advantage' in Catalan). Other frequent errors include dropping of letters as in *abosolut_ly*, or *immediat_ly*, or unnecessary letter duplication as in *affraid*, *dissapointment* typical for speakers of Danish and Swedish languages.

`Misuse of prepositions` is common for almost all of the English language learners. The typical confusions include such pairs of prepositions as *in – on*, *in – into*, *in – at*, *of – about*. These errors could be attributed to the complexity of the English language system. In addition, some languages use similar prepositions differently: for example, German preposition *in* has wider functionality than the English one. As a result, *in* is overused by German speakers. Some of the errors are clearly native language related: another error typical for German speakers is the use of *since* instead of *for* (*I have been*

learning English since seven years and started to learn Spanish). In German, in this case, preposition *seit* is used, which stands for both *since* and *for* in the temporal meaning.

The differences in the lexical systems result in the `misuse of certain verbs`. Verbs *say* and *tell*, *bring* and *take*, *do* and *make* are commonly confused. This is explained by the fact that in many languages they have one verb as a translation (*say* and *tell* are both *sagen* in German, and *decir* in Spanish, *do* and *make* are both *machen* in German, and *fer* in Catalan).

The English `verb tense` system is also a problematic area for the English language learners. The most common errors are caused by the agreement of tenses rules (*When I went to the meeting, we discuss ways of making our village more attractive*), the use of continuous aspect (*I'm always looking forward to a visit to the zoo*), and perfect tenses (*I have been lived here for three years*).

To summarise, all the Indo-European languages have similar problematic areas, such as the use of punctuation, verbs and prepositions. Misspellings are also common. These are the topics that need special attention when learning English. But since these errors are typical for all the considered languages and they do not vary across the languages, it is assumed that they are not discriminative as features.

Discriminative Error Types

To extract error types that are discriminative for the set of language pairs under consideration, information gain (IG) is used, as it was shown in Yang and Pedersen (1997) that this feature selection method is one of the most effective in text categorisation. Let et be an error type, and $\{l_i\}_{i=1}^{2}$ denote a pair of languages. Then $Pr(l_1|et)$ is the probability that a text is produced by a speaker of language l_1 if errors of type et are present in this text, and $Pr(l_1|\bar{et})$ is the probability of l_1 if errors of type et are absent. A logarithm to the base 2 is taken to measure the number of bits of information for language prediction obtained from presence or absence of the error type in the text. To calculate the information gain of an error type for a language pair, all the texts produced by speakers of one language within a pair are treated as a single text.

IG is calculated using Equation 3.2:

$$\begin{aligned} IG(et) = &-\sum_{i=1}^{2} Pr(l_i) log_2 Pr(l_i) \\ &+Pr(et)\sum_{i=1}^{2} Pr(l_i|et) log_2 Pr(l_i|et) \\ &+Pr(\bar{et})\sum_{i=1}^{2} Pr(l_i|\bar{et}) log_2 Pr(l_i|\bar{et}) \end{aligned} \tag{3.2}$$

Tables 3.2 to 3.7 present the top 10 informative error types for every language pair. Examples of the error types in the form *error|correction* are given in the last column of every table.

It should be noted that a great part of these discriminative error types include the use of determiners (codes ending with _D), prepositions (_T) and anaphoric pronouns (_A). The use of function words appears to be discriminative. Other discriminative error types include missing content (M_) or content that needs replacing (R_).

Error code	Description	IG	Example
R	**R**eplace word/phrase	0.082	Jenny's voice *appeared|could be heard*
U	**U**nnecessary word/phrase	0.055	Edinburgh is quite far *in the|_* north
DJ	**D**erivation of ad**J**ective	0.053	*sport|sports* facilities
UP	**U**nnecessary **P**unctuation	0.042	*mountain-spring-water|mountain spring water*
M	**M**issing word/phrase	0.038	*Not|It is not* necessary to say that ...
RC	**R**eplace **C**onjunction	0.035	farmers from Germany *as|and* France
L	Register error (**L**abel)	0.035	In the past people didn't read books all the time. Neither did their *kids|children*.
RY	**R**eplacing adverb	0.034	there are plenty of shops *especially|specially* for wedding clothes
UJ	**U**nnecessary ad**J**ective	0.03	the *urban|_* city of Zurich
MP	**M**issing **P**unctuation	0.028	satisfy *everyones|everyone's* wishes

Table 3.2: Top informative error types for *German – Swiss German*.

Error code	Description	IG	Example
U	**U**nnecessary word/phrase	0.114	We spent the second week *we spent|_* in Austria.
R	**R**eplace word/phrase	0.093	it was closed *because|for* refurbishment
M	**M**issing word/phrase	0.09	it is easier for *_|those of* us who live close enough
CE	**C**omplex **E**rror	0.036	*What is that for a* life?
FN	Wrong **N**oun **F**orm	0.036	chance to see some *kind|kinds* of animals
AGA	**A**naphor **AG**reement error	0.031	*It was|Those were* very nice days
RA	**R**eplace **A**naphor	0.023	Take good care of *you|yourself*.
MV	**M**issing **V**erb	0.0205	I *_|would* like you accept my Words of Wisdom
MN	**M**issing **N**oun	0.02	my fifteen *_|year* old son
DA	**D**erivation of **A**naphor	0.018	*Your|Yours* sincerely

Table 3.3: Top informative error types for *Danish – Swedish*.

Error code	Description	IG	Example
M	**M**issing word/phrase	0.092	I was interested in your job advertisement _|*for a* 'SUMMER JOB'
RJ	**R**eplace ad**J**ective	0.052	cages in a *common|ordinary* zoo are too small
UT	**U**nnecessary preposi**T**ion	0.046	once *in*|_ every two months
CN	**C**ountability of **N**oun error	0.043	We would be pleased to get *informations| information*
RY	**R**eplace adverb	0.041	*At first|First*, I want to say that ...
RD	**R**eplace **D**eterminer	0.041	Most of us won't have the chance in *their|our* whole lives
FV	Wrong **V**erb **F**orm	0.0405	I suggest *to go|going* to a huge shopping centre
MD	**M**issing **D**eterminer	0.038	As _|*a* present, I would like to have ...
RQ	**R**eplace **Q**uantifier	0.036	you can phone *every time|any time* you need help
RN	**R**eplace **N**oun	0.036	the *offer|price* of the sports facilities was not expensive

Table 3.4: Top informative error types for *High German – Low German*.

Error code	Description	IG	Example
AGV	**V**erb **AG**reement error	0.076	Therefore I can understand that many people *prefers|prefer* a car.
CE	**C**omplex **E**rror	0.0625	*During the fall|As he fell* his head hit the shelf
IV	Incorrect **I**nflection of **V**erb	0.037	many thanks for your letter which *gaves|gave* me enough information
RJ	**R**eplace ad**J**ective	0.029	technology is not very *big|great*
AGD	**D**eterminer **AG**reement error	0.024	when I look at *this|these* photos I am relaxed
S	**S**pelling error	0.022	*recive|receive*
RQ	**R**eplace **Q**uantifier	0.022	bicycle in the winter isn't *very|much* fun
RY	**R**eplace adverb	0.019	I've been very busy with work *at the moment| recently*
UP	**U**nnecessary **P**unctuation	0.014	the *entry-price|entry price* includes a free lunch
CN	**C**ountability of **N**oun error	0.013	The village must develop *accommodations| accommodation* for holidays.

Table 3.5: Top informative error types for *Western Germanic – Northern Germanic*.

Error code	Description	IG	Example
UD	**U**nnecessary **D**eterminer	0.023	it isn't in *a*|_ good condition
RN	**R**eplace **N**oun	0.021	Keeping animals in zoos is a common *costume*| *custom*
UT	**U**nnecessary preposi**T**ion	0.017	I would like *to*|_ you to give me more information
AS	**A**rgument **S**tructure error	0.014	You have to *be very concentrate in*|*concentrate hard on* how you are driving
AGA	**A**naphor **AG**reement error	0.014	I think the best *one*|*ones* are swimming and painting.
CL	**C**o**L**location error	0.012	people going *out and in*|*in and out* the places
MY	**M**issing adverb	0.011	My family were_|*usually* at home but that day they had gone to a party
AGD	**D**eterminer **AG**reement error	0.01	associations have been trying to show *this*|*these* troubles to people
IQ	**I**nflection of **Q**uantifier error	0.009	I am a member of *others*|*other* clubs
FN	Wrong **N**oun **F**orm	0.009	cars are more comfortable than *bicycle*|*bicycles*

Table 3.6: Top informative error types for *Spanish – Catalan*.

Error code	Description	IG	Example
MA	**M**issing **A**naphor	0.0395	Regarding your letter _|*which* I received yesterday
RA	**R**eplace **A**naphoric	0.033	*there*|*it* is only 50 kilometres to get there
M	**M**issing word/phrase	0.03	it doesn't matter if _|*it's* winter or summer
RV	**R**eplace **V**erb	0.028	we haven't *met*|*seen* each other for years
MC	**M**issing **C**onjunction	0.022	They usually appear on TV, _|*and* go to shows
MD	**M**issing **D**eterminer	0.022	you have _|*a* chance of finding a job
MT	**M**issing preposi**T**ion	0.02	I will arrive _|*on* Tuesday
UT	**U**nnecessary preposi**T**ion	0.0195	I visited it *for*|_ three years ago
FJ	Wrong ad**J**ective **F**orm	0.018	watching television is *best*|*better* than reading a book
FV	Wrong **V**erb **F**orm	0.017	get information about *to become*|*becoming* a future member of your club

Table 3.7: Top informative error types for *Romance – Germanic*.

Chapter 4

Evaluation

In this chapter, results of the binary classification performed on the data sets are presented. Results of the distributional analysis and error-based analysis experiments are reported and discussed in Sections 4.1 and 4.2, respectively.

All results are obtained using 5-fold cross-validation.

4.1 Distributional Analysis

4.1.1 Results

The following notation for the features presented in Section 3.2 is used in the results tables:

`t` – word unigrams

`t2` – word bigrams

`p` – PoS uni-, bi- and trigrams

`ch1` – character unigrams

`ch2` – character uni- and bigrams

`ch3` – character uni-, bi- and trigrams

`ch4` – character uni-, bi-, tri- and quadgrams

`c` – corpus-derived error rate

`e` – CLC error rate

`r` – PS rules

Table 4.1 shows the model's performance when using each feature independently on different language pairs. The best results using separate features are in bold. The bottom

Feature	G – Sw	HiG – LoG	Dan – Swe	W – N	Sp – Cat	Rom – Ger
t	86.25%	91.41%	88.33%	78.87%	60.00%	80.16%
t2	87.50%	**93.81%**	93.34%	77.09%	62.00%	**83.56%**
p	80.00%	84.50%	**95.83%**	75.30%	58.80%	79.93%
ch1	66.25%	77.33%	76.67%	64.29%	54.80%	69.39%
ch2	81.25%	86.60%	92.50%	69.35%	**66.40%**	76.19%
ch3	88.75%	89.00%	94.17%	75.59%	63.20%	79.70%
ch4	**90.00%**	90.05%	91.67%	**79.16%**	63.60%	80.16%
c	62.50%	48.54%	56.66%	53.57%	50.00%	50.79%
e	55.00%	57.60%	50.83%	47.92%	52.40%	59.98%
r	57.50%	65.54%	63.33%	65.48%	54.80%	65.42%
Best combination	(ch4) **90.00%**	(ch4+p) **95.19%**	(ch3+p+c+e+r) **97.50%**	(ch4) **79.16%**	(ch2+r) **68.40%**	(ch3+p+c) **84.35%**

Table 4.1: Distributional analysis results

row of the table contains the highest accuracy obtained on the language pair, and the best performing feature combination.

For all but one language pairs, the two classes are distributed equally both in the training and in the test set. Hence, the majority baseline for all the pairs except for the *Western Germanic – Northern Germanic* pair is 50%. For the *Western Germanic – Northern Germanic* pair, the baseline is 50.4%, with *Western Germanic* being the majority class.

In all the cases, classification using PoS n-grams, character-based n-grams and PS rules results in a higher performance than the baseline. The lowest accuracy across the language pairs is obtained when using the corpus-derived and CLC error rates. In a number of cases, the system performance using error rates is lower or equal to the majority baseline (see the accuracy of classification of *High German – Low German* using corpus-derived error rate `c`, or *Western Germanic – Northern Germanic* using CLC error rate `e`). Classification with any of the error rates does not result in a significantly higher performance than the baseline, where the significance is measured using paired t-test, $\alpha = 0.05$.

As has been noted earlier (Section 3.2.1), classification using token-based features has a potential of giving high accuracy, especially if the scripts in the two sets are produced in response to different prompts. This is the case with some of the language data sets. In the *German* and *Swiss German* sets only about 6% of the texts per set are written on the same topics, in the *High German* and *Low German* sets only 4%, and the *Danish* and *Swedish* sets do not have any common topics at all. An accuracy of classification with token-based n-grams for these language pairs ranges from 87.50% to 93.81%, which is, presumably, due to the fact that the texts' content differs to a considerable degree. This means, that tokens are topic-specific rather than related to the native language. Remarkably, the

accuracy of classification using token-based n-grams is lower on the other three language pairs where 46% to 49% of the texts per data set share topics with the opposed data set.

Token-based classification results are hard to beat. However, the accuracy of the best performing combination of features not including token-based n-grams is higher than the accuracy of classification using token-based n-grams for all the language pairs. Paired t-test shows that this improvement in the accuracy is statistically significant.

All of the best performing feature combinations contain character-based n-grams. Character-based n-grams also model text content to a certain degree. An average word length for the texts in our data sets ranges from 3.75 to 3.83 characters per word, with about 20% of all the words being of length 3. That means that a set of character n-grams of length up to 4 extracted from the texts contains a number of function words as well as morphological units like suffixes.

PoS n-grams are topic-independent features. For all but one language pairs, classification using PoS n-grams results in an accuracy higher than 75%. An accuracy of classification using PoS n-grams is lower than 60% for the *Spanish – Catalan* pair, but on this language pair even the best feature combination performs worse than 60%.

Classification using PS rules results in lower accuracy than classification with PoS or character-based n-grams, but in general, an accuracy of classification using PS rules is significantly higher than the majority baseline.

4.1.2 Feature contribution

We measure the significance of the contribution of different features running a number of paired t-tests with an α value of 0.05.

For every feature f from the feature set, a group of results obtained using feature combinations including the feature f is compared to the group of results obtained using the same feature combinations excluding the feature f. For example, to measure statistical significance of the contribution of the PoS n-grams `p`, all the combinations of features `ch1` through `ch4+c+e+r` are compared to the correspondent combinations `ch1+p` through `ch4+c+e+r+p`.

Tables 4.2 to 4.7 show the results in terms of the absolute gain in percentage points, statistical significance of adding a feature (with '+' or '−') and the correspondent p value. We also report the significance of incrementing the order of an n-gram by one: for example, we compare the results of the feature combinations `ch1`, `ch1+p`, `ch1+c` and the others with those of `ch2`, `ch2+p`, `ch2+c` and the other correspondent combinations.

Character n-grams significantly improve the classification accuracy in almost all cases. However, incrementing the order of n-grams does not always result in a better accuracy. For example, for *Romance* languages, increasing the length of an n-gram from 3 to 4 lowers

the accuracy. PoS n-grams significantly increase the accuracy in all cases. Error rates do not contribute significantly in any of the language pairs. Adding PS rules improves the results in all cases, but this improvement is statistically significant in half of them only.

Feature	Significance	p value	Percentage points
Adding `ch1`	+	0.0072	3.5
Adding `ch2`	+	0.0008	11.41
Adding `ch3`	+	< 0.0001	17.16
Adding `ch4`	+	< 0.0001	**19.08**
Incrementing `ch1` to `ch2`	+	0.0003	8.36
Incrementing `ch2` to `ch3`	+	< 0.0001	5.86
Incrementing `ch3` to `ch4`	+	< 0.0001	1.87
Adding `p`	+	0.0003	5.7
Adding `c`	−	0.5197	0.2
Adding `e`	−	1.0000	0.0
Adding `r`	−	0.3833	0.19

Table 4.2: Feature contribution for *German – Swiss German*.

Feature	Significance	p value	Percentage points
Adding `ch1`	+	0.0107	7.47
Adding `ch2`	+	0.0004	15.5
Adding `ch3`	+	< 0.0001	18.37
Adding `ch4`	+	< 0.0001	**18.8**
Incrementing `ch1` to `ch2`	+	< 0.0001	7.17
Incrementing `ch2` to `ch3`	+	< 0.0001	3.78
Incrementing `ch3` to `ch4`	+	0.0003	0.47
Adding `p`	+	< 0.0001	8.18
Adding `c`	−	0.5065	-0.17
Adding `e`	−	0.0539	0.16
Adding `r`	−	0.1197	1.2

Table 4.3: Feature contribution for *High German – Low German*.

Feature	Significance	p value	Percentage points
Adding `ch1`	+	0.0050	8.5
Adding `ch2`	+	0.0054	15.5
Adding `ch3`	+	0.0038	**16.5**
Adding `ch4`	+	0.0031	15.8
Incrementing `ch1` to `ch2`	+	0.0030	5.55
Incrementing `ch2` to `ch3`	+	< 0.0001	1.0
Incrementing `ch3` to `ch4`	−	0.0684	-0.7
Adding `p`	+	< 0.0001	12.64
Adding `c`	−	0.2373	0.23
Adding `e`	−	0.4580	-0.09
Adding `r`	−	0.2534	0.04

Table 4.4: Feature contribution for *Danish – Swedish*.

Feature	Significance	p value	Percentage points
Adding `ch1`	−	0.1218	2.42
Adding `ch2`	+	0.0141	5.75
Adding `ch3`	+	0.0355	7.68
Adding `ch4`	+	0.0077	**9.2**
Incrementing `ch1` to `ch2`	+	0.0017	3.46
Incrementing `ch2` to `ch3`	−	0.1721	1.26
Incrementing `ch3` to `ch4`	+	< 0.0001	2.59
Adding `p`	+	0.0002	5.32
Adding `c`	−	0.8867	-0.03
Adding `e`	−	0.7942	-0.02
Adding `r`	+	0.0043	1.77

Table 4.5: Feature contribution for *Western Germanic – Northern Germanic*.

Feature	Significance	p value	Percentage points
Adding `ch1`	+	0.0101	1.1
Adding `ch2`	+	0.0002	**8.2**
Adding `ch3`	+	< 0.0001	7.2
Adding `ch4`	+	0.0003	6.1
Incrementing `ch1` to `ch2`	+	< 0.0001	7.5
Incrementing `ch2` to `ch3`	−	0.0906	-1.1
Incrementing `ch3` to `ch4`	+	0.0019	-1.0
Adding `p`	+	0.0205	1.8
Adding `c`	−	0.0810	-0.2
Adding `e`	−	0.4113	-0.08
Adding `r`	+	< 0.0001	2.1

Table 4.6: Feature contribution for *Spanish – Catalan*.

Feature	Significance	p value	Percentage points
Adding `ch1`	+	0.0094	4.3
Adding `ch2`	+	0.0022	7.665
Adding `ch3`	+	0.0003	**11**
Adding `ch4`	+	0.0009	10.26
Incrementing `ch1` to `ch2`	+	0.0002	3.58
Incrementing `ch2` to `ch3`	+	< 0.0001	3.35
Incrementing `ch3` to `ch4`	+	0.0018	-0.667
Adding `p`	+	< 0.0001	8.275
Adding `c`	−	0.0587	-0.315
Adding `e`	−	0.2365	0.29
Adding `r`	+	0.0117	1.04

Table 4.7: Feature contribution for *Romance – Germanic*.

Table 4.8 summarises the results of the t-tests. The following notation is used:

↗ – increase in accuracy

↘ – decrease in accuracy

→ – no change in accuracy

\+ – statistically significant

− – not statistically significant

Feature	G – Sw	HiG – LoG	Dan – Swe	W – N	Sp – Cat	Rom – Ger
Adding `ch1`	+ ↗	+ ↗	+ ↗	− ↗	+ ↗	+ ↗
Adding `ch2`	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗
Adding `ch3`	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗
Adding `ch4`	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗
Incrementing `ch1` to `ch2`	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗
Incrementing `ch2` to `ch3`	+ ↗	+ ↗	+ ↗	− ↗	− ↘	+ ↗
Incrementing `ch3` to `ch4`	+ ↗	+ ↗	− ↘	+ ↗	+ ↘	+ ↘
Adding `p`	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗	+ ↗
Adding `c`	− ↗	− ↘	− ↗	− ↘	− ↘	− ↘
Adding `e`	− →	− ↗	− ↘	− ↘	− ↘	− ↗
Adding `r`	− ↗	− ↗	− ↗	+ ↗	+ ↗	+ ↗

Table 4.8: General feature contribution across language pairs.

4.2 Error-Based Analysis

In Section 3.3, we discussed using error type rates, error type distribution and error content as features. We have also described the types of errors that we take into account. These include error types considered in the previous studies, namely misuse of determiners, subject–verb disagreement and noun number disagreement. We have also selected discriminative error types for each of the language pairs (see Tables 3.2 through 3.7).

The results obtained using these feature types are presented in Tables 4.9 through 4.11.

In the tables, row $N = 1$ shows an accuracy of classification using features based on all error types, where 'all' refers to all the errors tagged in the CLC corpus. Rows $N = 3$ to $N = 5$ show the results obtained using smaller sets of language-specific error types: from 10 to 1 discriminative error types selected using IG. In addition, a set of 10 most discriminative error types across all the languages is collected. An accuracy of classification using features based on these error types is presented in row $N = 2$.

As we have noted above, we replicate the experiments of Wong and Dras (2009), considering determiners-related, subject–verb agreement and noun number agreement related

errors. The results obtained in our experiments are presented in rows $N = 6$ to $N = 8$. Since agreement is usually considered to be problematic for English language learners, we consider different types of agreement-related errors: besides subject–verb and noun number agreement, these also include determiner and anaphoric pronoun agreement errors. We use a combination of features based on the four agreement-related types of errors. The results are presented in row $N = 9$. We combine features based on the four agreement-related types of errors with the determiner-related errors and use them in classification. Row $N = 10$ shows the results of this classification. Finally, spelling errors are also considered and the results are presented in row $N = 11$.

The best result for every language pair using a particular type of features is in bold.

N	Type	G – Sw	HiG – LoG	Dan – Swe	W – N		Sp – Cat		Rom – Ger
1	All	56%	72%	69.17%	70.83%		58.8%		**68.88%**
2	Top 10 IG across languages	55%	73.5%	**77.5%**	**72.92%**		**64.29%**		54.8%
3	Top 10 IG for pair	58%	**75%**	72.5%	72.08%		57.6%		68.71%
4	Top 3 IG for pair	**60%**	71.5%	74.17%	69.58%		54%		65.31%
5	Top 1 IG for pair	R: 55% U: 59% DJ: 55%	M: 61% RJ: 61.5% UT: 58.5%	U: 62.5% R: 65.83% M: 62.5%	AGV: 64.58% CE: 60.83% IV: 62.08%		UD: 50.8% RN: 51.6% UT: 52.8%		MA: 55.95% RA: 52.55% RV: 64.97%
6	Det	53%	64.5%	53.33%	55.83%		50.8%		62.25%
7	S - V agr	50%	48.5%	51.67%	64.58%		50.4%		54.42%
8	NN agr	55%	57%	57.5%	57.92%		49.2%		51.87%
9	All agr-related	54%	44%	50%	67.08%		54.4%		54.93%
10	All 6 - 9	54%	64%	49.17%	66.25%		52.8%		55.78%
11	Spelling	50%	56.5%	50.83%	62.5%		52.8%		55.78%

Table 4.9: Error type rates results.

The results show, that classification using the features based on the types of errors selected with IG results in a higher performance than classification using the features based on the types considered by Wong and Dras (2009). As discussed earlier, such phenomena as the use of determiners, subject–verb agreement and noun number agreement are usually considered to be problematic for English language learners. However, our results confirm that these types of errors are not discriminative for classification, neither individually nor in combination. Only in rare cases a combination of these error types performs better than the other error types (e.g., *Spanish – Catalan*, Table 4.11, row $N = 9$).

N	Type	G – Sw	HiG – LoG	Dan – Swe	W – N		Sp – Cat		Rom – Ger
1	All	61.67%	68.98%	70%	72.32%		57.6%		**70.07%**
2	Top 10 IG across languages	57%	69.5%	72.5%	70.83%		**64.29%**		53.6%
3	Top 10 IG for pair	**64%**	**72%**	75%	**74.17%**		58.8%		68.54%
4	Top 3 IG for pair	62%	61.5%	**81.67%**	70%		54.4%		61.84%
5	Top 1 IG for pair	R: 56% U: 56% DJ: 60%	M: 60% RJ: 59% UT: 54.5%	U: 62.5% R: 70% M: 61.67%	AGV: 65.83% CE: 62.08% IV: 57.08%		UD: 54% RN: 54.8% UT: 54.4%		MA: 55.44% RA: 54.42% RV: 62.59%
6	Det	51%	63%	60.83%	54.17%		55.6%		56.12%
7	S - V agr	50%	51%	55%	65.83%		52%		55.27%
8	NN agr	54%	54%	55%	55.83%		50.4%		50%
9	All agr-related	56%	47.5%	56.67%	65.42%		46.8%		55.1%
10	All 6 - 9	51%	64%	59.17%	62.5%		56%		56.12%
11	Spelling	55%	58%	49.17%	63.33%		55.6%		58.78%

Table 4.10: Error type distribution results.

N	Type	G – Sw	HiG – LoG	Dan – Swe	W – N		Sp – Cat		Rom – Ger
1	All	**73%**	**74%**	**72.5%**	59.17%		49.17%		**69.39%**
2	Top 10 IG across languages	57%	56%	55%	64.17%		52.04%		52.4%
3	Top 10 IG for pair	65%	59%	62.5%	62.92%		45.2%		63.95%
4	Top 3 IG for pair	55%	52.5%	58.34%	**65%**		44%		59.52%
5	Det	49%	55.5%	52.78%	54.58%		52%		54.93%
6	S -V agr	52.5%	55.5%	53.33%	63.33%		50.4%		55.1%
7	NN agr	46.25%	58%	58.33%	52.08%		52.4%		55.78%
8	All agr-related	54%	55%	55%	61.67%		51.6%		51.36%
9	All 5 - 8	48.75%	63%	57.5%	60.42%		**54%**		56.8%
10	Spelling	53%	58%	55.83%	52.08%		47.6%		59.18%

Table 4.11: Error content results.

Within the distributional analysis, we considered the general error rates (see Section 4.1) which we calculated taking all the errors into consideration and not distinguishing between different error types. As compared to the general error rates, both error type rates and error type distribution used here result in a significantly better performace ($p = 0.035$ and $p = 0.0121$ compared to the corpus-derived error rate `c`, $p = 0.015$ and $p = 0.0085$ compared to the CLC error rate `e`, respectively). The results of error type rates and error type distribution (Tables 4.9 and 4.10) also show that, on the average, classification using features based on 10 IG-selected error types (row $N = 3$) yields higher accuracy than classification using the full set, i.e. 'all' error types (row $N = 1$). This improvement is, however, not statistically significant.

For the "error content"-based features (see Table 4.11), combination of all the error types for almost all of the language pairs yields higher accuracy than combination of IG-selected error types (cf. rows $N = 1$ and $N = 3$, $N = 1$ and $N = 4$).

With respect to classification using error type rates (Table 4.9), classification using features based on the top 10 IG error types selected across all the language pairs, on the average, yields better results than using the features based on the full set of error types (cf. rows $N = 1$ and $N = 2$). In all other cases (Tables 4.10 and 4.11), the features based on the top 10 cross-language error types perform worse both compared to the full set of error types and to the error types selected for each language pair (cf. rows $N = 2$ and $N = 1$, $N = 2$ and $N = 3$). In all the cases, however, these differences are not statistically significant.

4.3 Combining Two Approaches

Wong and Dras (2009) reported that adding syntactic error features including misuse of determiners, subject–verb disagreement and noun number disagreement, did not improve the results obtained using lexical types of features. In the previous section, we showed that features based on the IG-selected error types performed better than the syntactic error types considered by (Wong and Dras, 2009). Classification using features based on the full set of error types also results in a higher accuracy than classification using only these three error types.

To check, whether adding the error types would improve performance on our data, we add the features based on the error types to the best performing feature combinations identified using distributional approach (see Section 4.1). We consider features based on the error type rates, error type distribution and error content of the full set of error types.

The results are presented in Table 4.12, and the best accuracy for each language pair is in bold.

Features	G – Sw	HiG – LoG	Dan – Swe	W – N	Sp – Cat	Rom – Ger
Best distributional result	(ch4) 90.00%	(ch4+p) 95.19%	(ch3+p+c+e+r) 97.50%	(ch4) 79.16%	(ch2+r) **68.40%**	(ch3+p+c) **84.35%**
All error type rates	56%	72%	69.17%	70.83%	58.8%	68.88%
Combination	86%	96%	**100%**	**80%**	67.6%	83.67%
All error type distribution	61.67%	68.98%	70%	72.32%	57.6%	70.07%
Combination	56%	73%	**100%**	72.5%	56%	67.3%
All error type content	73%	74%	72.5%	59.17%	49.17%	69.39%
Combination	**92%**	**96.5%**	**100%**	**80%**	67.6%	83.67%

Table 4.12: Combination of the approaches.

The results show that adding error type based features, in particular, "error content"-based features can improve the results obtained using only distributional features for some language pairs. In general, the difference is not higher than 2.5 percentage points.

4.4 Classification with Spelling Errors

In Tsur and Rappoport (2007), it was hypothesised that the choice of words is influenced by the writer's native language. This hypothesis was confirmed, as an accuracy of 65.60% was obtained using only character bigrams on the 5-class task, i.e. distinguishing between 5 native languages (see Section 1.2). Experiments on the CLC show an accuracy in the range of 66.40% to 92.50% on the binary classification task when using character unigrams and bigrams. It is logical to assume, however, that if the choice of words is influenced by the language phonology, then the misspelled words would be strong indicators of different native languages.

Section 4.2 presents the results obtained when using misspellings tagged in the CLC corpus as features (see Table 4.11, row $N = 11$). However, all of those results are below 60%.

In this section, we present an alternative approach. The extraction of the misspelled words from the texts is based on comparison of the words contained in the texts to a big word list of standard English[1]: if a word is not contained in the word list, it is considered to be

[1]English Word List of 755,110 words, http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml

a misspelling. Then, sets of character n-grams from the misspelled words are compared to the standard English n-grams, and those missing from the English set are considered to be erroneous.

We adopt another approach to the features weighting. Feature vectors consist of binary values: 1 denotes presence of a particular misspelling or character n-gram in a text, and 0 – its absence. Thus, features are weighted with respect to presence/absence and not with respect to their frequency.

Table 4.13 presents the results of classification using misspelled words and character n-grams as features. The following notation is used:

`t` – misspelled words

`ch2` – misspelled character uni- and bigrams

`ch3` – misspelled character uni-, bi- and trigrams

`ch4` – misspelled character uni-, bi-, tri- and quadgrams

The best accuracy per language pair is in bold.

Feature	G – Sw	HiG – LoG	Dan – Swe	W – N		Sp – Cat		Rom – Ger
`t`	84.00%	81.50%	81.67%	72.25%		56.40%		68.24%
`ch2`	52.00%	57.00%	46.67%	52.24%		55.20%		57.15%
`ch3`	88.00%	**90.50%**	**100.00%**	73.88%		**60.00%**		72.94%
`ch4`	**89.00%**	89.50%	**100.00%**	**74.29%**		59.60%		**76.91%**

Table 4.13: Misspellings as features.

Misspelled words and character n-grams reveal phonological influence and native language spelling conventions clearly as opposed to the correctly spelled words. Table 4.14 presents some of the discriminative words and character n-grams for the *Romance – Germanic* pair.

Language group		Misspelling	Example
Romance		univesidad	Spanish: *univesidad* 'university'
Germanic		smal	Swedish: *smal* 'narrow'
Romance		-sctr-	Portuguese: *esctructure* 'structure'
		-rtvi-	Italian: *entertview* 'interview'
		-aua-	Italian: *becauase* 'because'
		-aue-	French, Catalan, Spanish: *becaue*, *becaues* 'because'

Table 4.14: Discriminative misspellings for *Romance – Germanic*.

Although an accuracy of above 80% is obtained with misspelled words on three language pairs, overall the results are significantly lower than classifying with all token unigrams.

However, we discussed earlier that this should be attributed to the fact that token unigrams are mostly topic-specific. Word misspellings, on the other hand, are less related to the topics.

Classification with combinations of erroneous unigrams and bigrams shows lower results than classification with all unigrams and bigrams (see Table 4.1, results for `ch2`), though for combinations of erroneous character n-grams including n-grams longer than 3 the difference is only minor and not statistically significant (cf. results for `ch3` and `ch4` here and in Table 4.1).

This approach provides high results as opposed to classification with the CLC-detected misspellings (see Table 4.11, row $N = 11$). This is partly due to the fact that the method of comparing words from the texts to a corpus of standard English helps finding misspellings as well as foreign words and peculiar named entities. For example, a Portuguese word *Oejros*, which is a name of some locality, is not contained in the English corpus and would be identified by this method. In the CLC, it is not tagged. It should be noted that even though named entities are not misspellings in the proper sense, when used by learners they reveal certain characteristics of language specific spelling.

This approach shows promising results. It has two differences as compared to the previous approach described in Section 4.2: the misspellings are extracted using a big corpus of English rather than manual error tagging, and features are weighted with respect to presence/absence and not with respect to their frequency. To make direct comparison to the CLC-based results (see Table 4.11, row $N = 11$), the erroneous character n-grams should be extracted from the CLC-detected misspellings, and the misspellings should be weighted by presence/absence.

Chapter 5

Conclusions and Future Work

In this project, the task of native language identification has been investigated on a set of Indo-European languages. As compared to previous works in this area, a wider set of languages has been considered and a more systematic approach to classification of closely related languages has been undertaken. A number of binary classification experiments with support vector machines have been carried out.

It has been shown that languages of two Indo-European language groups, namely *Germanic* and *Romance*, can be distinguished with an accuracy of 84.35%. On our data, classification accuracy for the language pairs within these branches ranges from 68.40% for the *Spanish – Catalan* pair to 100.00% for the *Danish – Swedish* pair. Since the previous studies have been performed on sets of other, mostly unrelated, languages, direct comparison of the results is not possible.

We have formed a hypothesis that multi-class classification would perform better on the group of closely related *Romance* languages. Multi-class classification is an interesting topic in its own right, and it is left for future research.

We have explored a wide range of different features. A hypothesis formulated by Tsur and Rappoport (2007) that native language exerts influence on the choice of words in English, has been confirmed: character n-grams of length 2 to 4 on all the language pairs performed better than any other features and contributed the most to the feature combinations. PoS n-grams have also been shown to perform well on the native language classification task.

A contrastive analysis hypothesis that error types encode the differences between the native language and English systems and, hence, are strong indicators of native languages, has also been studied. We showed that types of errors committed by learners depend on their native languages, and classification with error types selected from the data is more accurate than classification with a uniform set of typical errors as in Wong and Dras (2009). It has also been shown that classification with only erroneous character n-grams extracted from misspelled words yields results comparable to those with all the n-grams.

Lists of the most typical errors for each language have been collected from the data and discussed in the paper.

The results obtained in this project show that there is room for improvement. In this project, we applied linear classifiers. Further inspection of the data may show that some language pairs data is not linearly separable. Thus, classification performance might be improved if we use *kernel-based* learning methods.

Based on our results, a native language identification tool can be built that would rely on a set of binary classifiers to identify a writer's native language from a set of Indo-European languages, with the binary classifiers combined in a pipeline. We leave this for future research.

Bibliography

S. Argamon, M. Koppel, and A.R. Shimoni. Gender, Genre, and Writing Style in Formal Written Texts. *TEXT*, 23(321-346), 2003.

B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422 – 426, 1970.

T. Briscoe, J. Carroll, and R. Watson. The second release of the RASP system. In *Proceedings of the COLING/ACL*, volume 6, 2006.

T. Briscoe, B. Medlock, and Ø. Andersen. Automated assessment of ESOL free text examinations. Technical report, University of Cambridge, Computer Laboratory, 2010.

C. E. Chaski. Empirical evaluations of language-based author identification techniques. *International Journal of Speech Language and the Law*, 8(1), 2001.

M. Chodorow, J. R. Tetreault, and N. Han. Detection of Grammatical Errors Involving Prepositions. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*, 2007.

N. Christianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

S. P. Corder. The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 5(4):161 – 170, 1967.

M. Corney, O. de Vel, A. Anderson, and G. Mohay. Gender-preferential text mining of e-mail discourse. In *Proceedings of the 18th Annual Computer Security Applications Conference (ACSAC 2002)*, pages 282 – 292, 2002.

M. Coulthard. Author identification, idiolect and linguistic uniqueness. *Applied linguistics*, 25(4):431–447, 2004.

R. de Felice. *Automatic Error Detection in Non-native English*. PhD thesis, University of Oxford, 2008.

J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship Attribution with Support Vector Machines. *Applied Intelligence*, 19(1-2):109–123, 2003.

T. M. Ellison and S. Kirby. Measuring Language Divergence by Intra-Lexical Comparison. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006.

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

J. E. I. Farmer and M. Y. Erlewine. An Extensible Method for Authorship Identification. Technical report, University of Chicago, 2006.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Workshop Programme*, page 47, 2008.

M. Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics (COLING '04)*, 2004.

M. Gamon, J. Gao, C. Brockett, A. Klementiev, W. B. Dolan, D. Belenko, and L. Vanderwende. Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. In *Asia Federation of Natural Language Processing*, 2008.

G. E. Heidorn. *Intelligent Writing Assistance*, pages 181–207. R. Dale, H. Moisl, H. Somers eds., 2000.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings to the tenth European Conference on Machine Learning (ECML-98)*, pages 137–142, 1998.

T. Joachims. Learning to Classify Text Using Support Vector Machines. Dissertation, Kluwer, 2002.

M. Koppel and J. Schler. Exploiting Stylistic Idiosyncrasies for Authorship Attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico, 2003.

M. Koppel, S. Argamon, and A.R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), 2002.

M. Koppel, J. Schler, and K. Zigdon. 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*, 2005a.

M. Koppel, J. Schler, and K. Zigdon. *Automatically Determining an Anonymous Author's Native Language*, pages 209 – 217. P.Kantor et al., Springer-Verlag Berlin Heidelberg, 2005b.

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

J. S. Y. Lee. *Automatic Correction of Grammatical Errors in Non-native English Text.* PhD thesis, MIT, 2009.

N. McCombe. Methods Of Author Identification. B.A. Thesis, Trinity College, Dublin, 2002.

D. Nicholls. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics conference*, pages 572 – 581, 2003.

T. Odlin. *Language transfer: Cross-linguistic influence in language learning.* Cambridge University Press, 1989.

Mark Pagel. Human language as a culturally transmitted replicator. *Nature Reviews Genetics*, 10:405 – 415, 2009.

J. C. Richards. A non-contrastive approach to error analysis. *ELT Journal*, 25(3):204 – 219, 1971.

E. Stamatatos. Author Identification: Using Text Sampling to Handle the Class Imbalance Problem. *Information Processing and Management*, 44(2):790–799, 2008.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities*, 35(2):193–214, 2001.

J. H. Steiger. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245 – 251, 1980.

M. Swadesh. Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Phil.*, 96: 453 – 463, 1952.

J. R. Tetreault and M. Chodorow. Native Judgments of Non-Native Usage: Experiments in Preposition Error Detection. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, 2008a.

J. R. Tetreault and M. Chodorow. The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1, 2008b.

L. Mayfield Tomokiyo and R. Jones. 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)*, 2001.

O. Tsur and A. Rappoport. Using Classifier Features for Studying the Effects 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, 2007.

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

V. N. Vapnik. *The nature of statistical learning theory.* Springer, 1995.

E. J. Williams. The Comparison of Regression Variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 21(2):396 – 399, 1959.

I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques.* Second edition. Morgan Kaufmann, 2005.

S. J. Wong and M. Dras. Contrastive Analysis and Native Language Identification. In *Proceedings of the Australasian Language Technology Association Workshop*, 2009.

Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning*, 1997.

H. Yannakoudakis, T. Briscoe, and B. Medlock. A New Dataset and Method for Automatically Grading ESOL Texts. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.