Universität des Saarlandes
Phiolosophische Fakultät

Pluricentric Languages: Automatic Identification and Linguistic Variation

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Philosophie
der Philosophischen Fakultäten
der Universität des Saarlandes

vorgelegt von

Marcos Zampieri

aus Sao Paulo, Brasilien
Saarbrücken, 2016

Der Dekan: Prof. Dr. Erich Steiner

Erstberichterstatter: Prof. Dr. Josef van Genabith

Zweitberichterstatter: Prof. Dr. Stephan Busemann

Tag der letzten Prüfungsleistung: 20.10.2016

Pluricentric Languages: Automatic Identification and Linguistic Variation

ABSTRACT

Language Identification is a well-known research topic in NLP. State-of-the-art methods consist of the application of n-gram language models to distinguish languages automatically with well over 95% accuracy. This level of success is obtained when discriminating between languages that are typologically not closely related (e.g. Finnish and Spanish), or due to the contrast between languages with unique character sets such as Greek or Hebrew. Recent studies show that one of the main difficulties of n-gram based methods is the identification of closely related languages. The research presented in this thesis goes one step further and investigates computational methods to identify standard national varieties of pluricentric languages such as Portuguese, Spanish, French, and English. It explores different computational methods and different sets of features for this task that go beyond character and word language models. The main objective is to investigate the extent to which it is possible to identify language varieties automatically in both monolingual and in real-world (multilingual) settings and to establish what are the main challenges of this task in comparison to general purpose language identification models. This research shows, for example, that it is possible to discriminate between Brazilian and European Portuguese with 99.8% accuracy using journalistic texts. Another contribution of this thesis is the use of linguistically motivated features such as POS tags and morphological information to discriminate between language varieties with results of up to 83.1% accuracy in discriminating between Mexican and Peninsular Spanish texts. An additional aspect of this thesis is the use of classification output in corpus-driven contrastive linguistics research as explained in Chapter 6. Classification methods combined with linguistically meaningful features are able to provide empirical evidence on the convergences and divergences of language varieties in terms of lexicon, orthography, morphology and syntax.

Acknowledgements

First and foremost, I would like to thank my supervisors Josef van Genabith and Stephan Busemann for the support and extensive feedback they provided me. Without them, I would not have been able to complete this thesis.

I would like to thank all colleagues I collaborated with in the last years. I was fortunate to meet many great people along the way and most of them became very good friends. We published several papers together and I learned a lot from each of them. Special thanks to Liviu P. Dinu, Sascha Diwersy, Binyam Gebrekidan Gebre, Marco Lui, Shervin Malmasi, and Vlad Niculae for many fruitful discussions; to my colleagues from Saarland: Ekaterina Lapshinova-Koltunski, Santanu Pal, Liling Tan, and Mihaela Vela for the great collaboration, support and friendship; and to the VarDial workshop and DSL shared task organizers: Nikola Ljubešić, Preslav Nakov, and Jörg Tiedemann for making the workshop and the shared task successful.

I take this opportunity to thank the following people who provided me constructive and insightful feedback: Martin Becker (Chapter 2 and 6), Jon Dehdari (Chapter 3), Sascha Diwersy (Chapter 6), Binyam Gebrekidan Gebre (Chapters 1 and 3), and Nikola Ljubešić (Chapters 1 and 3). I tried to incorporate their feedback in this thesis as much as possible. Remaining errors and misconceptions are, of course, my own.

I would like to thank Heike Przybyl and Anne-Kathrin Schumann who helped me writing an intelligible German summary and a short German abstract respectively. I also thank José Martínez Martínez who helped me with the Spanish examples presented in Chapter 6 and Gareth Dwyer who proofread this thesis as a native speaker.

Last but not least, I thank my parents and my family for the support and encouragement they provided me at all times.

Contents

List of Tables

List of Figures

Introduction

The present work investigates the use of automatic language identification methods with a set of standard national language varieties. Automatic language identification, or simply language identification, can be defined as the application of computational methods to identify which language a given document is presented in. State-of-the-art language identification is modelled as a classification problem often relying on n-gram language models at the character and word level (Brown, 2014).

The task is a vital part of the pre-processing in many Natural Language Processing (NLP) applications as will be discussed in this dissertation. As evidenced in Chapter 1, however, very little has been said about the identification of closely related languages, language varieties, and dialects. My work aims to fill this gap by experimenting with different sets of features, languages and language varieties, and algorithms to advance the state-of-the-art in language identification and more specifically to improve the automatic identification of language varieties.

This thesis presents a number of experiments focusing on pluricentric languages such as Portuguese, Spanish, French and English. Pluricentric or polycentric languages are languages which possess more than one standard (national) variety such as Brazilian and European Portuguese or Argentinian and Peninsular Spanish. The distinction between varieties and dialects is not always trivial. In order to define the object of study of this dissertation, Chapter 2 discusses concepts such as pluricentricity, language varieties, dialects, *Ausbausprachen*, *Abstandsprachen*, diglossia, etc. A complete outline of the dissertation is provided in the end of this introduction.

Language identification may be seen exclusively as an NLP task and its origins can be traced back to the work of Ingle (1980). In my work, however, I investigate not only the NLP task itself but also the use of corpora annotated with linguistic information (e.g. morphosyntactic and part-of-speech information) in automatic classification. The use of linguistic features provides quantitative information for researchers in contrastive linguistics about convergence and divergence of languages, dialects, and language varieties. Differences in various levels of human language may be examined by using these methods including morphology, syntax, and the lexicon.

To perform the classification experiments, I used standard contemporary journalistic texts sampled from newspapers published in different countries. As will be discussed in this thesis, I contend that journalistic prose is representative of the standard version of a written national variety. Moreover, the corpora used in this thesis contain texts from different topics and therefore, provided that proper sampling techniques are applied, thematic bias that influences classification can be diminished.

Main Challenges in Language Identification

The most recent language identification studies report performance over 95% accuracy for multi-class classification. For example Simões et al. (2014) report 97% accuracy in identifying a set of 25 languages. However, the authors acknowledge that discriminating between Brazilian and European Portuguese is a difficult task and suggest as future work to 'remove the Brazilian Portuguese and/or merge it with the European Portuguese variant'.

Discriminating between very similar languages is one of the main challenges faced by researchers and developers in the field. To my understanding, at present there are four main directions that research in language identification is taking, as discussed in Zampieri et al. (2015b):

1. **Improving the coverage of language identification systems by increasing the number of languages that systems are able to recognize.**
   To accomplish this, it is neceassary to compile or use existing corpora containing a very large number of languages to train classification algorithms. Some examples include the work by Brown (2012), Xia et al. (2010), and Brown (2013) who trained systems to identify over 900 languages, and Brown (2014) who developed a language identification tool able to discriminate between over 1,300 languages.

2. **Improving the robustness of language identification systems by training systems on multiple domains and various text types.**
   The idea behind this is to identify features that are very discriminative of each particular language regardless of the domain or text type. Lui and Baldwin (2011) investigate the best features across multiple domains by examining the difference in information gain of each feature regarding the language and the domain. Some text types are of course more difficult than others and they deserve special attention, which leads to the third challenge.

3. **Handling non-standard texts (e.g. multilingual texts, computer-mediated communication content, code-switching).**
   Processing non-standard data is a challenge not only for language identification but also for a number of other NLP tasks such as parsing and POS tagging due to the abundance of, for example, non-standard spelling and code alternation

(Owoputi et al., 2013; Carter et al., 2013). This challenge motivated the organization of two recent shared tasks, TweetLID shared task (Zubiaga et al., 2014, 2015), and the shared task on Language Identification in Code-Switched Data (Solorio et al., 2014).

4. **Discriminating between very similar languages, varieties and dialects.**
   Language identification methods achieve very good results when applied to languages which are typologically not closely related (Palmer, 2010), but often fail to deliver high performance when discriminating between closely related languages. When my PhD research started there had been only a few studies on this matter (Ljubešić et al., 2007; Huang and Lee, 2008) and most of them focused on similar languages and not on language varieties or dialects. In the last few years, this task, which is a sub-task of language identification, has received more attention as evidenced by experiments with similar languages (Tiedemann and Ljubešić, 2012), language varieties (Lui and Cook, 2013; Goutte et al., 2016), dialects (Sadat et al., 2014; Malmasi et al., 2015), and the two editions of the DSL shared task which I co-organized (Zampieri et al., 2014, 2015b).

This thesis focuses on the fourth challenge by exploring different methods and features to discriminate between a set of standard national language varieties. An important aspect of this thesis goes beyond the scope of language identification itself. I propose the use of classification methods as a corpus-driven method for contrastive linguistics. Classification algorithms are designed to capture salient features of the input data to be able to classify them with satisfactory performance. Text classification methods have been used to answer a number of questions on language variation related to text types, the style and native language of authors, etc. (Koppel et al., 2002; Herring and Paolillo, 2006; Argamon et al., 2007; Wong and Dras, 2009), but to the best of my knowledge, before the start of my PhD research, these methods have not been applied to study the differences between national language varieties.

The *most informative features* obtained in classification can help scholars to identify important aspects of the data which in turn can explain differences between languages and varieties. In order to identify these differences, I propose the use of different kinds of features, not limited to the traditional character and word n-gram models. In the experiments presented in this thesis, POS and morphologic information were taken into account in classification. By taking word forms out of the classification setting, it is possible to avoid the biases towards named entities and proper nouns that influence classifiers' performance. The use of the most informative features for contrastive linguistics research is discussed in more detail in Chapter 6.

Research Questions

The research questions addressed in this thesis reflect the idea that the research presented here contributes to both computational linguistics and linguistics.

- **RQ1: Is it possible to automatically discriminate between language varieties with satisfactory performance?**
  Before the start of the research that led to this thesis, there have been only a few attempts to answer this question. The work by Huang and Lee (2008) for Chinese texts from Mainland, Taiwan, and Singapore was one such attempt. In this study, researchers conclude that the task is feasible for Chinese varieties using a bag-of-words approach, but is this true for varieties of other languages? To answer this question, I experiment with texts from different pluricentric languages using a number of algorithms and features. Results of up to 99.8% accuracy in discriminating between Brazilian and European Portuguese and 99.0% accuracy between Canadian and Mainland French (both of them are presented in Chapter 4) confirm that the task is feasible for Portuguese, French, and other languages as well.

- **RQ2: Can language varieties be integrated into real-world language identification systems?**
  If we assume that the answer to RQ1 for a given language L is 'yes', can we subsequently integrate varieties of L in a real-world language identification scenario? In other words, should a general-purpose language identification system be trained to recognize varieties of the language L instead of just L? Language identification systems disregard language varieties. The aforementioned remark by Simões et al. (2014) on Brazilian and European Portuguese provides us with an indication of how difficult it is to integrate language varieties in real-world language identification systems. Experiments in Chapter 4 containing up to 17 languages, as well as the results from the two DSL shared tasks, presented in Chapter 5, in which the best systems achieved a performance of above 95% accuracy for a set of 13 languages, indicate that state-of-the-art language identification methods are able to discriminate between language varieties in multilingual settings.

- **RQ3: What are the most efficient features and algorithms to discriminate between language varieties?**
  It is safe to assume, even if the answer to RQ1 is 'no', that some computational methods generally deliver better results for this task than others. Language varieties are very similar to each other and algorithms should be trained to recognize very subtle differences between them. In light of this, what are the best methods and features for this task and how do they relate

to the current research in general-purpose language identification? Results in this thesis indicate that machine learning algorithms such as Support Vector Machines (SVM) and Naive Bayes using characters as features are the methods that achieve the highest performance in this task. I also note that one of the most important differences between language variety identification and general-purpose language identification is that, for the latter, the use of words as features usually does not result in good performance. The results presented in this thesis, particularly in Chapter 4 and Chapter 5, indicate that for language variety identification, in some settings, results obtained using word-based representations are very similar to the ones obtained by character-based methods (e.g. for Portuguese a likelihood estimation method achieved 99.6% accuracy using word unigrams and 99.8% using characters 4-grams).

- **RQ4: Can we use the information obtained from automatic classifiers to study differences between language varieties?**
  There are a number of ways to investigate language variation using corpora. Most of them, both corpus-based and corpus-driven, are designed with a very specific linguistic goal in mind. This includes, for example, the study of lexical variation between language varieties (Peirsman et al., 2010; Soares da Silva, 2010). Text classification, on the other hand, is usually considered as an engineering task where obtaining the best possible performance is the main goal regardless of the kind of linguistic knowledge we may acquire from it. Nevertheless, can we still use the output of automatic classifiers to study language varieties? Are there optimal text representations that are both informative from a linguistic and computational point of view? To answer RQ4 I experimented with a number of text representations using linguistic information and carried out different kinds of analysis on the most informative features in classification. In Chapter 6, using the results obtained in the experiments with Brazilian and European Portuguese, I show that the most informative lexical features obtained in classification are very similar to those that can be obtained by using the keyword lists produced by a number of corpus processing software applications.

Outcomes

This thesis extends the knowledge on language identification and on pluricentric languages in the following ways:

- By using text samples and novel language (variety) settings for language identification. The experiments presented in Chapters 4, 5, and 6 include a number of pluricentric languages, namely: English, French, Spanish, and Portuguese.[1] To the best of my knowledge, discriminating between varieties of these languages was not yet proposed before I began my PhD research.

- By using new features including POS and morphological annotation to represent language varieties. In Zampieri (2013) I introduce the use of POS tags and morphological information as features to discriminate between Mexican and Peninsular Spanish with results of up to 83.1% accuracy. This kind of de-lexicalized approach was later used for the same task by Lui et al. (2014). Although the use of these features does not outperform word- and character-based representations, they are an important source of information for the study of linguistic differences between language varieties.

- By comparing different classification methods for a previously unexplored task. In Zampieri and Gebre (2012) I propose the use of likelihood estimation as simple, yet effective, classification method which achieves almost perfect performance in discriminating between Brazilian and European Portuguese. In Chapter 3, using the same dataset as Tiedemann and Ljubešić (2012), I show that this method achieves good performance comparable to other state-of-the-art methods for discriminating between similar languages. This is also confirmed by the performance obtained by the likelihood estimation entry in the 2015 edition of the DSL shared task presented in Chapter 5.

- By using text classification output to study differences between language varieties within the scope of contrastive linguistics research. Text classification methods have been used to study linguistic variation, but to my knowledge they have not been applied to language varieties. Zampieri et al. (2013) is a first step in this direction. Chapter 6 provides several examples of how corpora and the output of classification methods can be used to study differences between language varieties.

The research that led to this PhD thesis produced a few important resources for the research community:

- A comprehensive and concise overview of the main language identification challenges and particularly of the problem of discriminating similar languages, language varieties and dialects.

- The organization of the two editions of the DSL shared task. The DSL shared tasks were the first initiative of this kind. They featured more than 20 teams

[1] See Chapter 2 for a complete list of all corpora.

participating in each of the two editions and has shed light on different aspects of this task.

- The compilation of the DSL corpus collection (DSLCC) versions 1.0, 2.0 and 2.1. The DSLCC is the first corpus compiled for this purpose available to the research community and it has been used beyond the scope of language identification (e.g. Malmasi and Zampieri (2016) and Martinez and Tan (2016) use the corpus as a resource to train models for the identification of complex words in texts). The corpus was used for educational purposes as part of Computational Linguistics curriculum at Indiana University[2] and in students' projects in Machine Learning and Natural Language Processing courses at Stanford University.[3,4]

- The publication of eight research papers (three of them associated with the DSL shared task), an extended abstract, a survey, and an introductory book chapter on the topic. All publications appeared (or are due to appear) in international peer-reviewed conferences and periodicals. A list of papers is available at the end of this introduction.

Thesis Outline

The content of this thesis is organized in seven chapters as follows:

- **Chapter 1** presents the task of language identification. It begins by defining the task and its modelling as a classification problem. The chapter contains an extensive review of related work on automatic language identification, starting from the early approaches (Ingle, 1980) and going on to discuss state-of-the-art methods beginning with Dunning (1994). Models proposed to discriminate between texts from closely related languages, dialects, and language varieties are also discussed. In addition, the chapter presents two NLP tasks related to language identification, namely native language identification (NLI) and the automatic identification of lexical variation between similar languages and language varieties.

- **Chapter 2** presents the processes of sampling and data collection and describes the linguistic object of this dissertation: standard national language varieties. To precisely define my object of study, I delimit the boundaries between varieties, dialects, and other closely related language systems.

[2] http://cl.indiana.edu/~md7/14/715/

[3] http://cs229.stanford.edu/proj2015/335_report.pdf

[4] http://nlp.stanford.edu/courses/cs224n/2015/reports/24.pdf

The chapter discusses the concepts of pluricentric languages, language varieties, dialects, *Ausbausprachen*, *Abstandsprachen*, *Dachsprachen*, and diglossia among others. The fundamental concepts presented in Chapter 2 allow me to explain the reasons behind the choice of the languages, corpora, and features used in this research.

- **Chapter 3** presents the models and algorithms used in this thesis with emphasis on n-gram language models. Chapter 3 also contains two preliminary experiments carried out to validate the likelihood estimation algorithm and the corpora used in this thesis. The first preliminary experiment compares the performance of the likelihood estimation to another method designed for discriminating between similar languages, the one proposed by Tiedemann and Ljubešić (2012), using the same test set. The second preliminary experiment is used to investigate the variation between the corpora used in this thesis. I compare the results obtained when discriminating between texts from two different countries to the results obtained by discriminating between texts from two newspapers published in the same country. The results show that the algorithm obtains significantly higher performance when discriminating between texts published in two different countries, which indicates that diatopic variation is an important aspect present in the corpora collected for this thesis and it motivates the experiments presented in Chapter 4.

- **Chapter 4** describes several language variety identification experiments carried out using words and characters as features and a set of machine learning algorithms. These experiments model the task as single-label classification (one label is assigned to each text). I experiment with different classification settings, features, languages and corpora. The evaluation is performed with standard NLP and text classification evaluation metrics, namely: accuracy for binary classification settings and precision, recall, and f-measure for multi-class classification settings. The best results are obtained in binary classification. In this chapter I report results of 99.8% accuracy in discriminating between Brazilian and European Portuguese and 99.0% accuracy in discriminating between Canadian and Mainland French.

- **Chapter 5** presents the results of the two editions of the DSL shared task. The DSL shared task is to the best of my knowledge the first shared task of its kind and it fills an important gap in language identification research applied to similar languages and language varieties. It allows researchers to investigate the problem using the first standardized dataset (DSLCC) compiled for this purpose. The results along with descriptions of the systems that participated in the two editions of the DSL challenge are an important source of information for researchers and developers interested in the task.

- **Chapter 6** discusses the results obtained in the experiments using linguistically-motivated features (or knowledge-rich features). As previously mentioned, in this work I use not only traditional character and *n*-gram language models, but also knowledge-rich features that rely on POS tags and morphological information to discriminate between language varieties. The motivation behind the use of knowledge-rich features stems from the need to represent differences that cannot be grasped by either character or word-based representations. This chapter also discusses how the output of the experiments described in this thesis is relevant to contrastive linguistics and presents different ways of representing these features with examples from Portuguese and Spanish.

- **Chapter 7** concludes this dissertation and presents avenues for future research.

Publications

This thesis contains only my original work except where indicated. Citation to all previously published material and results have been included. Tables and images which were previously published were explicitly indicated.

This thesis has the following overlap with the papers I (co-)authored:

- Earlier drafts of sections from Chapter 1 and Chapter 3 were summarized and included in an introductory chapter on language identification (Zampieri, 2016) in the handbook 'Working with Text: Tools, Techniques and Approaches for Text Mining' by Tonkim and Tourte (2016).

- Chapter 4 is based on the following publications: Zampieri and Gebre (2012); Zampieri et al. (2012, 2013); Zampieri (2013).

- Chapter 5 is based on the two DSL shared task reports (Zampieri et al., 2014, 2015b) and my own DSL shared task entry (Zampieri et al., 2015a).

- Chapter 6 contains results published in Zampieri et al. (2013) in Section 6.2.

A complete list of papers which have been published resulting from the research presented in this thesis, including one introductory book chapter and an extended abstract, is presented next:

- Zampieri, M., and Gebre, B. G. (2012). Automatic Identification of Language Varieties: The case of Portuguese. In Proceedings of KONVENS, pp. 233-237. Vienna, Austria.

- Zampieri, M., Gebre, B. G., and Diwersy, S. (2012). Classifying Pluricentric Languages: Extending the Monolingual Model. In Proceedings of the Fourth Swedish Language Technlogy Conference (SLTC), pages 79-80. Lund, Sweden.

- Zampieri, M. (2013). Using Bag-of-words to Distinguish Similar Languages: How efficient are they? In Proceedings of the 14th IEEE International Symposium on Computational Intelligence and Informatics (CINTI), pp. 37-41. Budapest, Hungary.

- Zampieri, M., Gebre, B. G., and Diwersy, S. (2013). N-gram language models and POS distribution for the identification of Spanish varieties. In Proceedings of TALN, pp. 580-587. Sable d'Olonne, France.

- Zampieri, M., and Gebre, B. G. (2014) VarClass: An Open Source Language Identification Tool for Language Varieties. In Proceedings of Language Resources and Evaluation (LREC). pp. 3305-3308. Reykjavik, Iceland.

- Tan, L., Zampieri, M., Ljubešić, N., and Tiedemann, J. (2014). Merging Comparable data sources for the discrimination of similar languages: The DSL corpus collection. In Proceedings of the Workshop on Building and Using Comparable Corpora (BUCC). pp.6-10 Reykjavik, Iceland.

- Zampieri, M., Tan, L., Ljubešić, N., and Tiedemann, J. (2014). A Report on the DSL Shared Task 2014. In Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial), pp. 58-67. Dublin, Ireland.

- Zampieri, M., Gebre, B.G., Costa, H., and van Genabith, J. (2015) Comparing Approaches to the Identification of Similar Languages. In Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial), pp. 66-79. Hissar, Bulgaria.

- Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., and Nakov, P. (2015). Overview of the DSL Shared Task 2015. In Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial), pp. 1-9. Hissar, Bulgaria.

- Zampieri, M. (2016). Automatic Language Identification. In Tonkim, E. and Tourte, G., editors, Working with Text: Tools, Techniques and Approaches for Text Mining. pp. 189-205. Chandos Publishing, Elsevier.

German Summary (Zusammenfassung)

Die vorliegende Arbeit untersucht den Einsatz von automatischen Methoden zur Spracherkennung bei einer Reihe von Standardsprachvarietäten. Automatische Spracherkennung, oder auch einfach Spracherkennung, kann als Anwendung von rechnergestützten Methoden zur Erkennung der Sprache eines Dokuments definiert werden. Moderne Spracherkennung wird als Kassifikationsproblem modelliert und basiert oft auf n-gram Sprachmodellen auf Zeichen- und Wortebene (Brown, 2014).[5]

Spracherkennung spielt bei der Vorverarbeitung in vielen Anwendungen der Sprachtechnologie (Natural Language Processing, NLP), die in dieser Dissertation besprochen werden, eine wichtige Rolle. Wie in Kapitel 1 gezeigt wird, wird in der Literatur die Erkennung von eng verwandten Sprachen, Sprachvarietäten und Dialekten bisher jedoch sehr wenig behandelt. Die vorliegende Arbeit soll diese Lücke schlieen, indem unterschiedliche Merkmale, Sprachen und Sprachvarietäten sowie Algorithmen getestet werden, mit dem Ziel die moderne Spracherkennung und insbesondere die automatische Erkennung von Sprachvarietäten zu verbessern.

In dieser Dissertation werden diverse Experimente zu plurizentrischen Sprachen wie Portugiesisch, Spanisch, Französisch und Englisch durchgeführt. Plurizentische oder polyzentrische Sprache besitzen mehr als eine Standardvarietät (Nationalsprache) wie beispielsweise brasilianisches oder europäisches Portugiesisch, sowie argentinisches oder spanisches Spanisch. Die Unterscheidung zwischen Varietät und Dialekt ist nicht immer einfach. Um den Untersuchungsgegenstand dieser Dissertation zu definieren, werden in Kapitel 2 Konzepte wie Plurizentrizität, Sprachvarietäten, Dialekte, Ausbausprachen, Abstandsprachen und Diglossie diskutiert.

Spracherkennung wird oft als reine NLP-Aufgabe definiert. Die Ursprünge knnen auf die Arbeit von Ingle (1980) zurückverfolgt werden. In meiner Arbeit untersuche ich jedoch nicht nur die reine NLP-Aufgabe, sondern auch den Einsatz von mit linguistischen Informationen annotierten Korpora (z.b. Informationen zu Morphosyntax oder grammatische Kategorien (Part-of-Speech, POS)) für die automatische Klassifikation. Der Einsatz von linguistischen Merkmalen liefert quantitative In-

[5] This summary is a requirement of the Graduate School from Saarland University.

formationen für Wissenschaftler im Bereich der kontrastiven Linguistik hinsichtlich Konvergenz und Divergenz von Sprachen, Dialekten und Sprachvarietäten. Unterschiede auf verschiedenen Ebenen menschlicher Kommunikation können durch den Einsatz solcher Methoden, einschließlich Morphologie, Syntax und Lexik, untersucht werden.

Zur Durchführung von Klassifikationsexperimenten wurden in vorliegender Arbeit zeitgenössische journalistische Texte aus Zeitungveröffentlichungen unterschiedlicher Länder verwendet. Wie in dieser Dissertation diskutiert wird, können journalistische Texte als Standardvariante der geschriebenen Sprachvarietät eines Landes angesehen werden. Des Weiteren umfassen die in dieser Dissertation verwendeten Korpora Texte unterschiedlicher Themenbereiche. Unter Anwendung der richtigen Auswahlmethode soll somit eine Beeinflussung durch thematische Unausgewogenheit verringert werden.

Aktuelle Studien zur Spracherkennung zeigen eine Genauigkeit von mehr 95%. Simoes et al. (2014) zeigen beispielsweise 97% fehlerfreie Erkennung der richtigen Sprache aus 25 verschiedenen Sprachen. Die Autoren besttigen jedoch die Schwierigkeit, brasilianisches und europisches Portugiesisch zu unterscheiden.

Die Unterscheidung von sehr ähnlichen Sprachen ist eine der schwierigsten Herausforderung vor denen Forscher und Entwickler in diesem Bereich stehen. Meines Wissens nach gibt es aktuell vier Hauptforschungsrichtungen im Bereich Spracherkennung, wie in Zampieri et al. (2015b) aufgezeigt:

1. **Verbesserung des Erfassungsbereiches von Spracherkennungssystemen durch eine höhere Anzahl von durch Systeme erkannte Sprachen.**
   Um dieses Ziel zu erreichen ist es notwendig, Korpora mit einer großen Anzahl an Sprachen durch Klassifikationsalgorithmen zu trainieren. Als Beispiele hierfür sind die Arbeiten von Brown (2012), Xia et al. (2010), und Brown (2013) zu nennen, die Systeme zur Erkennung von mehr als 900 Sprachen trainiert haben. Des Weiteren ist Brown (2014) anzuführen, der ein Spracherkennungstool zur Unterscheidung von mehr als 1300 Sprachen entwickelt hat.

2. **Verbesserung der Verlässlichkeit von Spracherkennungssystemen durch Systeme, die mit mehreren Fachgebieten und unterschiedlichen Texttypen trainiert wurden.**
   Der Grundgedanke hierbei ist, spezifische Merkmale einer bestimmten Sprache, unabhängig vom Fachgebiet oder Texttyp, zu identifizieren. Lui und Baldwin (2011) untersuchen die besten Merkmale über verschiedene Fachgebiete hinweg. Sie untersuchen Information Gain für jedes Merkmal hinsichtlich Sprache und Fachgebiet. Für einige Texttypen ist dies natürlich schwieriger als für an-

dere, denen dann besondere Aufmerksamkeit gebührt. Dies führt zur dritten Herausforderung.

3. **Bearbeitung von nicht standardisierten Texten (z.B. mehrsprachige Texte, computer-vermittelter Kommunikationsinhalt, Code-Switching).** Die Verarbeitung von nicht standardisierten Daten ist nicht nur bei der Spracherkennung eine Herausforderung. Auch bei zahlreichen NLP-Arbeitsschritten wie Parsen oder POS-Taggen stellen nicht standardisierte Daten aufgrund einer Vielzahl von nicht standardisierten Schreibweisen oder Wechseln zwischen verschiedenen Codes eine Schwierigkeit dar (Owoputi et al., 2013; Carter et al., 2013). Diese Herausforderung führte in den letzten Jahren zur Durchführung von zwei Shared Tasks: TweetLID Shared Task (Zubiaga et al., 2014, 2015) und Language Identification in Code-Switched (CS) Data (Solorio et al., 2014).

4. **Unterscheidung zwischen sehr ähnlichen Sprachen, Varietäten und Dialekten.** Methoden der Spracherkennung liefern sehr gute Ergebnisse, wenn sie auf typologisch sehr unterschiedliche Sprachen angewandt werden (Palmer, 2010). Bei der Unterscheidung von eng verwandten Sprachen versagen sie jedoch häufig. Zu Beginn meiner Dissertationsforschung gab es nur wenige Studien zu diesem Thema (Huang und Lee, 2008). Die meisten dieser Studien konzentrierten sich zudem auf ähnliche Sprachen, nicht auf Sprachvarietäten oder Dialekte. In den letzten Jahren erhielt dieses Thema, ein Unterbereich der Spracherkennung, mehr Aufmerksamkeit. Ergebnisse aus Studien mit ähnlichen Sprachen (Tiedemann und Ljubesic, 2012), Sprachvarietäten (Lui und Cook, 2013; Goutte et al., 2016), Dialekten (Sadat et al., 2014; Malmasi et al., 2015), und zwei DSL Shared Tasks, die ich mitorganisert habe (Zampieri et al., 2014, 2015b), zeigen dies.

Der Schwerpunkt dieser Dissertation liegt auf der vierten Herausforderung: der Erforschung unterschiedlicher Methoden und Merkmale zu Unterscheidung von einigen Standardvarietäten von Nationalsprachen. Ein wichtiger Aspekt dieser Dissertation geht über die eigentliche Spracherkennung hinaus. Ich schlage den Einsatz von Klassifikationmethoden als *corpus-driven* Methode der kontrastiven Linguistik vor. Klassifikationsalgorithmen sind so ausgelegt, dass sie zentrale Merkmale der eingebrachten Daten mit ausreichender Verlsslichkeit einordnen können. Testklassifikationsmethoden werden zur Beantwortung zahlreicher Fragen zur Sprachvariation hinsichtlich Texttyp, Autorenstil, Muttersprache von Autoren, u.a. genutzt (Koppel et al., 2002; Herring und Paolillo, 2006; Argamon et al., 2007; Wong und Dras, 2009). Vor Beginn meiner Dissertation wurden diese Methoden meines Wissens nach jedoch nicht zur Untersuchung von Unterschieden verschiedener Sprachvarietäten verwendet.

Die aussagekräftigsten Merkmale aus der Klassifikation können Wissenschaftlern helfen wichtige Aspekte der Daten zu erfassen, die bei der Erklärung der Unterschiede zwischen Sprachen und Varietäten helfen können. Zur Ermittlung dieser Unterschiede schlage ich den Einsatz von verschiedenen Merkmalen vor. Diese gehen weiter als die aus der Literatur bekannten, traditionellen n-gram Modelle auf Zeichen- und Wortbasis. Die Experimente in dieser Dissertation nutzen POS sowie morphologische Information bei der Klassifikation. Durch den Ausschluss von Wortformen in der Klassifikation ist es möglich Tendenzen hin zu Eigennamen, die die Leistung der Klassifikatoren beeinflussen, auszuschließen. Der Einsatz der aussagekräftigsten Merkmale in der kontrastiven Linguistik wird in Kapitel 6 genauer beschreiben.

Die Forschungsfragen dieser Dissertation sollen einen Beitrag sowohl zur Computerlinguistik als auch in der Linguistik leistet.

- **FF1: Ist es möglich Sprachvarietten mit ausreichender Verlässlichkeit zu bestimmen?**
  Vor Beginn der Forschungsarbeit, die in diese Dissertation mündet, gab es nur wenige Versuche diese Frage zu beantworten. Die Arbeit von Huang und Lee (2008) für chinesische Texte aus China, Taiwan und Singapur war ein solcher Versuch. Das Ergebnis dieser Studie zeigt, das die Aufgabe für Varietten des Chinesischen mit dem Bag-of-Words-Ansatz lösbar ist. Gilt dies auch fr Varietäten anderer Sprachen? Um diese Frage zu beantworten, experimentiere ich mit Texten verschiedener plurizentischer Sprachen sowie verschiedenen Algorithmen und Merkmalen. Ergebnisse von bis zu 99,8% Genauigkeit bei der Unterscheidung zwischen brasilianischem und europischem Portugiesisch, sowie 99,0% Genauigkeit bei der Unterscheidung zwischen kanadischem und französischem Französisch (beide Ergebnisse werden in Kapitel 4 vorgestellt) zeigen, das diese Aufgabe für Portugiesisch, Französisch und andere Sprachen lösbar ist.

- **FF2: Können Sprachvarietäten in allgemeine kommerzielle Spracherkennungssysteme integriert werden?**
  Wenn wir annehmen, dass FF1 für eine gegebene Sprache S mit *ja* beantwortet werden kann, können wir dann Varietten von S in ein kommerzielles Spracherkennungsszenario integrieren? Anders ausgedrückt: Sollten allgemeine Spracherkennungssysteme darauf trainiert werden, statt nur die Sprache S auch Varietäten von S zu erkennen? Spracherkennungssysteme missachten Sprachvarietäten. Die vorgenannte Studie von Simoes et al. (2014) zu brasilianischem und europäischem Portugiesisch zeigt uns wie schwierig es ist Sprachvarietäten in kommerzielle Spracherkennungssysteme zu integrieren. Die Experimente in Kapitel 4 umfassen bis zu 17 Sprachen. Diese Experimente

sowie die Ergebnisse der beiden DSL Shared Tasks, bei denen die besten Systeme bis zu 95% Genauigkeit bei 13 Sprachen erzielten, zeigen, dass moderne Spracherkennungsmethoden zwischen Sprachvarietäten im mehrsprachigen Kontext unterscheiden können.

- **FF3: Welche Merkmale und Algorithmen eignen sich am besten um zwischen Sprachvarietten zu unterscheiden?**
  Wir können annehmen, dass, selbst wenn die Antwort auf FF1 *nein* ist, einige computergestützte Methoden bei dieser Aufgabe gewöhnlicherweise bessere Ergebnisse liefern als andere. Sprachvarietäten sind einander sehr ähnlich. Algorithmen sollten darauf trainiert werden, schon geringe Unterschiede zu erkennen. Welches sind vor diesem Hintergrund die besten Methoden und Merkmale zur Unterscheidung von Sprachvarietäten und wie werden diese in aktueller Forschung zu allgemeiner Spracherkennung berücksichtigt? Die Ergebnisse dieser Dissertation weisen darauf hin, dass Algorithmen wie Suppert Vector Machines (SVM) und Naive Bayes, die Zeichen als Merkmale nutzen, die besten Ergebnisse liefern. Ein weiteres Ergebnis meiner Untersuchungen zeigt, dass der Hauptunterschied zwischen Erkennung von Sprachvarietäten und allgemeiner Spracherkennung beim Einsatz von Wörtern als Merkmalen liegt: Bei allgemeiner Spracherkennung liefern Wörter als Merkmale normalerweise keine guten Ergebnisse. Die Ergebnisse dieser Arbeit zeigen, dass bei der Erkennung von Sprachvarietäten, je nach Einsatz, wortbasierte Reprsentationen ähnliche Ergebnisse liefern wie zeichenbasierte Methoden (z.B. erzielt die Likelihood-Methode fr Portugiesisch 99,6% Genauigkeit bei der Verwendung von Wortmonogrammen und 99,8% für 4-Gramme).

- **FF4: Können die Ergebnisse aus der automatischen Klassifizierung genutzt werden, um Unterschiede zwischen Sprachvarietäten zu untersuchen?**
  Sprachvariation kann mithilfe von Korpora in unterschiedlicher Art und Weise untersucht werden. Die meisten Ansätze, sowohl korpusbasierte als auch corpus-driven Methoden, verfolgen eine spezielle linguistische Fragestellung. So auch die Untersuchungen zur lexikalischen Variation von Sprachvarietäten (Peirsman et al., 2010; Soares da Silva, 2010). Textklassifikation wird hingegen normalerweise als technische Aufgabe angesehen, bei der es darum geht, das beste Ergebnis zu erzielen, ohne weitere Berücksichtigung von linguistischem Wissensgewinn. Kann der Output von automatischen Kassifikatoren dennoch zur Untersuchung von Sprachvarietäten genutzt werden? Gibt es optimale Textrepräsentationen, die sowohl für die Linguistik als auch für den Bereich der Informatik informativ sind? Um FF4 zu beantworten wurden zu einigen Textrepräsentationen mit linguistischen Informationen Experimente und

zu den aussagekräftigsten Klassifikationsmerkmalen unterschiedliche Analysen durchgeführt. Mit den Ergebnissen aus den Experimenten mit Brasilianisch und Portugiesisch wird in Kapitel 6 gezeigt, dass die aussagekräftigsten lexikalischen Merkmale aus der Klassifikation und die Ergebnisse aus Versuchen mit Keywordlisten, generiert durch verschiedene Korporaanalysetools, vergleichbar sind.

Chapter 1

Language Identification

1.1 Introduction

The first chapter of this thesis presents the task of automatic language identification. Language identification methods are vital to many NLP and information retrieval applications. They are necessary, for example, to aid document collection creation in scenarios where the languages of documents are not known. A clear example are documents crawled from the Internet which are often unlabelled regarding their language and language identification methods are applied to identify the language of each document (Ljubešic et al., 2014). Machine translation (MT) also benefits from language identification methods, because to translate a document to a target language, it is first necessary to determine its source language. These features are present in different translation tools and web-browsers such as *Google Chrome*. Finally, corpus creation for low-resource languages is another area that uses language identification methods (Scannell, 2007; Emerson et al., 2014).

The outline of this chapter is as follows. I first define the task as a classification problem in which texts receive labels, each of them corresponding to the language of the text. I then provide a historical overview of the main language identification approaches leading up to the most recent state-of-the-art methods in use today. I describe in more detail three methods, each of them influential in different decades. Ingle's short word approach (Ingle, 1980), the Out-of-place Metric by Cavnar and Trenkle (1994), and the approach behind *langid.py* (Lui and Baldwin, 2012) which uses machine learning and Information Gain (IG) to estimate the best features for language identification. These three approaches were selected due to their impact in different stages of the development of language identification research.

A subsection is dedicated to reviewing previous research on the identification of similar languages, language varieties and dialects, which is the main focus of this thesis. I present a number of approaches that deal with the discrimination between similar languages and dialects such as: Malay and Indonesian; Serbian, Bosnian, and Croatian; and Arabic dialects.

As will be evidenced in this literature review, two of the four research questions of this thesis, namely **RQ1:** Is it possible to automatically discriminate between language varieties with satisfactory performance? and **RQ2:** Can language varieties be integrated into real-world language identification systems? have not been completely answered before the start of the research that led to this thesis. The results that will be presented in Chapter 4 and Chapter 5 answer these questions. One of the previous attempts to address **RQ1** in the literature can be traced back to Huang and Lee (2008) who applied computational methods to discriminate between three Chinese varieties: Mainland, Taiwan, and Singapore. To my knowledge, no studies before the first DSL shared task in 2014 explored the integration of language varieties into multilingual settings **(RQ2)**.

At the end of this chapter, I present two tasks related to language identification, namely native language identification (NLI) and automatic identification of lexical variation between similar languages and varieties. NLI is a text classification task that consists of the identification of the native language of an author based on his/her production of second language text. The task is by no means trivial as the algorithms have to learn intrinsic properties of texts to be able to classify them accurately. The interest in NLI has been growing over the last few years mostly for English, but also for languages with a large number of speakers. The subsection that closes this chapter is concerned with previous work on lexical variation.

Chapter 1 is organized as follows:

- **Section 1.2** presents and defines the task of automatic language identification.

- **Section 1.3** formalizes the task of language identification as a single-label multi-class classification problem based on the research that is described in Medlock (2008).

- **Section 1.4** presents a historical overview of the task and of several approaches that have been proposed in the last decades (from early approaches such as the one by Ingle (1980) to state-of-the-art approaches (Lui and Baldwin, 2012)).

- **Section 1.5** discusses the problem of identifying similar languages, language varieties, and dialects.

- **Section 1.6** presents two tasks related to language identification. The first of them is Native Language Identification (NLI) and the second involves computational approaches to lexical variation such as cognates and false friends between similar languages, varieties, and dialects.

- **Section 1.7** summarizes the content of this chapter.

1.2 Language Identification and Language Varieties

Automatic language identification or simply language identification can be broadly defined as the task of automatically identifying the language(s) contained in a given document.[6] This task is a well-known research topic in computational linguistics and its origins can be traced back to the work of Ingle (1980). In this chapter I give a concise historical overview of the task from early approaches to more current research. I focus on the identification of similar languages as one of the challenges of language identification and discuss this aspect in more detail.

There are a number of situations in which the source language of a document is unknown and computational methods can be applied to determine it. This makes language identification a relevant task that can be integrated with most NLP applications such as MT (for cases in which knowing the source language or language variety of a document is vital for further processing) or Information Retrieval (IR) (e.g. English terms in the domain of Informatics are used in a number of languages, but only documents of a particular language are relevant for a given search).

As discussed by Lui (2014) language identification systems are typically divided into four main steps. Given a set of documents written in different languages the system will implement the following:

1. Data representation: select a text representation (e.g. characters, words, or a combination of the two);

2. Language modelling: calculate or derive a model from documents known to be written in each language;

3. Classification function: define a function that best represents the similarity between a document and each language model;

4. Prediction or output: compute the highest-scoring model to determine the language of the document.

State-of-the-art methods for language identification apply n-gram-based language models at the character or word level to distinguish a set of languages automatically. The most successful approaches model the task as a supervised single-label

[6] In this chapter and throughout the thesis I will be exclusively discussing language identification methods applied to text. There are several systems that perform the same task for speech data, including studies that discriminate varieties of the same language such as Koller et al. (2010) for Portuguese. However, although methods applied to text and speech are often similar, due to the scope of this dissertation, speech processing applications will not be discussed.

classification (one label for each document) problem. The average multi-class accuracy obtained by these methods is usually over 95% (Lui and Baldwin, 2012; Brown, 2013).

As stated in Palmer (2010), it is very common for language identification methods to perform perfectly (or almost perfectly) when distinguishing between languages which are typologically not closely related (e.g. Finnish and Portuguese or Bulgarian and Spanish) as well as when recognizing languages with unique character sets such as Hebrew. The difficulty therefore lies in discriminating between similar languages and languages that use similar character sets.

> For languages with a unique alphabet not used by any other languages, such as Greek or Hebrew, language identification is determined by character set identification. Similarly, character set identification can be used to narrow the task of language identification to a smaller number of languages that all share many characters such as Arabic vs. Persian, Russian vs. Ukrainian, or Norwegian vs. Swedish (Palmer, 2010, p.13).

This explains the success obtained by most state-of-the-art general purpose language identification methods that will be presented in this chapter. Brown (2013), for example, describes a system trained to identify texts from 1,100 languages based on character n-grams with results reaching 99.2% accuracy in a multi-class setting. These methods work with a large language set and most of the languages are typologically unrelated. For example, it is extremely unlikely that a state-of-the-art language identification method would label French texts as Persian or Japanese as Italian, but it might, in some cases, tag a Portuguese text as Spanish.

At this point it is not difficult to recognize a number of challenges faced by the language identification systems. One of them is the identification of closely related languages that share similar character sequences and lexical units (e.g. Croatian and Serbian and to a lesser extent Portuguese and Spanish or Danish and Swedish). Problems that systems face when discriminating similar languages also occur when discriminating between language varieties and dialects.[7] For this reason, Section 1.5 will concentrate on experiments aimed at discriminating between similar languages, varieties and dialects including some recently completed shared tasks. Even though there have been a few attempts to distinguish varieties and dialects in the literature, up to now this aspect of language identification has not received much attention. Pluricentric languages, for example, are often modelled as a unique class and no distinction is made between language varieties.

[7] Standard national variety is defined in more detail in Chapter 2. For NLP systems, however, the methods are the same as to those applied to similar languages and dialects. For this reason, Section 1.5 discusses similar languages, varieties, and dialects as the same problem.

Another challenge faced by language identification systems is identifying the language of short excerpts from texts particularly those containing non-standard language. Systems have difficulty when confronted with small excerpts from texts (a few words) that do not provide enough data for algorithms to classify them correctly. This difficulty is even more evident for texts available on the Internet, such as *tweets*, because these are often noisy and contain non-standard spelling and/or code alternation (Nguyen and Dogruoz, 2013). Methods designed in recent years specifically to deal with these two aspects of language identification will be presented in this chapter.

1.3 A Classification Problem

Language identification is a text classification task which follows the aforementioned steps described by Lui (2014). Text classification 'is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set' (Sebastiani, 2005). Classes are represented by a finite set of labels. In language identification, documents are texts whose source language is unknown and the finite set of labels is the set of languages (varieties and dialects) that the system is able to recognize.

The formal definition of automatic classification is the computational process of assigning class labels to objects. A class is defined as a finite set of objects, each represented by a unique class label which is an arbitrary descriptor for the object.

Formally, single-label classification can be represented by the following function, adapted from Medlock (2008):

$$f_{class} : \chi \to \lambda \tag{1.1}$$

In Function 1.1, χ is the sample space and λ is a set of class labels. The classification function assigns a label $y \in \lambda$, which in the case of language identification is the set of languages that the method is trained to recognize, to all instances of a given dataset.

The type of classification used for language identification is called single-label classification. This kind of classification allows only one label to be attributed to each instance. The language of a text can be either English or German or Portuguese or Chinese, but not two or more of them at the same time. The term single-label classification is used in contrast to multi-label classification which allows more than one label to be attributed to each instance.[8]

[8] Language identification applied to multilingual documents may be modelled as multi-label classification allowing more than label (language) to be attributed to each instance (text). In

Conceptually, language identification is not different to other text classification tasks (e.g. text categorization and spam detection) as discussed in Sebastiani (2002) and Medlock (2008). The kinds of features and algorithms used differ, but the idea behind the task is the same.

Although language identification is usually modelled as a single-label classification task, it is also possible to imagine a scenario in which language identification applied to monolingual documents is modelled as a multi-label and multi-class classification. Multi-label classification allows at least two levels of classification, the first being the language (e.g. English, Portuguese, or French) and the second being the variety (e.g. British, American, or Canadian). This classification could, in theory, be used in cases in which it is paramount to know the language of the document and desirable but not essential to identify the language variety.[9]

1.4 Historical Overview

In this section I will present a historical overview of language identification approaches published over the years starting from general-purpose methods and proceeding to methods designed to discriminate between similar languages.

The study published by Ingle (1980) is the first well-known attempt to solve the language identification problem. Ingle applied Zipf's law to order the frequency of short words in texts and used this information to perform language identification.[10] The study published by Beesley (1988) is regarded to be the first to use character n-grams for language identification.[11] The basic intuition behind Beesley's approach is similar to the example provided in Section 1.2:

> For example, the probability of *TH* occurring in English is relatively high, but in Spanish or Portuguese the probability approaches zero; the probability of *SZ* is relatively high in Polish and Hungarian but low in English, French, Spanish, and Portuguese. Such information can be

this thesis, I use only monolingual documents and therefore I approach the task as a single-label classification problem.

[9] In Chapter 5 I present the results of the two editions of the DSL shared task. The best performing systems in the DSL shared tasks used a two-stage approach such as the one described here.

[10] Some studies mention Gold (1967) as the pioneer paper in language identification. However, the notion of language identification discussed by Gold focuses on learnability rather than on pure language classification or discrimination. In this thesis I consider the work of Ingle (1980) as the first systematic automatic language identification study.

[11] Adams and Resnik (1997) cite Dunning (1994) as the first to propose n-gram methods for language identification.

> quantified precisely for a representative corpus of the language, and one can even go on to compute probabilities for 3-grams, 4-grams, etc (Beesley, 1988, p.7).

Among the pioneers in the use of n-grams for language identification there is Dunning (1994). The study published by Dunning (1994) reports over 99% accuracy in distinguishing between English and Spanish texts by calculating the likelihood of character n-grams using Markov models, and applying Bayesian decision rules to minimize errors. After Dunning, n-gram language models became the basis of almost the vast majority of language identification systems.

Automatic language identification followed the trend observed in the field of computational linguistics from the beginning of the 1990s in which statistical language modelling and stochastic methods became more popular than symbolic approaches. This paradigm shift in computational linguistics can be observed in different NLP tasks and applications such as statistical machine translation systems and parsing.

One of the most widely cited n-gram-based methods often used as baseline for state-of-the-art language identification systems is that by Cavnar and Trenkle (1994). Cavnar and Trenkle (1994) applies n-gram methods that make use of a list of the most frequent character n-grams in different corpora and calculates what the authors refer to as the 'out-of-place' metric. This approach will be explained in more detail later in this chapter. TextCat[12] is a tool that implements this approach. TextCat contains language models for 76 languages and can be adapted or customized to a user's needs as it allows users to train the system with their own data.

A few comparative studies have been published on language identification, one of them is Grefenstette (1995). Grefenstette (1995) compares two language identification methods: a trigram approach similar to the one published by Beesley (1988) and Cavnar and Trenkle (1994) and the frequent (short) word approach proposed by Ingle (1980). Based on these experiments, the author highlights the simplicity of both methods and the advantage of character-based approaches when dealing with short texts (those comprising fewer than 15 words). According to Ingle (1980), shorter sentences are section headings and titles which might not contain any of the short words used in Ingle's approach.

Other comparative studies include the one by Vojtek and Belikova (2007) which compares two language identification methods based on Markov processes, such as the method proposed by Dunning (1994). Padró and Padró (2004) compare the performance of three language identification methods: Markov models, n-gram based text categorization (Cavnar and Trenkle, 1994), and trigram frequency vectors. Gro-

[12] `http://odur.let.rug.nl/vannoord/TextCat/`

ethe et al. (2008) compare language identification methods based on three features: character n-grams, frequent words, and short words.

A number of classification algorithms have been proposed for language identification. Examples include Monte Carlo sampling (Poutsma, 2001), Markov-based methods (Xafopoulos et al., 2004), and machine learning methods which have been widely used in language identification. Combrinck and Botha (1994) proposed the use of machine learning as an alternative to Markov-based approaches. Takçı and Güngör (2012) applied a centroid-based classification approach, widely used in text classification reporting results of 97.5% accuracy. Although most language identification studies involve supervised learning, there have been a few attempts to perform language identification using unsupervised methods, for example the work by Amine et al. (2010) which proposes a hybrid unsupervised method that includes the popular k-means clustering algorithm (de Amorim and Mirkin, 2012).

The Internet is a very interesting application for language identification and also one of the most challenging scenarios for state-of-the-art systems. This is mainly because documents available on the Internet are often unidentified regarding source language which makes the use of language identification methods a vital part of most applications developed to process Internet data. Moreover, individual documents may contain more than one language and, particularly in the case of user-generated content, texts often contain non-standard spelling, making it difficult for NLP applications to process them.

In recent years, a number of language identification methods were proposed to identify the language of webpages and microblog posts, including Martins and Silva (2005), Rehurek and Kolkus (2009), Chew et al. (2011), Tromp and Pechnizkiy (2012), and Vogel and Tresner-Kirsch (2012). The two last methods target short Internet texts (e.g. microblog posts and *tweets*) using the LIGA algorithm. Ceylan and Kim (2009) aim to identify languages of short query texts input by users in search engines. The authors use logs from *Yahoo!* to train machine learning algorithms for this purpose. They train two decision tree classifiers: one that uses only linguistic features and another that includes non-linguistic features.

Martins and Silva (2005) propose a method to identify web pages from 12 languages (Danish, Dutch, English, Finnish, French, German, Italian, Japanese, Portuguese, Russian, Spanish, and Swedish). The results reported by Martins and Silva (2005) varied by language, ranging from 80% accuracy for Italian and 99% accuracy for English. Their study provides an example of one of the aforementioned challenges of language identification: the identification of closely related languages. The performance obtained when identifying Italian is particularly representative of this difficulty; among 500 texts classified, 20 were tagged as Portuguese and 42 were labelled as Spanish. Italian, Portuguese, and Spanish are all Romance languages and this is why algorithms have difficulty classifying Italian documents.

More recent language identification studies include Lui and Baldwin (2012), who developed a tool called *langid.py*.[13] The system contains language models for 97 languages, using various data sources such as the EMEA biomedical corpus, the EuroPARL Corpus and Wikipedia. The study reports 91.3% accuracy for a set of 67 languages using Wikipedia data. The tool is off-the-shelf and can be tailored to a user's needs by training the algorithm to identify new classes (languages). Brown (2013) applies language identification methods to a collection of documents written in 1,100 languages (later extended to over 1,300 languages in Brown (2014)). Each document contained at most 65 characters and the performance of this algorithm reached 99.2% accuracy using smoothing and 98.2% without smoothing.

In the following sections, I discuss in more detail three well-known language identification approaches, namely the approach proposed by Ingle (1980), the out-of-place metric proposed by Cavnar and Trenkle (1994), and *langid.py* Lui and Baldwin (2012). The focus on these approaches is due to the impact they have had on the development of language identification methods.

1.4.1 Ingle's Short Word Approach

Ingle (1980) proposed a language identification method that relies on the frequency of short words, arguing that these can be good features for language identification systems. Short words are often grammatical words such as determiners, conjunctions and prepositions and they appear very frequently in natural language corpora. It is very unlikely that long words will ever be as frequent as short words in any corpus. As an example, Table 1.1 presents the most frequent words in the English collection of the Project Gutenberg.[14]

Rank	Word	Rank	Word
1	the	11	with
2	of	12	is
3	and	13	it
4	to	14	for
5	a	15	as
6	in	16	had
7	that	17	you
8	was	18	not
9	he	19	be
10	his	20	on

Table 1.1: Most frequent words in Project Gutenberg

[13] `https://github.com/saffsd/langid.py`

[14] `http://www.gutenberg.org/`

The table shows that nearly all 20 most frequent words are short words with a maximum of three characters, except *that* and *with*. Among the 100 most frequent words, all are shorter than 5 characters up to position 87 which is held by the 6 character word *before*. The idea behind Ingle's approach is that short words are highly discriminative because the frequency of words in a corpus follows a Zipfian distribution (Zipf, 1949).

According to Zipf's law, for every corpus, the frequency of a word is inversely proportional to its rank in a frequency list. Zipf's law also indicates that short words are on average more frequent than long ones because humans try to optimize communication[15] using the 'Principle of Least Effort'. For example, a personal pronoun in its first person singular form, which is used very often in a language, would probably never be more than two or three characters long (e.g. *Ich* DE, *Io* IT, *I* EN, *Je* FR,*eu* PT). There is some criticism of this theory, as in Piantadosi (2014), but nonetheless the large-scale distribution of words in a language is 'robustly Zipfian' as argued by Piantadosi.

> First, the method of plotting word frequency distributions has obscured an important fact: word frequencies are not actually so simple. They show statistically-reliable structure beyond Zipfs law that likely will not be captured with any simple model. At the same time, the large-scale structure is robustly Zipfian (Piantadosi, 2014, p.18).

As another example, Grefenstette (1995) presents a short word list calculated from the ECI corpus. Table 1.2 presents the top 5 words for Dutch, English, French, German, Portuguese and Spanish.

Dutch	**English**	**French**	**German**	**Portuguese**	**Spanish**
de	the	de	der	de	de
van	and	la	die	a	la
het	to	le	und	que	que
een	of	et	den	o	el
en	a	des	in	e	en

Table 1.2: Top five short words presented by Grefenstette (1995)

Although the frequency of many short words is very high, which makes them good features to discriminate between languages, one can see a limitation of this method by looking at Table 1.2. Approaches based on short words have difficulty in providing

[15] It should be noted that Piantadosi et al. (2011) looks at the relationship between word length and frequency and conclude that information content is a more important predictor of word length than frequency.

information to distinguish between closely related languages. French, Portuguese, and Spanish are Romance languages and the most frequent word in these three languages is *de*. The list of Portuguese and Spanish five most frequent words, for example, both feature *que*, whereas the lists containing French and Spanish most frequent words share the word *la*.

1.4.2 Cavnar and Trenkle's Out-of-place Metric

The language identification approach proposed by Cavnar and Trenkle (1994) is conceptually a ranking method that relies on what the authors called the 'out-of-place metric'. The metric establishes n-gram profiles and calculates statistics that determines how far out of place an n-gram in one profile is from its place in the given category. The basic idea can be understood by looking at the example presented in Figure 1.1.

According to Cavnar and Trenkle's description, the n-gram 'AND' is at rank 5 in the document, but at rank 6 in the category (language), and therefore 1 rank out of place. If an n-gram (e.g. 'IT') is not in the category profile, it takes an arbitrarily defined maximum out-of-place value. The sum of all of the out-of-place values for all n-grams is the distance between the document and the category. The algorithm then applies a function, that the authors call 'Find Minimum Distance', to perform classification. 'Find Minimum Distance' takes the distance measures from all of the category profiles to the document profile, and picks the smallest one.

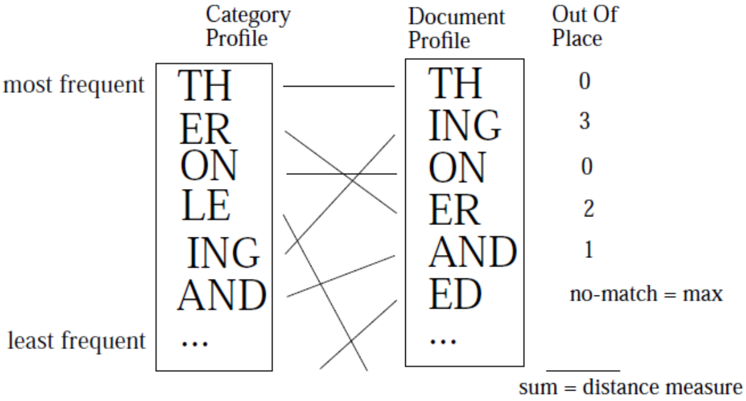

Figure 1.1: Calculating the Two Out-of-Place Measure (Cavnar and Trenkle, 1994)

The system was used to classify documents from 14 countries written in the following eight languages: Dutch, English, French, German, Italian, Polish, Portuguese, and Spanish. The authors report an overall performance of 99.8% accuracy. The approach by Cavnar and Trenkle (1994) is often used as baseline performance for state-of-the-art language identification methods.

1.4.3 Lui and Baldwin's *langid.py*

The *langid.py* (Lui and Baldwin, 2011, 2012) is a readily available general-purpose language identification tool. It achieves results of up to 94.7% accuracy, outperforming similar tools such as *TextCat* (Cavnar and Trenkle, 1994) and *GoogleAPI* on a standard dataset (e.g. Wikipedia and Europarl) and on a dataset containing microblog messages.

The approach uses a multinomial Naive Bayes classifier combined with information gain (IG) for feature selection to minimize the impact of the topic influence in language identification. As described in Lui and Baldwin (2011), this approach considers the IG of particular n-gram features among the set of all languages, within a given language, and within the domain the data was obtained from (e.g. Wikipedia, newspaper texts, microblogs). The method applies this information to identify features that best help the system to identify a language regardless of the domain. The best features for a given language typically have high IG with respect to language but low IG with respect to domain. A study evaluating the use of *langid.py* in the context of language modelling is presented in Cook and Lui (2014).

The performance of *langid.py* was further tested when it was selected to be one of the algorithms used by the UniMelb-NLP team in the Discriminating between Similar Languages (DSL) shared task (Lui et al., 2014). The UniMelb-NLP team was ranked fourth (out of eight systems) in the closed submissions and first (out of two systems) in the open submission. The researchers compare different approaches and a number of resources for language identification applied to similar languages and language varieties. The authors report that *langid.py* modelled hierarchically in two-levels (first identifying the language group and then the language or variety) achieving the best performance among all methods tested.[16]

1.5 Similar Languages, Varieties and Dialects

Although general purpose methods for automatic language identification have been substantially explored, the same is not true for methods designed to deal specifically with similar languages or varieties. The identification of closely related languages seems to be the weakness of most n-gram based models and the interest of the NLP community in this problem has been growing in the last few years, as evidenced by recent studies starting with Ljubešić et al. (2007).

Along with the recently published studies that will be discussed in this section, the growth of interest in varieties and dialects within the NLP community is evidenced by recent events held at international NLP conferences. These events include

[16] See Chapter 5 for more details.

the DIALECTS workshop[17] at the 2011 edition of EMNLP, 'Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants' held at RANLP 2013[18], LT4CloseLang[19] at EMNLP 2014, the VarDial[20] workshop at COLING 2014, and the most recent event, LT4VarDial[21] held at RANLP 2015. The latter two workshops hosted the two editions of the DSL shared task, which I co-organized, and which will be described in detail in Chapter 5.

The study by Ljubešić et al. (2007) is among the first to focus on the discrimination between texts from similar languages. Ljubešić et al. (2007) propose a computational model for the identification of Croatian texts in comparison to other Slavic languages. This study reports 99% recall and precision in three processing stages. The last stage includes a 'black list', a list of forbidden words for Serbian and Croatian. The study published by Tiedemann and Ljubešić (2012) improves this method and applies it to Bosnian, Serbian and Croatian texts.

Ranaivo-Malançon (2006) presents a semi-supervised character-based model to distinguish between Indonesian and Malay, two closely related languages from the Austronesian family. The study uses three sets of features: 1) the frequency and rank of character trigrams derived from the most frequent words in each language; 2) a list of exclusive words; and 3) the format of decimal numbers (Indonesian uses comma whereas Malay uses a dot). The author compares the performance obtained by this method with the performance obtained by TextCat.

Huang and Lee (2008) present a bag-of-words approach to discriminate between Chinese texts from the mainland, Singapore, and Taiwan with results of up to 92% accuracy. Trieschnigg et al. (2012) describe a classification experiment for Dutch dialects from the Dutch Folktale Database, which also contains historical texts. Researchers report micro average f-measure results of 79.9% with the best f-measure result reaching 98.7% for one of the classes.

A few studies have been published and are included as part of this PhD thesis. Examples include Zampieri and Gebre (2012) on Portuguese varieties and Zampieri et al. (2013) on Spanish varieties. In the first paper, we propose the likelihood estimation method to identify two varieties of Portuguese (Brazilian and European). The approach was trained and tested in a binary setting using journalistic texts, with accuracy results above 99.5% for character *n*-grams. The algorithm was later adapted to classify Spanish texts using not only word and character *n*-grams but also POS distribution (Zampieri et al., 2013).

[17] `http://www.ofai.at/~dialects2011/`

[18] `http://c-phil.informatik.uni-hamburg.de/view/Main/RANLPLangVar2013`

[19] `http://www.c-phil.uni-hamburg.de/view/Main/LTforCloseLang2014`

[20] `http://corporavm.uni-koeln.de/vardial`

[21] `http://ttg.uni-saarland.de/lt4vardial2015/`

Among recent studies, Lui and Cook (2013) proposes a method to distinguish between Australian, Canadian, and British English. This study investigates the performance of a classifier across different domains and the results obtained suggest that the characteristics of each variety are consistent across all domains. Malmasi and Dras (2015a) apply SVM classifiers to discriminate between Persian and Dari texts. Ciobanu and Dinu (2016) use text classification methods to discriminate between Romanian dialects and Hollenstein and Aepli (2015) propose the use of character n-grams to discriminate between five Swiss German dialects (Aarau, Basel, Bern, Ostschweiz, and Zurich).

In recent years there has been a significant increase of interest in the computational processing of Arabic. This is evidenced by a number of research papers on several NLP tasks and applications including the identification of Arabic dialects (Elfardy and Diab, 2013; Zaidan and Callison-Burch, 2014; Malmasi et al., 2015), machine translation of Arabic dialects (Zbib et al., 2012; Sajjad et al., 2013; Salloum and Habash, 2013), and the compilation of Arabic dialectal corpora (Zaidan and Callison-Burch, 2011; Cotterell and Callison-Burch, 2014; Mubarak and Darwish, 2014). Arabic is particularly interesting for researchers of language variation due to the fact that the language is often in a diaglossic situation in which the standard form called Modern Standard Arabic (MSA) coexists with a number of regional dialects used in everyday communication.

Among the studies published on this topic, Elfardy and Diab (2013) propose a supervised approach to distinguish between Modern Standard Arabic (MSA) and Egyptian Arabic which achieved up to 85.5% accuracy. The proposed approach discriminates between MSA and Egyptian Arabic at the sentence level using the Arabic online commentary dataset. A more recent study that achieved higher accuracy results using the same dataset is the one by Tillmann et al. (2014). In this study authors proposed a linear SVM classifier and report 89.1% accuracy. Finally, Salloum et al. (2014) explores the use of sentence level Arabic dialect identification for machine translation. The authors report an improvement of 1.0% BLEU score compared to a baseline system. This extrinsic evaluation is a good example of how the task of identifying dialects or language varieties can be integrated into different NLP applications in order to increase performance.

1.5.1 Shared Tasks

Before the organization of the first DSL shared task presented in Chapter 5, shared tasks in language identification focused mostly on general-purpose language identification or on other aspects of the task rather than on discriminating similar languages or language varieties. This is evidenced by a number of shared tasks organized such as the ALTW language identification shared task (Baldwin and Lui, 2010) focusing on general-purpose language identification, the *tweetLID* shared task (Zubiaga et al., 2014) which is concerned with user-generated content using *Twitter* data, and finally the shared task on language identification in code-switched data (Solorio et al., 2014).

One of the few examples of shared tasks involving the identification of language varieties is the 2010 DEFT challenge. Unlike the DSL task, the DEFT shared task was restricted to French texts. Participating teams received a training corpus comprising journalistic texts published in different francophone countries and with different publication dates (Grouin et al., 2010). Systems were trained to answer the following questions:

1. Where was the text published?
2. When was the text published?
3. In which newspaper was the text published?

Therefore the DEFT 2010 shared task involves the following tasks: 1) the identification (or discrimination) of language varieties, 2) temporal text classification, and 3) the identification of the medium in which the text was published. The first task is the subject of this thesis and I discuss the identification of French varieties from Canada and France in Chapter 4. The second task, temporal text classification, along with the recognition of time specific expressions were the subject of a recent SemEval shared task (Popescu and Strapparava, 2015) entitled 'Diachronic Text Evaluation (DTE)'. The third task is the most difficult of the three and to the best of my knowledge it has not been the subject of other related shared tasks. Even though newspapers might have specific traits (words, topics, etc.) that allow algorithms to identify them automatically, one can assume that the differences between texts published by different newspapers in the same year and in the same country are not particularly prominent.[22]

1.6 Related Tasks

This final section of this chapter presents two related NLP tasks. The first one shares substantial overlap with language identification and the second one consists of the application of corpora for the study of lexical variation. The first task is called Native Language Identification (NLI). This task consists of identifying the language of a writer based on his second language production assessed through written texts. The methods applied to NLI are similar to those applied to language identification

[22] In chapter 4, I carried out a controlled experiment and trained an algorithm to differentiate between two newspapers published in the same year and in the same country (in Spain and in England). The poor results obtained give us an indication on how difficult this task is.

but the task is more challenging than language identification because it relies on the identification of very subtle features in text, often not perceived by humans, that may reveal information about the writer of the text and his or her native language.

1.6.1 Native Language Identification

NLI is the task of automatically identifying the native language of a writer based on the writer's foreign language production. The task is by no means trivial and is based on the assumption that the mother tongue influences second language acquisition and production (Lado, 1957).

When an English native speaker hears someone speaking English, it is not difficult for him to identify whether this person is a native speaker or not. Moreover, it is, to some extent, possible to assert the mother tongue of non-native speakers by their pronunciation patterns, regardless of their language proficiency. In NLI, the same principle that seems intuitive for spoken language, is applied to text. If it is true that the mother tongue of an individual influences speech production, it should be possible to identify these traits in written language as well.

NLI methods can be used both to discriminate between native and non-native texts as well as to determine the native language or language family (e.g. Slavic, Germanic, Romance) of an individual. The task is often regarded as part of the broader task of authorship profiling (Rangel et al., 2013). Authorship profiling methods try to assert attributes of an author such as native language, age (Nguyen et al., 2013), gender (Cheng et al., 2011), and even income (Preoţiuc-Pietro et al., 2015) by identifying patterns in texts.

Several studies on NLI have been published in recent years, using a variety of methods and approaches. Many of these approaches overlap with recent work on language identification and similar language identification and they are therefore worth discussing. This section summarizes a few important studies on this task. A compreheensive overview on the task methods can be found in Malmasi (2016).

Examples of NLI approaches include Tomokiyo and Jones (2001) who proposed a Naive Bayes classifier to discriminate between native and non-native English texts written by Chinese and Japanese speakers, Koppel et al. (2005) who applied machine learning to discriminate between five native languages using the International Corpus of Learner English (ICLE) (Granger et al., 2009), Tsur and Rappoport (2007) who investigated the influence of mother tongue' phonology in NLI, and finally Kochmar (2011) who investigated the influence of different features in NLI proposing an approache able to discriminating between native speakers from Germanic and Romance languages with 84.35% accuracy.

The NLI Shared Task[23] (Tetreault et al., 2013), to my knowledge the only shared task on NLI organized so far, featured 29 systems attempting to identify the native language using students essays from the TOEFL11 collection (Blanchard et al., 2013). Texts writtens by native speakers of eleven languages were included in the dataset: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish.

Among the participanting teams, Gebre et al. (2013) used TF-IDF weighting combined with three algorithms: perceptrons, SVM, and logistic regression with results reaching 81.4% accuracy using SVM. The latter two algorithms were proposed to language identification in the DSL shared task 2015 with very good results. The SVM approach ranked second among ten systems (Zampieri et al., 2015a).[24] The overall best result in the NLI shared task, 83.6% accuracy, was obtained by Jarvis et al. (2013) using an SVM classifier on a set of word and sequential POS combinations.

Among the most recent NLI approaches, Ionescu et al. (2014) investigate character-based approaches for NLI, Malmasi and Cahill (2015) examine feature diversity in NLI, Malmasi and Dras (2015c) investigate the application of NLI methods on multiple languages, and finally Ionescu and Popescu (2016) propose the use of String Kernels to NLI.

1.6.2 Lexical Variation

One of the dimensions that can be explored in the automatic identification of language varieties is lexical variation. Lexical choices may sometimes reflect personal idiosyncrasies but are usually motivated by the social function and use of language in a given community of speakers. Language varieties have unique characteristics in terms of the use of the lexicon that can make lexical variation a distinctive feature for classification algorithms to distinguish between two or more language varieties. Some studies have been published over the past few years to discover patterns of word usage based on word distribution and word frequency.

One of these studies is by Peirsman et al. (2010). This study applies distributional lexical semantics to synonymy retrieval and it was used for the identification of distinctive features (which the authors call *lectal markers*) in Dutch and Flemish. Experiments measuring lexical variation, focusing on convergences and divergences between varieties, were carried out for Brazilian and European Portuguese by Soares da Silva (2010). A study to recognize lexical and grammatical variation between standard German and the German variety spoken in the Italian region of South

[23] `https://sites.google.com/site/nlisharedtask2013/`

[24] The approach by Gebre et al. (2013) is discussed in Chapter 5.

Tyrol was carried out by Anstein (2012, 2013) using the software Vis-A-Vis.

Finally, other studies worth mentioning deal with the recognition of cognates and false friends between similar languages and varieties. Ljubešic and Fišer (2013) proposes a method to identify false friends between Slovene and Croatian, Torres and Aluísio (2011) investigate automatic methods to process cognates and false friends between Portuguese and Spanish; and Ciobanu and Dinu (2014) propose a method to identify cognates between Romanian, four other Romance languages (French, Italian, Spanish and Portuguese) and Turkish.[25]

1.7 Chapter Summary

This chapter serves as a concise yet comprehensive overview of the current state of research in language identification. I started by providing a brief historical overview describing a number of language identification approaches focusing mostly on the identification of similar languages. I looked in more detail into three language identification approaches which I consider to be very influential in different decades: Ingle's short word approach (Ingle, 1980), the Out-of-Place metric behind *TextCat* (Cavnar and Trenkle, 1994), and the information gain-based approach behind *langid.py* (Lui and Baldwin, 2012).

The review went on to describe the identification of similar languages, varieties and dialects which is a known challenge in language identification, and one which up to now has not received significant attention. Finally, I closed the chapter by discussing two related NLP tasks, namely: NLI and the automatic identification of lexical variation.

As evidenced in this chapter, and particularly in section 1.5, two of the four research questions of this thesis have not been previously answered in the literature. These are **RQ1:** Is it possible to automatically discriminate between language varieties with satisfactory performance? and **RQ2:** Can language varieties be integrated into real-world language identification systems? This thesis contributes to filling this gap.

[25] The decision to work with Turkish is, according to the authors, motivated by historical reasons and language contact, rather than by any typological connection between Romanian and Turkish.

Chapter 2

Object of Study and Data Collection

2.1 Introduction

In Chapter 1 I focused on previous work on language identification and the computational aspects of the task. The present chapter aims to clearly delimit my linguistic object of study: standard national language varieties.

This chapter introduces a number of important linguistic concepts such as pluricentric languages, a term applied to describe languages with different standard national varieties (e.g. English, French or Portuguese). The concept of pluricentricity is fundamental in this work, as I aim to explore the integration of language varieties into language identification systems. Through the use of the methods proposed in this thesis, I investigate the extent to which classification methods can be used to provide relevant quantitative and qualitative information to contrastive linguistics research, particularly applied to language varieties. Some related concepts will be discussed in this chapter as well, such as the relation between language varieties and dialects, diglossia, and the terms *Ausbausprache*, *Abstandsprache*, and *Dachsprache* coined by Kloss (1952) and presented in this thesis in Section 2.2.1.

As to the corpora collected, I opted for the use of journalistic texts which ideally reflect standard contemporary language of each language variety. The reasons behind this choice and the question of sampling are discussed in detail. As to the languages, this work focuses on the identification of national varieties of pluricentric languages from the Romance branch, namely Portuguese, Spanish, and French. Other Indo-european languages are also explored such as English.

This chapter is organized as follows:

- **Section 2.2** presents a concise introduction to pluricentric languages and national language varieties. I discuss related concepts such as dialects, diglossia, *Ausbausprache*, and *Abstandsprache* (Kloss, 1952).

- **Section 2.3** delimits my object of study, standard national language varieties, and presents the criteria I adopted for corpus compilation.

- **Section 2.4** contains the sources, facts and figures relating to the corpora compiled for this thesis.

- **Section 2.5** closes this chapter and summarizes it.

2.2 Pluricentric Languages

The term pluricentric language was first introduced by Heinz Kloss in 1952 (Kloss, 1952). Inspired by the work of Kloss, the definition of pluricentric language adopted in this thesis is the one presented by Clyne (1992):

> The term pluricentric was employed by Kloss (1978: 66-67)[26] to describe languages with several interacting centres each providing a national variety with a least some of its own (codified) norms (Clyne, 1992, p.1).

Based on this definition, a language is considered pluricentric if it possesses different national standard versions (sometimes several), both in spoken and in written forms. Commonly, these varieties are standard national versions of the given language and they are used in legal contexts, media, and also in spoken everyday communication. Examples of language varieties include American, Australian and British English; Brazilian and European Portuguese or Canadian and Hexagonal French[27].

Chambers and Trudgill (1998) in their book, *Dialectology*, are concerned mainly with 'dialects' and 'accents' rather than standard national varieties. The authors, therefore use the term 'variety' in an *ad hoc* manner as stated:

> We shall use 'variety' as a neutral term to apply to any particular kind of language which we wish, for some purpose, to consider as a single entity. The term will be used in *ad hoc* manner in order to be as specific as we wish for a particular purpose (Chambers and Trudgill, 1998, p.5).

[26] The 1978 reference is the second edition of Kloss' 1952 book *Die Entwicklung neuer germanischer Kultursprachen seit 1800.* In this dissertation I refer to the 1952 version and to his 1967 paper *'Abstand Languages' and 'Ausbau Languages'.*

[27] To avoid using French French and Spanish Spanish, I refer to the French national variety spoken in Mainland France as *Hexagonal French* and to the Spanish variety spoken in Spain as *Peninsular Spanish.* The first term originated from the French term *'français hexagonale'*, sometimes referred to as Mainland French. The latter is sometimes referred to as European or Iberian Spanish. More consensual is the use of *European Portuguese* to refer to the national variety spoken in continental Portugal.

Central to this discussion, and to the delimitation of the object of study of this thesis, is the following: given the terminology adopted by sociolinguists, I consider here varieties in terms of national varieties or standard national varieties. I therefore consider a variety to be a linguistic system officially used in all communicative contexts in a country or region with its own set of well-defined and codified norms which constitute a standard.

The definition of 'codified norms' is also important to this discussion. Leitner (1989) points out that standard varieties show that existing grammars are basically monovarietal and mononormative and they create the widespread notion of a 'core' standard language along with other systems that deviate from this norm. In his example, the national varieties of Australia, New Zealand and South Africa are determined largely by their deviations from standard British English or in others by the absence of British features. Standard British English would therefore be considered the 'core' standard language whereas all other varieties would be somehow deviations of the 'core'. This is, however, not how the current literature portrays language varieties. Instead, recent studies consider standard language varieties as unique linguistic systems that are not mere variations dependent on the 'core' language.

It is impossible to systematically define the concepts of language varieties and pluricentric languages, without mentioning the term *dialect*. Language varieties and dialects have a number of aspects in common and in some cases the boundaries between these two definitions are not well defined. The concept of a dialect is also not particularly easy to define. Kabatek and Pusch (2009) discuss the differences between dialects and languages and state that a simplistic definition for dialect is a 'diatopical variation' of a language. This definition of course allows different interpretations according to what one is interested in and should be refined for better understanding. According to Kabatek and Pusch (2009), when we look at the development of languages, it is often said that a dialect may be a 'primary form' of a language. This is true for some languages. The dialect from Castilla is, for example, the origin of modern-day Spanish. But this is not always true and the question of status and function of languages or dialects should be taken into account.

Given what has been discussed so far, it seems correct to state that the main difference between language varieties and dialects lies at their status level rather than at any intrinsic linguistic characteristic that might be studied empirically. Clyne states that:

> National varieties, those of nations or national groups, are differentiated from dialect – local and regional varieties – at the status level though not always in their linguistics indice (Clyne, 1992, p.2).

This statement does not, however, prevent scholars from empirically studying the linguistic differences between varieties and dialects. Even though the motivation for classifying or labelling a linguistic system as dialect or variety is not mainly linguistic, linguistic differences exist and they can be empirically studied at different levels (lexical, syntactic, etc.). The distinctive element in grammar and lexicon, however, may be small.

As to the social function of languages, Clyne points out that national varieties are associated with a particular nation and they are accepted by its community of speakers as a standard. At the same time, according to Clyne, these varieties are used to exclude 'non-nationals'. There are a number of cases in which national varieties are fully mutually intelligible and very similar to each other. According to Clyne these varieties often need to increase distance through corpus planning to promote the symbolic function of language as national standard. One good example is the case of Serbo-Croatian. After the end of the former Republic of Yugoslavia in 1992, each one of the new countries of the former Yugoslavia adopted its language variety (all of which are mutually intelligible among themselves) as their national language. To strengthen these differences, Serbian for example uses the Cyrillic script whereas Croatian still uses Roman characters.

Chambers and Trudgill (1998) discuss the question of mutual intelligibility. They point out that a way of looking at the relation between languages and dialects is by considering a language as 'a collection of mutually intelligible dialects'. This notion is similar to what Kloss (1967) describes as *Dachsprache* or roofing language. Chambers and Trudgill (1998), however, provide a counter example to the notion of *Dachsprache* by discussing the case of Scandinavian languages. Norwegian, Swedish and Danish are usually considered to be different languages and are to some degree mutually intelligible. They point out, however, that mutual intelligibility may not be equal in both directions stating that 'Danes understand Norwegians better than Norwegians understand Danes'. The same phenomenon can be observed in Romance languages, specifically in the case of Portuguese and Spanish, which are fully mutually intelligible in their written forms, but not in their spoken forms where allegedly Portuguese speakers understand Spanish better than the Spanish speakers understand Portuguese.

I started this chapter aiming to make a very clear distinction between varieties and dialects and stating that the experiments in this dissertation will work solely with standard national varieties. From a pure NLP perspective, however, the methods, algorithms and features presented here may be replicated for dialects as well, provided that there is enough written material for training and testing the computational models. In my thesis I focus primarily on discriminating between varieties of Romance languages, namely Portuguese, Spanish and French. I will deliberately avoid, however, discussions about the status of other languages of the Iberian peninsula such as Galician and Catalan. These are considered to be languages rather than dialects, for historical reasons, and this is how they will be considered here.

Although it is beyond the scope of this work to discuss the status of Galician within a Lusophone or Ibero-Romance context, I do acknowledge that the case of Galician is of particular interest to my work. Galicia is regarded to be the birthplace of the Portuguese language and both languages are mutually intelligible (Castro, 1991). Galician is unofficially regulated by the Royal Galician Academy, but independent organizations as the Galician Association of Language regard Galician as part of the Galician-Portuguese (*Galego-Portugues*) language. An official orthography for the Galician language is therefore not entirely consensual, with some organizations adopting a Spanish based orthography and others a Portuguese one. These orthographical differences make it very difficult to include Galician texts in the set of experiments proposed here that are based on written corpora compiled from samples of standard written language with consistent orthography.

The next section discusses two terms coined by Kloss (1967), *Ausbausprache* and *Abstandsprache*. The terms were proposed by Kloss as an attempt to define the constitution and status of different linguistic systems: e.g. languages, dialects, and varieties.

2.2.1 *Ausbausprache* and *Abstandsprache*

Given what has been discussed so far, it would be perfectly possible to call, for example, Brazilian Portuguese a language in its own right and name it Brazilian, or Quebec French, naming it Quebecian or Canadian. If there was enough political motivation to do so, as there was in the aforementioned case of Serbo-Croatian, these varieties could be adopted as official languages in their respective countries. The most common way of doing this is by deliberate language planning carried out by official organizations. Even though this is possible, would it represent any change in these languages' linguistic properties?

Kloss (1967) introduces the terms *Ausbausprache* and *Abstandsprache* with the main aim of describing and delimiting this kind of situation. To understand what *Ausbau* language means we should first differentiate it from the notion of *Abstand* language but also from what Stewart (1968) referred to as polycentric standard language. For this important distinction, I will use here the four diagrams presented in Kloss (1967), in which circles correspond to written standard and squares correspond to spoken language.

In Figure 2.1, we see what may be called 'the normal situation' with Kloss' own example of Breton.
The aforementioned situation is considered by Kloss as the standard situation, which he defines as follows.

> A standard based on some of the spoken speech forms and neither sub-

> divided in two major variants nor exposed to the competition of another standard based on other Breton dialects. (Kloss, 1967, p.31)

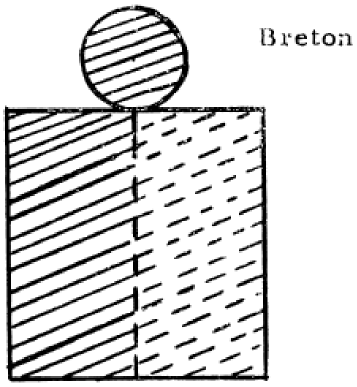

Figure 2.1: The Normal Situation (Kloss, 1967)

The second scenario is presented in Figure 2.2 and refers to the polycentric standard language (or pluricentric standard language). The second case is the most representative and will be explored in this work. According to Kloss, in Figure 2.2 we have 'two variants of the same standard, based on the same dialect or a near-identical dialect'. He exemplifies this with the case of Serbo-Croatian.[28]

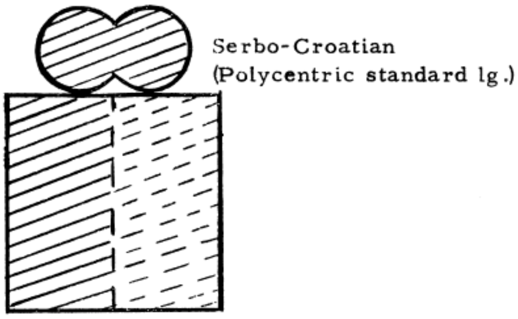

Figure 2.2: Polycentric (pluricentric) Standard Language (Kloss, 1967)

In the case of pluricentric languages, the existence of the two varieties does not prevent us from considering them as a single language. Kloss gives the example of Moldavian and Romanian which are varieties of the same standard language rather than two separate languages. The case of Moldavian and Romanian is, however, not the standard case of polycentric or pluricentric languages which often occur when these languages are found in two or more geographically separated countries,

[28] This corresponds to the situation of Serbo-Croatian before 1992. After 1992, Serbian and Croatian may be considered *Ausbau* languages.

usually a consequence of colonization as in the case of British and American English; Brazilian and European Portuguese; and Hexagonal and Quebec French.

The third case represented by Figure 2.3, is the case of *Ausbau* languages. The concept of *Ausbau* language is according to Kloss (1967) primarily a sociological one. Figure 2.3 portrays the case of Czech and Slovak, two *Ausbau* languages spoken in the former Czechoslovakia and currently spoken in the Czech Republic and in Slovakia or the Slovak Republic.

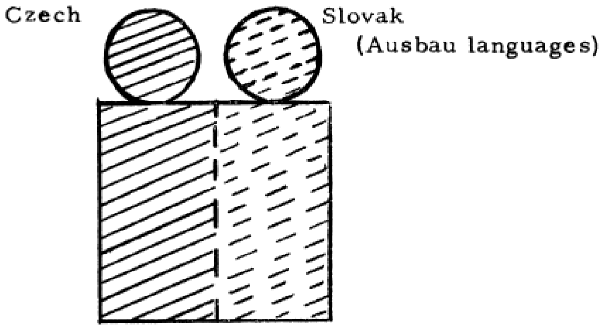

Figure 2.3: Two Ausbau Languages (Kloss, 1967)

Examples of *Ausbausprachen* codified as separate languages include Hindi and Urdu, Indonesian and Malay, Czech and Slovak and the current situation of Croatian and Serbian. This third situation represents the hypothetical scenario which I discussed at the beginning of this section. The possibility of the creation of a 'Brazilian' or a 'Quebecian' language would be considered in this framework as the creation of an *Ausbau* language or language by development.

The fourth and last case presented here is represented in Figure 2.4. It is the case of what Kloss defines as *Abstandsprachen* or language by distance. The diagram represents the case of German and Dutch.

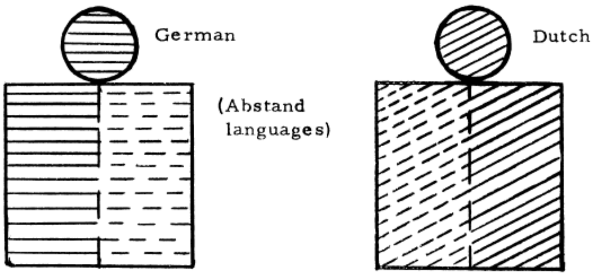

Figure 2.4: Two *Abstandsprachen* (Kloss, 1967)

The example of German and Dutch is to some degree comparable to Portuguese and Spanish, two languages which are typologically related and therefore present a degree of mutual intelligibility (particularly in their written forms) but are no longer considered to be varieties of each other.

> The term *Abstandsprache* is paraphrased best as 'language by distance', the reference being of course not geographical but to intrinsic distance (Kloss, 1967, p.29).

Unlike most other concepts presented in this section, *Abstand* language is a predominantly linguistic concept. Kloss acknowledged, however, that the criteria applied by linguists to measure the distance between languages was beyond the scope of his work, and he assumed that linguists would apply reliable and uniform criteria for this task. The article discussing all concepts presented so far was published in 1967, prior to the widespread use of corpora and corpus-based and corpus-driven methods which allow researchers to measure linguistic distance in a much more reliable manner.

As to the relation between *Abstand* and *Ausbau* languages, Kloss points out that many languages can be classified as both.

> Many of the leading tongues of the world, among them English, French and German, are both *Abstand* and *Ausbau* languages, i.e., they are called languages both because of having been made over and because of their intrinsic distance from all other languages (Kloss, 1967, p.30).

At this point it is important to mention another term coined by Kloss (1967), the last in this section, which is the term *Dachsprache*. A *Dachsprache* can be translated as a 'roofing language', which is a language that serves as standard language for different dialects, as in a dialect continuum. Cases of languages under a common *Dachsprache* include, for example, modern standard Arabic, which comprises the speakers of many different Arabic dialects or varieties. Under a *Dachsprache* it is often the case that the dialects are so different from each other that they are not mutually intelligible. This is very common in dialect continua that spread throughout a large geographical area. In this case, the dialects spoken at the extremities of the continuum are often no longer mutually intelligible.

The question of diglossia is also taken into account to define my object of study. In this thesis I only include texts from language varieties which are not in a diglossic situation[29]. The main reason is that languages and varieties in diaglossic situations have a different status if compared to those which are not. As an example, French spoken in the Maghreb region coexists with dialectal and standard Arabic, whereas

[29] See Section 2.3 for the criteria used when choosing the language varieties and compiling the corpora.

Brazilian Portuguese is the only language widely spoken in Brazil and it is spoken and written in all communicative situtations[30].

Ferguson (1959) defines diglossia as follows.

> Diglossia is a relatively stable language situation in which, in addition to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically more complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversations. (Ferguson, 1959, p.336)

Examples of diglossia include the situation of Arabic and Swiss German. In both of these situations, there is a high variety and a low variety. In Switzerland, a standard German variety is the language used in schools, newspapers and public administration and different Swiss German dialects are the languages spoken in informal contexts. The linguistic situation in Switzerland is particularly special with standard German, French, Italian and Romansh coexisting as standard languages.

In the case of Arabic, a triglossic situation can be observed. Classical Arabic is the language of religion, modern standard Arabic is regarded as the variety with higher prestige used in official contexts, schools and academia and regional Arabic or dialect Arabic is the daily spoken communication language.

2.3 Delimiting the Object of Study

The present chapter serves to define the object of study in this dissertation. I take the sociolinguistic definition of standard national language varieties as a starting point. Based on this definition, I define my object of study as standard language varieties in contrast to dialects, sociolects and other kinds of language variety defined in the literature.

The reasons behind this choice can be summarized into two main aspects:

1. The classification or identification of language varieties, dialects, and closely related languages in real-world NLP applications has been mostly neglected. Apart from some recent studies, this question has not been substantially explored as evidenced in the previous chapter.

[30] There are a number of indigenous languages in Brazil, particularly in the northern region. They have, however, very small community of speakers which are not comparable to the status of dialectal Arabic, Swiss German or other languages in diglossic situation.

2. The use of corpus-driven methods such as the one presented here is of interest to contrastive linguistics. As previously mentioned, this study does not leave the linguistic aspect of statistical language modelling aside. I intend to explore the extent to which these experiments provide insights into the differences and convergences of the language varieties studied here.

Given what has been discussed so far, the decision of working with standard language varieties is, on its own, not enough. There are other aspects that should be taken into consideration before moving any further. The object of study of this dissertation needs to be refined to be precise enough to allow generalizations.

The findings of the experiments presented here should, to some extent, be representative of the languages and language varieties studied, which means that these results should ideally not be biased by sampling. It is, however, understood that working with perfect samples is an impossible task and that corpus compilation aims to minimize these effects, while always taking into account that a certain degree of bias or variation will be present in the samples.

Before looking more closely at the question of sampling, there are a few remarks on the language varieties that will be studied here that are worth mentioning. All language varieties studied in this thesis are considered to be the following:

- Standard national language varieties.
- Languages which are official in their respective countries.
- Represent synchronic and contemporary language.
- Languages which are in a non-diglossic situation.

Provided that I have clearly defined the object of study, the question of sampling still remains. Which textual material should be compiled to serve as a sample of the language that will be studied? How big should this corpus be? And where should this material come from?

One limitation at this point is that, apart from a few exceptions (e.g. French), languages other than English do not possess large size balanced corpora. And even if they did, it is still not consensual whether or not the criteria adopted by, for example, the British National Corpus (BNC) or by the International Corpus of English (ICE) are optimal. These corpora are compiled with balanced text types and genres to serve as reference corpora for a given language and the criteria behind the compilation of these corpora is scientifically well supported. Nevertheless, they are not free of criticism. The BNC, for example, comprises 90% written material and 10% spoken material, mostly because of the practical difficulty and costs of collecting and transcribing spoken data rather than due to any scientific or linguistic motivation.

2.3.1 Collecting Corpora

The definition of the samples to be studied is for the reasons discussed so far, not trivial. Some decisions should be taken and consistently observed in all experiments to avoid thematic or any other kind of bias. According to these decisions the samples used here should be:

- Written samples,
- Extracted from newspapers,
- Balanced in terms of topics and text types,
- Represented by a consistent orthography.

Written samples extracted from newspapers are used for practical reasons, but also to minimize regional variation. It is understood that the journalistic genre is a standardized version of language. This standardization enables the study to compare standard versions of varieties rather than sub-standard regional versions.

The balance between text types and topics is obtained through random sampling. The newspapers used contain different sections (e.g. economy, sports, science, politics) and random sampling ensures that all text types will be represented. The question of consistent orthography does not influence most varieties such as Brazilian Portuguese or Argentinian Spanish. These varieties have official organizations that regulate orthography. This is not the case, for example, with Galician in which for political reasons, multiple orthographies coexist.

2.4 Corpora

Nineteen corpora were collected to perform the experiments described here. All samples consist of contemporary journalistic texts compiled from 2002 onwards. Nine of these corpora originated from the SETimes Corpus (Tyers and Alperen, 2010). The sample of the SETimes corpus I used was based on the one made available by Nikola Ljubešić, who carried out post-processing of the data to eliminate meta-information and to correct inconsistencies.[31] SE stands for *south-east European*, and it contains texts from the SETimes news portal which was a website that ceased to exist in March 2015.

The other ten corpora were retrieved from local newspapers published in the different language varieties. All of them were published in the year of 2008, except for the Brazilian Portuguese corpus which was published in 2004 and made available

[31] `http://www.nljubesic.net/resources/corpora/setimes/`

by *Folha de São Paulo*[32]. Using Python and Perl scripts, I carried out the extraction, compilation, cleaning and indexing of all articles prior to the experiments[33].

An overview of all corpora can be seen in Table 2.1. I used the ISO 3166 alpha-3 country code which is used throughout this dissertation except were indicated.[34] Language varieties are displayed in bold.

Language	ISO Code	Source	Year
Albanian	ALB	SETimes Corpus	2002 - 2010
American English	**USA**	**New York Times and Washington Post**	**2008**
Argentinian Spanish	**ARG**	**La Nacion**	**2008**
Bosnian	BIH	SETimes Corpus	2002 - 2010
Brazilian Portuguese	**BRA**	**Folha de São Paulo**	**2004**
British English	**GBR**	**The Guardian and The Independent**	**2008**
Bulgarian	BGR	SETimes Corpus	2002 - 2010
Croatian	HRV	SETimes Corpus	2002 - 2010
Hexagonal French	**FRA**	**Le Monde**	**2008**
Greek	GRC	SETimes Corpus	2002 - 2010
Macedonian	MKD	SETimes Corpus	2002 - 2010
Mexican Spanish	**MEX**	**El Universal**	**2008**
Peruvian Spanish	**PER**	**El Comércio**	**2008**
European Portuguese	**PRT**	**Diario de Noticias**	**2008**
Quebecian France	**CAN**	**Le Devoir**	**2008**
Romanian	ROU	SETimes Corpus	2002 - 2010
Serbian	SRB	SETimes Corpus	2002 - 2010
Peninsular Spanish	**ESP**	**El Mundo and El Pais**	2008
Turkish	TUR	SETimes Corpus	2002 - 2010

Table 2.1: Corpora: Languages, Sources and Year of Publication

Quantitative detail about the corpora (e.g. number of documents, types, tokens, and average token per document) is presented in Table 2.2. I compiled and processed a total of 301,241 documents written in 19 languages and language varieties resulting in slightly over 17 million tokens.

As can be seen in Table 2.2 the number of texts varies across languages. Texts are also of a different length depending on the data source. It is well known, however, that both the amount of training material and the length of documents play a crucial role in language identification and text classification tasks in general. For

[32] It should be noted that the best possible scenario in this task is to work with texts that were published in the same year or even in the same month, if possible. This is to diminish the impact of time-sensitive information in classification. Even so, a feature analysis of the results presented in Zampieri and Gebre (2012), showed that the variation between the Brazilian and European corpora were mostly diatopic and not diachronic.

[33] This step was carried out with the help from Sascha Diwersy who provided part of the material.

[34] In some experiments I grouped two varieties into the same class, to represent the language. I applied an *ad-hoc* code instead of country codes.

this reason, instances from the aforementioned corpora were subsequently sampled to perform each experiment presented in Chapter 4. Sampling was necessary in order to: 1) account for a balanced representation between language varieties (same or similar amount of texts for each class); and 2) make sure that the average length of texts was similar.

Even so, it is worth noting that the texts available in the corpora compiled from language varieties are substantially longer than those from non-pluricentric languages. This was done intentionally since the focus of this thesis is language varieties and therefore a larger sample for pluricentric languages is required. The corpora compiled from non-pluricentric languages were included to emulate a real-world language identification scenario (see Section 4.5.2).

Language	Documents	Tokens	Types	Avg. Tokens
ALB	6,776	150,613	17,656	22.22
USA	**2,034**	**1,073,589**	**41,095**	**527.82**
ARG	**1,055**	**861,439**	**29,023**	**816.52**
BIH	4,308	164,239	29,596	38.12
BRA	**8,441**	**2,641,961**	**74,595**	**312.99**
GBR	**3,032**	**1,419,803**	**46,158**	**468.27**
BGR	63,221	1,285,084	59,261	20.32
HRV	31,619	666,570	58,432	21.08
FRA	**1,518**	**690,158**	**44,260**	**454.64**
GRC	65,967	1,486,202	62,852	22.52
MKD	51,917	1,105,454	52,011	21.29
MEX	**1,593**	**660,406**	**35,339**	**414.56**
PER	**1,632**	**694,339**	**42,358**	**425.45**
PRT	**3,042**	**1,403,851**	**52,979**	**461.48**
CAN	**1,525**	**806,778**	**42,942**	**529.034**
ROU	7,365	164,905	18,181	22.39
SRB	36,108	805,808	63,443	22.31
ESP	**2,589**	**841,151**	**36,913**	**324.89**
TUR	7,769	140,921	24,206	18.13
Total	**301,241**	**17,063,271**	**-**	**-**

Table 2.2: Corpora: Facts and Figures - Number of Documents, Tokens, Types, and Average Number of Tokens per Document

Parallel to the corpora used in the dissertation, I also compiled the DSL corpus collection (DSLCC) in collaboration with the co-organizers of the DSL shared task. For the DSL shared task, we had to collect samples that could be used by all participants and later be redistributed. This was unfortunately not the case for some of the corpora I compiled for this thesis (e.g. Brazilian Portuguese), which, due to copyright reasons, cannot be redistributed.

2.5 Chapter Summary

In this chapter I defined my object of study, and the methods for compiling the corpora which will be used in the experiments presented in the next chapters. A question that the reader might pose before reading this chapter is: why experiment with Brazilian and European Portuguese and not, say, Spanish and Portuguese or Serbian and Croatian? This chapter was written to answer this question. A concise literature review on language varieties and pluricentric languages based on the work of Heinz Kloss was necessary to delimit my object of study.

Based on the literature discussed in this chapter, I argue that the distinction between dialects, language varieties, and languages in their own right is to a large extent a political one. Even so, it is possible to differentiate between dialects, language varieties, and similar languages by looking at some aspects such as contexts of language use (e.g. colloquial, official, media), *diglossia*, the presence of a widespread standard orthography, etc. I take these aspects into account to delimit my object of study and to compile the samples used in this thesis.

Given the motivation of my thesis, I contend that journalistic texts are the most appropriate text types for the experiments I will present in the next chapters.

Chapter 3

Evaluation and Computational Techniques

3.1 Introduction

This chapter presents the computational techniques behind the automatic language identification systems used in this thesis. The chapter begins by explaining the methods and the metrics I used to evaluate classification performance. For the evaluation I use standard NLP metrics such as precision, recall, f-measure and accuracy.

Next I present a concise overview of n-gram language models and bag-of-words (BoW) which were used as features in the experiments presented in this thesis. The chapter presents likelihood estimation (LE), a simple, fast, and effective probabilistic classifier that combines Laplace smoothing and a Bayesian classifier proposed for this task in Zampieri and Gebre (2012). I also discuss other machine learning algorithms used in this thesis such as Support Vector Machines (SVM), Naive Bayes, and Decision Trees.

To justify the use of the likelihood estimation method and the experimental setting I propose, I close this chapter by presenting two preliminary experiments. The first compares the results obtained using LE to those obtained by Ljubešić et al. (2007); Tiedemann and Ljubešić (2012) for South Slavic Languages. The second validates the data collected. I compare the performance of LE discriminating between texts from two newspapers from the same country to the performance obtained discriminating between texts published in two different countries. The assumption is that diatopic variation is stronger than the possible stylistic variation between two newspapers published in the same country.

Chapter 3 is organized as follows:

- **Section 3.2** presents the evaluation metrics used in this thesis.

- **Section 3.3** provides a concise overview of n-gram language models and smoothing techniques based on Manning and Schütze (1999) and Hammond (2007).

- **Section 3.4** explains bag-of-words (BoW) models and how they can be useful for identifying language varieties.
- **Section 3.5** describes the four classification algorithms that are used in this dissertation: Decision Trees, Likelihood Estimation, Naive Bayes, and Support Vector Machines.
- **Section 3.6** presents the results of two preliminary experiments that pave the way for the experiments that will be presented in Chapters 4, 5 and 6.
- **Section 3.7** summarizes this chapter.

3.2 Evaluation Metrics

To evaluate the extent to which the methods used in this dissertation are adequate for language identification, I use standard metrics used in NLP and text classification to report results in terms of precision, recall, f-measure and accuracy. Precision, recall and f-measure are used to evaluate the performance in multi-class classification experiments whereas accuracy is used for binary classification.[35]

The metrics used are presented next and they are based on the possible outcomes of a confusion matrix.

Predicted Class	**Actual Class**	
	Positive	Negative
Positive Prediction	True Positive	False Positive
Negative Prediction	False Negative	True Negative

Table 3.1: Example of Confusion Matrix

The confusion matrix contains four possible outcomes: tp, tn, fp and fn or *true positives*, *true negatives*, *false positives*, and *false negatives.* The results are obtained per class. To evaluate the performance of the classifier across all classes it is necessary to calculate the average (mean) performance of all classes. This allows us to evaluate how well the classifier is performing when identifying each individual class as well as when distinguishing all classes.

The first evaluation metrics are precision and recall. Precision refers to the percentage of positive predictions that are correct, whereas recall calculates the

[35] Reporting accuracy for binary classification is a common-practice in text classification settings involving balanced classes (same amount of data from both classes).

percentage of instances correctly identified divided by those who should have been identified. Prevision and recall are calculated as follows:

$$Precision = \frac{tp}{tp + fp} \tag{3.1}$$

$$Recall = \frac{tp}{tp + fn} \tag{3.2}$$

After the calculation of precision and recall, it is common practice to calculate a score that takes both metrics into account. This is usually done either by calculating f-measure or accuracy which provide a unified metric of success for each class. F-measure, for example, takes precision and recall into account and it can be customized to emphasize one or the other (Van Rijsbergen, 1979).

$$F = \frac{(\beta + 1) \times P \times R}{(\beta \times P) + R} \tag{3.3}$$

The importance of precision and recall can be tuned by the variable β: if β is set to 1 then recall and precision are counted equally. When β is set to 2, recall is twice as important as precision whereas when β equals 0.5, precision is twice as important as recall. Throughout this dissertation, precision and recall have equal weights and therefore f-measure is calculated as follows:

$$F = \frac{2PR}{P + R} \tag{3.4}$$

The fourth metric is accuracy which takes into account the number of instances correctly classified ($tp + tn$) divided by all instances classified.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{3.5}$$

3.3 N-gram Language Models

The vast majority of state-of-the-art language identification methods rely on (character) n-gram language models. The application of n-gram language models to language identification can be traced back to the work of Dunning (1994). N-gram language models are simple statistical language models calculated based on the co-occurrence of words or characters across text samples (Shannon, 1951). These models estimate the probability of different words occurring alone or in sequence in a text or a corpus. The same is applied to character sequences arranged in the form of n-grams.[36]

Chapter 6 from Manning and Schütze (1999) and Chapter 9 from Hammond (2007) provide a detailed explanation of n-gram language models and smoothing techniques. I used both chapters as a basis for the following sections together with other useful sources such as the PhD thesis by Dehdari (2014), some sections of the book on Statistical Machine Translation by Koehn (2009), and the book by Kelleher et al. (2015) on Fundamentals of Machine Learning. Examples of unigram and higher-order n-gram calculations were adapted from Hammond (2007).

Unigrams treat units in isolation. They are therefore the simplest kind of language models. Each word, token or character is assigned a probability and, just like the bag-of-words model, unigrams assume independence between them.[37]

To understand the calculation of unigrams I present the following three sentences as an example corpus:

(1) a. John and Mary met at the mall.
b. John goes to the mall once a week.
c. Mary never goes there.

These are three well-formed sentences containing a total of nineteen words (seven words in sentence *a*, eight words in sentence *b*, and four words in sentence *c*).[38] The word *week* occurs only once in the example corpus. We can therefore estimate that *week* has a probability of $\frac{1}{19}$ or 0.052 of occurring. The words *John*, *goes*, and *mall* occur twice in the example corpus and each of them has a probability of $\frac{2}{19}$ or 0.105 of occurring.

This information is then used to calculate the probability of words in a sequence. The probability of *John goes to* is obtained by multiplying the individual probabilities of each word: $0.105 \times 0.105 \times 0.052 = 0.000573$. This calculation is very simple and does not take word order into account. *John to goes* or *to goes John* would also receive the exact same score, although the latter two are not well-formed combinations and probably would never (or very rarely) appear in an English corpus.[39]

[36] When processing texts, not only words and characters can be represented as n-grams but also morphosyntactic patterns using for example part-of-speech tags. I discuss this point in Chapter 6.

[37] The fundamental difference between word unigram models and BoW approaches presented in this thesis is smoothing. The calculation of probabilities of unigrams include a simple Laplace smoothing whereas the BoW approach does not.

[38] For the sake of simplicity, in this example I do not take punctuation into account. Most tokenization methods, however, do consider punctuation as a token.

[39] The fact that a particular token or combination does not appear in a corpus does not mean it will not appear in another sample. I discuss this issue later in this chapter when I talk about smoothing techniques.

An example of the use of unigrams in language identification is the work of Souter et al. (1994). In this study, texts were analysed word by word with a likelihood calculated for each. After the text is analysed, the program returns the most likely language according to the word unigram probabilities.

There are, however, more sophisticated ways of calculating language models which take context into account. Natural language has a number of word order restrictions and for this reason it is naive to assume independence between the words of a sentence.[40] In English, personal pronouns are very often followed by verbs whereas adjectives are often placed before nouns. Returning to the previous example, *John goes to* is an acceptable combination in English, whereas *John to goes* is not. The same is true for character combinations, for example *th* are *ng* are very frequent character combination in English, whereas *ht* and *gn* are not.

Language models can therefore take ordering restrictions that are intrinsic to any natural language into account for more accurate probability estimation. One of the ways of doing so is by using bigrams and higher-order n-grams (trigrams, 4-grams, 5-grams etc.).

Bigrams are known as first order language models. They aim to capture some of the ordering restrictions that occur in natural language by considering the probability of a word occurring as a function of its immediate context, as shown below:

$$P(w_1 w_2) = P(w_1)P(w_2|w_1) \tag{3.6}$$

The calculation of bigrams allows us to understand the logic behind higher-order language models as well. Higher-order models take larger context into account and often deliver good results on large datasets. Before discussing higher-order n-gram models I will give an example of character n-grams. This example is adapted from the influential language identification paper by Cavnar and Trenkle (1994).[41] It takes the word *walk* and models it as bigrams, trigrams and 4-grams. The symbol * is used to represent blanks.

- bigrams: *w, wa, al, lk, k*
- trigrams: *wa, wal, alk, lk*, k**
- 4-grams: *wal, walk, alk*, lk**, k***

The example shows that the higher the order of the n-gram model, the more complete lexical units the model will consider. The word *walk* is considered as a whole in a

[40] In some applications, the simplicity of unigrams works well due to data sparsity observed in higher-order n-grams.

[41] The same example appears in Cavnar and Trenkle (1994) with the word *text*.

4-gram model. The 4-gram language model in the example would therefore not only be considering character sequences, but also a significant number of complete words.

Most language identification methods perform best by using character trigrams. Nevertheless bigrams and 4-grams have also been substantially explored in the literature.[42] String probabilities based on character and word trigrams can be calculated using the following formula:

$$P(w_i w_2 ... w_n) = P(w_1) \times P(w_2|w_1) \times \prod P(w_n|w_{n-2}w_{n-1}) \tag{3.7}$$

Given this calculation, it is possible to estimate n-gram probabilities for higher-order n-grams using similar methods as those used for unigrams. The simplest way to estimate this probability is by using Maximum Likelihood Estimation (MLE). In MLE the occurrence of a given n-gram is considered to be a random variable in which each n-gram is independent of the next (binomial distribution). This is a practical yet untrue assumption. For the aforementioned reasons, languages possess a number of ordering restrictions that influence the distribution of lexical items, characters, etc.

The probability of unseen events in MLE will be explained later in this chapter, when presenting smoothing techniques. Manning and Schütze (1999) state that: 'MLE does not waste any probability mass on events that are not in the training corpus, but rather it makes the probability of observed events as high as it can subject to the normal stochastic constraints'. MLE of n-gram probabilities are calculated as follows.

$$P_{MLE}(w_1 ... w_n) = \frac{C(w_1 ... w_n)}{N} \tag{3.8}$$

Where C is the frequency of $w_1 ... w_n$ in the training data and N is the total number of training instances. The formula presented above applied to w_n considers its prefix $w_1 ... w_{n-1}$ as follows.

$$P_{MLE}(w_n|w_1 ... w_{n-1}) = \frac{C(w_1 ... w_n)}{C(w_1 ... w_{n-1})} \tag{3.9}$$

The main criticism about the use of MLE is the problem of sparseness of data, even if a large corpus is used.

> While a few words are common, the vast majority of words are very uncommon and longer n-grams involving them are thus much rarer again.

[42] In this thesis I show that for language varieties, higher-order n-gram models (e.g. 5-grams) also perform well.

> The MLE assigns a zero probability to unseen events, and since the probability of a long string is generally computed by multiplying the probabilities of subparts, these zeroes will propagate and give us bad (zero probability) estimates for the probability of sentences when we just happened not to see certain n-grams in the training text (Manning and Schütze, 1999, p.198).

To cope with this problem, in practice, it is usual to not calculate n-grams for all words in the training corpus but only those which are most common. It is possible to set a threshold and all words with a frequency below this threshold are not included in the calculation. These words are therefore considered to be 'out of vocabulary' items. As words in a corpus follow a Zipfian distribution, the technique reduces the parameter space significantly. This is often done, for example, for the so-called *hapax legomena*, words that appear only once in the corpus.

3.3.1 Smoothing Techniques

Starting with the n-gram calculations discussed so far, what would happen if a word exists in the language but does not appear in a given corpus? Should it be part of the language model? No matter how big a corpus is, some words might be very rare and simply not occur. However, this does not mean they will not occur in other samples. Therefore, attributing a zero probability to them would spoil the calculation. The same is true for character sequences, due to the simple fact that in any given language some character sequences are more frequent than others.[43]

Manning and Schütze (1999) state 'regardless of how the probability is computed, there is still the need to assign a non-zero probability estimate to words or n-grams that are not present in our training corpus'. This technique is called smoothing and there are a number of smoothing techniques that are used in natural language processing and in language identification. A very simple one used in Dunning (1994) and Zampieri and Gebre (2012) is Laplace smoothing (Kotz et al., 2001), also referred to as 'add one smoothing'.

In the context of language modelling, Laplace smoothing is calculated as follows:

$$P_{lap}(w_1...w_n) = \frac{C(w_1...w_n) + 1}{N + B} \tag{3.10}$$

The formula is similar to the aforementioned MLE modified by adding 1 to the numerator (to assign a non-zero probability), and B representing the number of total possible unique n-grams in the denominator.

[43] For the Indo-European languages I work with in this thesis, the size of any vocabulary is, of course, much larger than the size of their alphabets.

The biggest criticism regarding the simplicity of the Laplace smoothing is that it leads to overestimation of the probabilities of unseen n-grams. One common alternative to Laplace is Lidstone's law of succession, where a positive value λ is added (Manning and Schütze, 1999) to both numerator and denominator of the equation.

$$P_{lid}(w_1...w_n) = \frac{C(w_1...w_n) + \lambda}{N + B\lambda} \tag{3.11}$$

There are other smoothing techniques worth mentioning that are used not only for language identification but also in other NLP tasks. Chapter 4 of Jurafsky and Martin (2009) provides a very intuitive introduction to several smoothing techniques. One example of such a technique is Good-Turing discounting, proposed by Good (1953) (Good credits Alan Turing for the original idea, hence Good-Turing). The basic intuition of this method is to use the count of items that appear once in a dataset to estimate the count of unseen items. Other smoothing techniques include back-off models, such as the one proposed by Katz (1987), widely used in speech processing, and absolute discounting (Ney et al., 1994). Kneser-Ney discounting (Kneser and Ney, 1993) improves absolute discounting by using a more sophisticated way of handling back-off distribution.

As to the relevance of smoothing to language identification, Giwa and Davel (2013) test different smoothing methods and show that smoothing substantially improves accuracy of language identification algorithms compared to a simple Naive Bayes baseline model.

The calculation of language models presented in this section are an important part of state-of-the-art n-gram-based language identification methods. These language models serve as the primary source of information to calculate the probability of a document belonging to a given class (language).

3.4 Bag-of-words

Bag-of-words (BoW) is a simple way of representing data that unlike higher order language models, assumes independence between words. The origin of the term bag-of-words in a linguistic context is attributed to the American linguistic Zellig Harris in his seminal paper 'Distributional Structure' (Harris, 1954).

BoW has been widely used in information retrieval (IR) and in several NLP tasks such as word sense disambiguation (WSD) as well as in a number of text categorization problems. In a bag-of-words model, texts (in the example of automatic classification, instances to be classified) are represented by a word vector with n number of entries (words). These n entries correspond to all words found in the corpus and catalogued in a dictionary. All entries n receive a number y in a vector

v depending on the presence or absence of n in the instance.

To clarify this idea, take the following three sentences as an example of a 'corpus':

(2) John likes football.

(3) Mary doesn't like football.

(4) John loves football.

In order to represent these three sentences (instances) in a bag-of-words model, the BoW algorithm has to convert each of them to a vector v. These vectors are binary and refer to the presence or absence of a given word in each sentence. For this calculation, the first thing that a BoW approach needs to do is to create a dictionary[44] containing all words as follows:

Position	**Word**	**Count**
1	football	3
2	John	2
3	likes	1
4	Mary	1
5	doesn't	1
6	like	1
7	loves	1

Table 3.2: Example of BoW Dictionary

Based on the dictionary above, the BoW model creates a vector v of length n for each instance, based on the presence or absence of the word at the position p. The length of the vector is equal to the number of entries in the dictionary, in this example, seven.

1. (1, 1, 1, 0, 0, 0, 0)
2. (1, 0, 0, 1, 1, 1, 0)
3. (1, 1, 0, 0, 0, 0, 1)

Sentences are represented as a collection of digits separated by commas and this representation is then used as input in NLP applications (in the case of language identification they serve as input for classification algorithms).[45] Techniques such as the popular TF-IDF weighting, term frequency - inverse document frequency, can be

[44] The dictionary presents types and not lemmas. Hence the presence of *like* and *likes*. Creating lemmatized BoW is also possible provided that a lemmatization tool is available.

[45] Note that in this example none of the words appear more than once in each instance, hence the uses of only ones and zeros.

applied to these text representations to capture words that that are more relevant in a particular document in comparison to the rest of the text collection.

As to the use of BoW to discriminate between language varieties, apart from the study published by Huang and Lee (2008), very little has been published to date. This is mainly because language identification methods developed to distinguish similar languages and language varieties use the same methods applied to general-purpose language identification and therefore rely on n-gram language models.

3.5 Classification Methods

In this section I present the algorithms used in this thesis. I used implementations of popular machine learning classifiers, namely: Naive Bayes, Support Vector Machines, and Decision Trees along with a variation of a Bayesian probabilistic classifier named likelihood estimation method proposed for this task in Zampieri and Gebre (2012).[46]

The likelihood estimation method is a simple discriminative n-gram-based method similar to a Naive Bayes classifier. The code is implemented in Python and uses functions available at the Natural Language Toolkit (NLTK) (Bird et al., 2009) such as *nltk.FreqDist* and *nltk.ngrams* to calculate language models. For smoothing, the classifier uses a simple Laplace probability distribution, explained in the previous section. As a reminder, the formula is repeated next.

$$P_{lap}(w_1...w_n) = \frac{C(w_1...w_n) + 1}{N + B} \tag{3.12}$$

With this information the algorithm calculates language models or profiles for each of the languages present in the training set. Given an input text, the algorithm uses this information to estimate the probability of a given document belonging to one class (language or language variety) or the other. For probability estimation, the algorithm uses a log-likelihood function (Dunning, 1993) shown below:

$$P(L|text) = \arg\max_{L} \sum_{i=1}^{N} \log P(n_i|L) + \log P(L) \tag{3.13}$$

N is the number of n-grams in the test text, and L stands for the language models. When provided with a text in the test set, the model calculates the probability for each of the language models and the language model with the highest probability determines the identified (or assigned) language of that given text.

[46] An implementation with language models is described in Zampieri and Gebre (2014).

This model is generally very similar to a Naive Bayes classifier and its basic difference lies in n-gram probability calculation. While a standard Naive Bayes algorithm uses conditional probabilities e.g. $P(w_3|w_1, w_2)$, likelihood estimation uses the product of individual unigram probabilities e.g. $P(w_1, w_2, w_3)$ to calculate bigrams and higher-order n-grams. This requires less calculations than a standard Naive Bayes implementation making likelihood estimation slightly faster than the most common Naive Bayes implementations particularly when using higher-order n-grams. The decision to not use conditional probabilities was taken to: 1) make the algorithm faster when handling large datasets; 2) improve generalization; and 3) to test whether a simpler way of modelling language would obtain satisfactory results. Throughout this thesis I show that the performance of likelihood estimation is comparable to other state-of-the-art approaches, and that the method is in most cases faster.

Along with the likelihood estimation method, I also experiment with three other popular machine learning classifiers for this task: Multinomial Naive Bayes (MNB) (Frank and Bouckaert, 2006), the Support Vector Machine (SVM) (Joachims, 1998; Cristianini and Shawe-Taylor, 2000; Joachims, 2006) adaptation called Segment Minimal Optimization (SMO) (Platt, 1998) and the J48 algorithm, which is an adaptation of the C4.5 algorithm (Quinlan, 1993). These classifiers are widely used in text classification and deliver very good results. Implementations of the three classifiers are available in machine learning packages such as the WEKA Machine Learning Workbench (Witten and Frank, 2005) and LIBLINEAR (Fan et al., 2008).

The Multinomial Naive Bayes (MNB) classifier, as the name suggests, is based on Bayes' theorem. Broadly speaking, Naive Bayes classifiers work under the assumption that the presence or absence of a particular feature of a class is not related to the presence or absence of any other feature. The independence assumption makes (Multinomial) Naive Bayes classifiers particularly useful for supervised learning, and it makes them very fast both at the training and prediction stages when compared to other learning algorithms. Bayes' theorem is represented by the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.14}$$

Where $P(A|B)$ is a conditional probability of A given B. As discussed by Kibriya et al. (2004), Bayes' theorem applied to text classification computes class probabilities for a given document and a set of classes represented by C. It assigns a text document t_i to the class with the highest probability $P(c|t_i)$ given by the equation below for $c \in C$. Adapting the formula to texts and classes we have:

$$P(c|t_i) = \frac{P(t_i|c)P(c)}{P(t_i)} \tag{3.15}$$

Frank and Bouckaert (2006) discusses the use of Naive Bayes for text classification in situations in which classes are unbalanced (e.g. scenarios in which there are more documents from one class than the other(s)). The authors point out that the MNB classifier considers each document as a collection of words, and that the order of the words is not particularly relevant. As presented by Frank and Bouckaert (2006) the probability of a class c is computed using the following:

$$P(c|d) = \frac{P(c)\Pi_{w \in d} P(w|c)^{n_{wd}}}{P(d)} \tag{3.16}$$

Where n_{wd} is the number of times a word w occurs in a document d and $P(w|c)$ is the probability of word w given a class c.

Kibriya et al. (2004) discuss the use of MNB and the transformation steps that led to 'transformed weight-normalized complement Naive Bayes' (TWCNB) for the task of text classification applied to four datasets. The researchers observed that MNB and TWCNB obtained better performance when using TF-IDF frequency compared to a simple bag-of-words approach. In the experiments I present in this thesis using MNB, I do not investigate the influence of TF-IDF as discussed by Gebre et al. (2013) and I leave this aspect for future work.

SVMs, on the other hand, are non-probabilistic linear binary classifiers. An SVM model can be understood as points in multidimensional space. These points are mapped and the algorithm tries to find the best possible plane that maximally separates these points. In this section, I adapt here the description and reproduce the illustration from Ben-Hur and Weston (2010).

A linear classifier is based on a linear discriminant function such as:

$$f(x) = w^T x + b \tag{3.17}$$

Where w is known as the *weight vector*, and b is known as the *bias*. For the sake of simplicity imagine that we have two classes: +1 (positive) and -1 (negative) and that $b = 0$. The bias b translates the hyperplane dividing the space into two according to the function presented in $f(x)$ (Equation 3.17). The set of points x if $w^T x = 0$ will all be perpendicular to the weight vector w (see Figure 3.1 for a graphical representation).

SVMs have been used for a wide range of text categorization problems obtaining very good results. Joachims (1998) presents four arguments in favour of the use of SVM in text classification which are summarized below. SVMs can help overcome some of the challenges presented by textual processing tasks, including:

- High dimensional input space: According to the literature, SVMs use overfitting protection, which is independent from the number of features (Joachims,

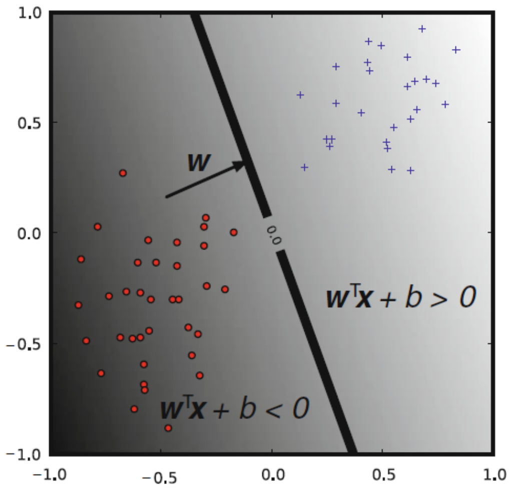

Figure 3.1: A linear classifier. The hyper-plane (line in 2-d) is the classifier's decision boundary (Ben-Hur and Weston, 2010)

1998). This means that SVMs are able to handle large feature spaces. This is particularly helpful when classifying texts because algorithms often have to deal with a very large amount of features. For example, in bag-of-words models every different word (type) in a dataset is represented as an independent feature resulting in a very large feature space.[47]

- Few irrelevant features: In text classification, there are often very few features which can be considered completely irrelevant for classification. A good classifier should therefore combine many features to learn a concept. A more coarse or aggressive feature selection that disregards many features may eventually result in a loss of information.

- Sparse document vectors: SVMs handle sparse vectors very well. Imagine a BoW model built for a big dataset containing a very large vocabulary of 10,000 types for example. If each document in this dataset contains only a few sentences and therefore a few hundreds of tokens, each document will be represented by a vector containing only few entries which are not zero.

- Finding linear separators: In text classification, most classes are linearly separable as it is the case of language (variety) identification. As the central idea behind SVM classifiers is to find these linear separators of classes, SVMs are appropriate for this task.

Finally, another algorithm I use in this thesis is the J48 algorithm. The J48 algorithm is a decision tree algorithm which is an adaptation of the popular C4.5

[47] A more detailed explanation about bag of words is found in Section 3.4.

classifier developed by Quinlan (1993). C4.5 itself is an extension of the ID3 (Iterative Dichotomiser 3) algorithm developed by the same author. C4.5 builds decision trees using the concept of information entropy (a measure of uncertainty of a random variable) and information gain.

Decision trees are simple and easily interpretable classification algorithms that take decisions on a top-down approach. The decision process starts from the root node of the tree. Conditions are tested in each node and the decision process follows the appropriate branch based on the outcome of each test. It proceeds either to another internal node (branch) to test other conditions or directly to a leaf node which corresponds to the label assigned to each instance (output).

The famous 'play tennis given the weather forecast' decision tree is a very good example of how conditions are tested in a decision tree. There are two labels to be attributed to each instance, *yes* or *no*, given a set of conditions determined by three attributes *humidity, outlook*, and *wind*. In this example (see Figure 3.2) each attribute has two or three values. Conditions are tested to determined whether given the weather forecast it is best to play tennis or not (e.g. if outlook is sunny and humidity normal: play tennis; if the outlook is rain and the windy is strong then don't play tennis).

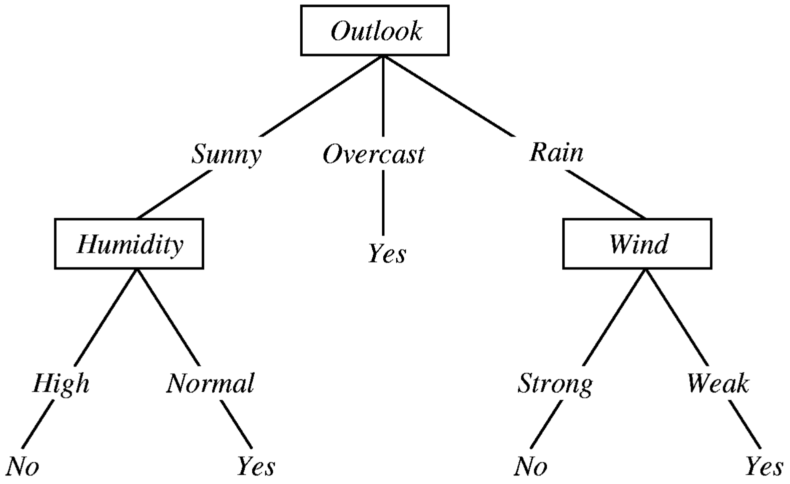

Figure 3.2: Decision tree: play tennis example (Mitchell, 1997)

The particular implementation I use in this thesis is the aforementioned variation of the ID3 algorithm.

As is the case with most decision tree classifiers, in the experiments presented in this thesis the J48 algorithm was significantly slower and did not achieve the same performance as the other algorithms (I also demonstrated a similar outcome for word sense disambiguation in Zampieri (2012)). Even so, I contend that decision trees are worth testing for various reasons. One of the reasons is that in the last few years there has been a revival of decision tree-based methods in NLP due to the application of random forest (RF) classifiers (Liaw and Wiener, 2002). RFs are ensemble classifiers that combine the output of multiple decision trees to obtain more robust performance. Another well-known advantage of decision trees, and another reason to consider them in my experiments, is that the output of these classifiers can be interpreted much more easily than other machine learning algorithms.

3.6 Preliminary Experiments

A number of experiments were carried out in this thesis and will be presented from Chapter 4 onwards. The remainder of this chapter features two preliminary experiments necessary to validate: 1) the likelihood estimation method; 2) the text samples I use throughout this thesis.

With the first preliminary experiment I aim to confirm whether the method proposed in Zampieri and Gebre (2012) which obtained almost perfect results for Portuguese varieties delivers performance comparable to other methods proposed for the same task. With the second experiment I aim to test the variation of the chosen corpora comparing the performance of algorithms when discriminating between A) texts from two different newspapers from different countries; and B) texts from two different newspapers from the same country.

In the first preliminary experiment I compare the performance of likelihood estimation with the performance of an existing approach to discriminate between similar languages (Ljubešić et al., 2007; Tiedemann and Ljubešić, 2012). There were no datasets compiled for the purpose of discriminating between similar languages before the release of the DSLCC (see Chapter 5). One exception is the work by Tiedemann and Ljubešić (2012) who along with the system description paper, released a test set including Bosnian, Croatian, and Serbian texts. I use this dataset to compare their methods with the likelihood estimation. One of the reasons that motivated me to organize the DSL shared task is to facilitate reproducibility which for this particular task was not feasible before 2014. The shared task and the DSLCC filled this gap.

The second preliminary experiment consists of applying a classification method to distinguish texts from the same language variety. The corpora used to calculate the language models are journalistic and compiled with random samples of different topics to avoid thematic bias in classification. Nevertheless, I try to investigate whether this variable interferes in the classification by comparing the classifiers' performance discriminating between texts published in two newspapers of the same language variety.[48]

[48] This preliminary experiment is similar to the one proposed at the DEFT 2010 shared task presented in Chapter 1.

3.6.1 Validating Likelihood Estimation

As a first preliminary experiment, it seems reasonable to compare the performance of the Likelihood Estimation algorithm with other state-of-the-art methods. As explained in Chapter 1, when I began my PhD research there were only a few methods developed for language varieties and closely related languages. Most of these are not available to enable comparison with the exact same dataset. Ljubešić et al. (2007) is an exception to the rule. The dataset[49] used in their experiments was made available online and contains a total of 600 documents, 200 texts from each language each text containing 500 tokens on average. The languages used in this experiment are Croatian (HRV), Bosnian (BIH), and transliterated Serbian (SRB).

To validate the performance of likelihood estimation, I compare the results obtained by this method to the results obtained by the two methods described in Ljubešić et al. (2007) and later in Tiedemann and Ljubešić (2012): Markov-chain and a Naive Bayes classifier.[50] The confusion matrices for the three methods are presented in Table 3.3.

Likelihood Estimation				**Markov-chain**				**Naive-Bayes**			
Doc	Predicted			Doc	Predicted			Doc	Predicted		
Lang	**BIH**	**HRV**	**SRB**	Lang	**BIH**	**HRV**	**SRB**	Lang	**BIH**	**HRV**	**SRB**
BIH	(182)	16	2	**BIH**	(173)	17	10	**BIH**	(181)	11	8
HRV	12	(188)	0	**HRV**	30	(170)	0	**HRV**	7	(193)	0
SRB	22	2	(176)	**SRB**	1	0	(199)	**SRB**	0	0	(200)

Table 3.3: Comparison Between Likelihood Estimation and Ljubesic et al. (2007)

The likelihood estimation scored 91.0% accuracy using character 4-grams, the Markov-chain based method reached 90.3% and Naive-Bayes outperformed both methods by reaching 93.7% accuracy. By this comparison, I contend that likelihood estimation, used in several experiments in this thesis, delivers comparable performance to other state-of-the-art methods developed to discriminate between language varieties and closely related languages. More comparisons involving likelihood estimation that support this claim (including processing time) are presented in Chapter 5. Serbian, Croatian and Bosnian language models will be used again in Chapter 4 in experiments emulating a real-world language identification setting.

[49] http://www.nljubesic.net/resources/tools/bs-hr-sr-language-identifier/ (last seen October 2015)

[50] Note that I used a different setting from the one in Tiedemann and Ljubešić (2012). They trained their model using another dataset and evaluate with these 200 instances. I run the likelihood estimation classifier in a 10-fold cross validation.

3.6.2 Topic Influence in Monolingual Settings

The second preliminary experiment is used to validate the dataset I collected (presented in Chapter 2) and the structure of the task itself. Most studies on identification of language varieties, including those I present in my thesis, use standard corpora sampled from newspapers and magazines (Huang and Lee, 2008). They do not, however, address the question of textual genres and stylistics that underlie these samples. A study that shed light in this direction is undertaken by Lui and Cook (2013) who compared the influence of different datasets on a method to distinguish Australian, British, and Canadian English.

Based on what was discussed in the previous chapters, I understand that the language contained in newspapers is the closest we can get to the written standard of a language or language variety. Moreover, newspaper topics (e.g. sports, politics, economics) are balanced in these samples, and this should ideally decrease the thematic influence on the classification results. Even so, I want to investigate the extent to which sampling influences classification. I do so by using two samples of the same language in a controlled experiment partially presented in Zampieri et al. (2012).

In this controlled experiment, likelihood estimation was first trained to distinguish between two Peninsular Spanish corpora, one of them containing texts from *El País* (PAI) and the other with texts from *El Mundo* (MUN) published in 2008. Secondly, the likelihood estimation algorithm was trained to distinguish two British English corpora, one of them from *The Guardian* (GUA) and the other from *The Independent* (IND). I subsequently compared the results obtained in the controlled experiments with the results obtained when classifying Argentinian texts against Peninsular Spanish texts, and American English against British English.

The features used for this experiment were word unigrams and character trigrams and 4-grams on a set of 1,000 documents divided into 500 for training and 500 for testing. Accuracy results are shown in Table 3.4.

Feature	ARGxESP	PAIxMUN	Difference	GBRxUSA	GUAxIND	Difference
W 1-grams	0.948	0.614	-33.4%	0.903	0.726	-17.7%
C 3-grams	0.948	0.654	-29.4%	0.969	0.707	-26.2%
C 4-grams	0.944	0.728	-21.6%	0.911	0.787	-12.4%

Table 3.4: Classification Using Newspapers from the Same Variety for English and Spanish

For Spanish, results are 21.6 to 33.4 percentage points worse than the classification of Argentinian and Peninsular Spanish depending on the features used. For English, results obtained when discriminating between the two British samples were 12.4 to 26.2 percentage points poorer when compared with the classification results involving American English. The poor results obtained when discriminating between newspapers of the same country suggest that language varieties present a stronger variation than the difference in style and topic bias between two newspapers from the same countries. Therefore, it is safe to assume that the classifiers are actually detecting diatopic variation and discriminating between language varieties and not between newspapers.

It is well known that named entities influence the performance of classifiers. For example, words like *Madrid* or *Barcelona* will be more frequent in texts from Spain than in those from Argentina which instead will have a high frequency of other named entities like *Buenos Aires* or *Boca Juniors*. Nevertheless, based on this experiment it seems reasonable to assume that even though text types and genres may to a certain extent influence classification performance, the impact of the style and topics of newspapers is not strong enough to overcome the differences between language varieties. I return to this question by looking in more detail at the influence of named entities in Chapter 5 using the DSL corpus collection.

3.7 Chapter Summary

This chapter presented the evaluation methods and the computational techniques used in this thesis as well as two preliminary experiments that pave the way for the experiments that will be presented in the remainder of this thesis.

I first described in detail evaluation metrics widely used to evaluate classification performance in NLP, namely: accuracy, precision, recall, and f-score. Subsequently I explained the calculation of *n*-grams and bag-of-words models which are used as features for classification, and finally the classification algorithms used in this thesis: Decision Tree, Likelihood Estimation, Naive Bayes, and SVMs. As in this thesis I am interested in applying existing computational methods rather than developing new methods, for the sake of brevity and scope, this chapter presented a concise and not exhaustive explanation on the classification algorithms. References to literature on these methods are, however, abundant in this chapter and throughout the thesis.

I closed this chapter by proposing two preliminary experiments. The first compares the performance of LE to the two methods proposed by Tiedemann and Ljubešić (2012). In this experiment LE scored 91.0% accuracy compared to 90.3% accuracy obtained by the Markov-chain classifier and 93.7% accuracy obtained by the Naive-Bayes method, thus confirming that the LE delivers competitive performance compared to state-of-the-art algorithms. The second preliminary experiment compares the performance of LE for identifying language varieties to its performance for identifying newspapers from the same country. Results are 33.4 percentage points higher for Argentinian and Peninsular Spanish texts in comparison to two newspapers from Spain and 26.2 percentage points higher for American and British

English texts in comparison to two newspapers published in Britain. This confirms that diatopic variation is indeed much stronger that stylistic differences between newspapers.

This chapter provided the first indication to answer **RQ1:** Is it possible to automatically discriminate between language varieties with satisfactory performance? Results discriminating between Argentinian and Peninsular Spanish and British and American English confirm that the task is feasible for these two languages. I investigate this question in more detail in the next chapters.

Chapter 4

Discriminating Language Varieties Using Words and Characters

4.1 Introduction

Words and characters are the most common features used in state-of-the-art language identification systems as well as in many other text classification tasks. They have shown to deliver the best performance so far in language identification system dealing with several languages and various text types (Lui and Baldwin, 2011, 2012; Brown, 2013, 2014; Simões et al., 2014). For this reason I investigate the application of word-based and character-based methods for the task of discriminating between language varieties.

This chapter presents the results of a number of classification experiments using word and character n-gram models as well as bag-of-words for discriminating between varieties of pluricentric languages such as French, Spanish, and Portuguese. The findings of this chapter address **RQ3:** What are the most efficient features and algorithms to discriminate between language varieties?

Using the likelihood estimation method, I integrate several language varieties to a real-world language identification setting (containing up to 17 classes). The question of integrating language varieties to general-purpose language identification systems has been, to the best of my knowledge, mostly neglected and my work contributes to this body of research by answering **RQ2:** Can language varieties be integrated into real-world language identification systems?

A part of the research presented in this chapter is based on research papers I presented in international conferences which appeared in peer-reviewed conference proceedings (Zampieri and Gebre, 2012; Zampieri et al., 2012; Zampieri, 2013). Unless otherwise specified, all texts used in the following experiments were randomly sampled from the journalistic corpora previously presented in detail in Section 2.4.

For the sake of clarity, in this chapter, every experiment contains a table describing the experimental settings (see Table 4.1 for an example). In these tables I include the most important information such as: the algorithm(s) used, the features, the amount of documents in total, training/test set split, and the length of each document. All results are reported using either a test set or cross-validation.

Algorithm	Naive Bayes
Features	Word bigrams
Number of Texts	1,000
Training/Test Split	10-fold cross validation
Document Length	max. 500 tokens

Table 4.1: Experimental Setting - Example

The reader will notice that in each section of this chapter I vary the conditions of the experiments by changing the aforementioned experimental settings. This was done to explore the impact of different variables in the performance of computational methods in this task. In the tables presenting results, for the sake of clarity I display the best result in bold whenever appropriate.

Chapter 4 is organized as follows:

- **Section 4.2** presents the results obtained in the experiments discriminating between Brazilian and European Portuguese texts.

- **Section 4.3** describes experiments carried out to distinguish between four Spanish language varieties, namely: Argentina, Mexico, Peru, and Spain.

- **Section 4.4** presents the results obtained in experiments discriminating between French texts published in Canada and France.

- **Section 4.5** discusses the integration of language varieties into real-world language identification settings with multiple languages by presenting a number of experiments.

- **Section 4.6** closes this chapter and summarizes its results and main findings.

4.2 Binary Experiments with Portuguese Varieties

This section discusses the results obtained when discriminating between Brazilian and European Portuguese texts. The research presented here was published in Zampieri and Gebre (2012) and uses the likelihood estimation algorithm and the Portuguese dataset presented in Section 2.4.

Brazilian and European Portuguese are regarded to be substantially different in terms of phonetics, lexicon, syntactic structures, and orthography. In the experiments presented here I try to model syntactic, lexical, and orthographic variation by using different sets of features.

First, I use word unigrams as features to perform classification taking into account lexical differences between the two varieties of Portuguese. Accuracy results are reported using 1,000 texts with a maximum of 300 tokens each.

Algorithm	Likelihood Estimation
Features	Word Unigrams
Number of Texts	1,000
Training/Test Split	50% test - 50% train
Document Length	max. 300 tokens

Table 4.2: Experimental Setting - Likelihood Estimation: Word Unigram Results for Portuguese Varieties

Document Length	**Accuracy**
300 words	0.996

Table 4.3: Results - Likelihood Estimation: Word Unigram Results for Portuguese (Zampieri and Gebre, 2012)

Using a unigram model, likelihood estimation can discriminate between European and Brazilian Portuguese with 99.6% accuracy. It is clear that when using unigram models, named entities play an important role in classification. For example, texts from Portugal will contain many named entities that are almost exclusively used in Portugal and vice-versa (e.g. place names like *Lisboa*, names of important or famous people from Brazil/Portugal). As previously stated, the influence of named entities will be be investigated in more detail in Chapters 5 and 6.

Next I use a word bigram model. I investigate how the length of texts affects the performance of likelihood estimation in this task. The results obtained by the classifier were grouped according to the maximum text length, ranging from 100 to 700 tokens. The best results were obtained with a maximum length of 500 tokens per text. With longer texts the model seems to indicate saturation, as can be seen in Table 4.5.

My explanation for this outcome is related to the corpora used in this experiment. Only a few journalistic texts in both corpora contain more than 500 words. The average text length in the Brazilian sub-corpus is 312.99 tokens whereas the average length of Portuguese texts is 461.48 tokens. Adding these few texts to the classification therefore brings no improvement in the algorithm's performance.

The best results obtained in my experiments for Portuguese relied on character n-gram models and using texts of maximum 300 tokens. Results reached 0.998 accuracy for 4-grams, and they are presented in Table 4.7. As discussed in Chapter 1, character n-grams are shown to be the data representation that achieves the best results in general-purpose language identification.

Algorithm	Likelihood Estimation
Features	Word bigrams
Number of Texts	1,000
Training/Test Split	50% test - 50% train
Document Length	from 100 to 700 tokens

Table 4.4: Experimental Setting - Likelihood Estimation: Text Size and Word Bigrams for Portuguese Varieties

Document Length	**Accuracy**
100 tokens	0.851
200 tokens	0.886
300 tokens	0.889
400 tokens	0.904
500 tokens	**0.912**
600 tokens	**0.912**
700 tokens	0.905

Table 4.5: Results - Likelihood Estimation: Text Size and Word Bigrams for Portuguese Varieties (Zampieri and Gebre, 2012)

The good results obtained by character n-grams in comparison to the word n-gram models indicate that the orthographical differences (along with lexical variation) between Brazilian and European Portuguese are a strong factor for distinguishing these varieties. A discussion on the most informative features in the classification using Portuguese varieties is presented in Chapter 6.

Algorithm	Likelihood Estimation
Features	Word n-grams (2 to 6)
Number of Texts	1,000
Training/Test Split	50% test - 50% train
Document Length	max. 300 tokens

Table 4.6: Experimental Setting - Likelihood Estimation: Character N-grams Results for Portuguese Varieties

N-Grams	Accuracy
2-Grams	0.994
3-Grams	0.996
4-Grams	**0.998**
5-Grams	0.988
6-Grams	0.990

Table 4.7: Likelihood Estimation: Character *n*-gram Results for Portuguese Varieties (Zampieri and Gebre, 2012)

4.3 Multiclass Experiments with Spanish Varieties

In this section I present results obtained using words and characters as features to discriminate between Spanish varieties. These results were previously published in Zampieri et al. (2013). I also explore morphosyntactic variation using knowledge-rich features, representing each token using a combination with POS tags and morphology, to be presented in Chapter 6.

The first experiments I present in this section aim to discriminate between four Spanish varieties (Argentina, Mexico, Peru, and Spain). For these experiments I once again used the likelihood estimation method on a set of 2,000 texts (500 texts from each variety). Texts were randomly sampled from the corpus and split into 50% - 50% resulting in 250 documents for training and 250 documents for testing in each class. Each document contained a maximum of 500 tokens.

I evaluate the classification performance using standard metrics presented in Section 3.2. Precision, recall and f-measure were used in settings containing more than two classes, whereas accuracy was reported for binary classification. Results are presented in Table 4.9.

The results obtained by word bigrams, 0.880 precision, were surprisingly high, outperforming word unigrams and all character *n*-gram models.[51] In Chapter 6, I look in more detail at the most informative features in classification. One possible source of variation to explain this good performance, along with named entities, is the use of compound past tenses composed by an auxiliary plus a main verb (e.g. *han dicho*) that are more prominent in some Spanish varieties than others, and specifically in Peninsular Spanish.

Results range from 0.813 f-measure for character 4-grams to 0.876 f-measure for word bigrams. The results for each class remained constant for all features and

[51] It should be noted that the performance obtained using word unigrams and word bigrams was very similar. Statistical significance tests can be used to evaluate whether the performance obtained by these two groups of features is substantially different.

the Peninsular Spanish class seemed to be the most difficult for the algorithm to identify in this setting. As an example, Table 4.10 presents a confusion matrix with the classification output using character 4-grams in which the algorithm obtained its worst performance. From the 250 texts from Spain used for testing, only 109 were correctly classified, while 140 were tagged as Argentinian and 1 as Mexican.

Algorithm	Likelihood Estimation
Features	Character n-grams (2-5), Word n-grams (1-2)
Number of Texts	500
Training/Test Split	50% test - 50% train
Document Length	max. 500 tokens

Table 4.8: Experimental Setting - Likelihood Estimation: 4-Class Classification Results for Spanish Varieties

Feature	**P**	**R**	**F**
C 2-grams	0.835	0.804	0.819
C 3-grams	0.848	0.806	0.826
C 4-grams	0.842	0.787	0.813
C 5-grams	0.854	0.811	0.832
W 1-grams	0.879	0.848	0.848
W 2-grams	**0.880**	**0.870**	**0.876**

Table 4.9: Likelihood Estimation: 4-Class Classification Results for Spanish Varieties (Zampieri et al., 2013)

Document	Predicted			
Language	**ARG**	**MEX**	**PER**	**SPA**
ARG	(248)	0	0	2
MEX	0	(190)	60	0
PER	0	10	(240)	0
SPA	140	0	1	(109)

Table 4.10: Confusion Matrix: Four Spanish Varieties (Zampieri et al., 2013)

Next I trained models to discriminate between the four language varieties in binary settings this time using 1,000 texts (500 for training and 500 for testing). All the results obtained in binary settings were substantially higher than the 4-class classification setting.

The best results, on average 0.999 accuracy, were obtained when discriminating between texts from Argentina and Mexico. The most difficult language variety pair was again Spain and Argentina in which LE achieved an average result of 0.842 accuracy. Results are reported in terms of accuracy in Table 4.11.

It is once again surprising that very good results were obtained using word bigrams. As presented in the previous section, for Portuguese varieties, word-based models did not obtain as good results as the classification using character n-grams. This could be caused by two factors: 1) a stronger lexical variation among the Spanish varieties; 2) the direct impact of the two different orthographies (Brazilian and European Portuguese) which favours character-based representations when compared to Spanish. The most informative features in classification will be investigated in more detail in Chapter 6.

Feature	**ARGxMEX**	**ARGxPER**	**MEXxPER**	**ESPxARG**	**ESPxMEX**	**ESPxPER**	**Average**
C 2-grams	0.999	0.996	0.860	0.852	0.957	0.940	0.934
C 3-grams	0.999	**1.000**	0.911	0.847	0.987	0.991	0.956
C 4-grams	**1.000**	0.999	0.922	0.827	0.992	**0.996**	0.965
C 5-grams	0.999	0.999	0.927	0.802	0.991	0.993	0.952
W 1-grams	0.999	0.999	0.945	0.851	0.994	0.992	0.963
W 2-grams	0.999	0.997	**0.951**	**0.881**	**0.998**	0.989	**0.969**
Average	0.999	0.998	0.919	0.843	0.986	0.983	0.955

Table 4.11: Likelihood Estimation: Binary Classification for Spanish (Zampieri et al., 2013)

4.4 Binary Experiments with French Varieties

As presented in Zampieri et al. (2012), this section reports results obtained when discriminating between French texts from Canada (Quebec) and France. Experiments were carried out using likelihood estimation and a set of 1,000 documents randomly sampled from the corpus and divided in 500 texts for training and 500 documents for testing. Results are reported in terms of accuracy in Table 4.13.

The results suggest that, on average, these two French corpora have a stronger variation than, for example, the Spanish varieties of Argentina and Spain presented in the previous section. French scores were higher in most groups of features except character 5-grams and 6-grams. The results for French are, however, lower than those obtained for Portuguese, which reached 0.998 accuracy for character 4-grams.

Algorithm	Likelihood Estimation
Features	Character n-grams (2-6), Word n-grams (1-2)
Number of Texts	1,000
Training/Test Split	50% test - 50% train
Document Length	max. 300 tokens

Table 4.12: Experimental Setting - Likelihood Estimation: Binary Classification for French

Feature	FRA x CAN
Word 1-grams	0.968
Word 2-grams	0.956
Character 2-grams	0.956
Character 3-grams	**0.990**
Character 4-grams	0.968
Character 5-grams	0.960
Character 6-grams	0.934

Table 4.13: Likelihood Estimation: Binary Classification for French

4.5 Towards a Real-world Setting

So far I have presented experiments to evaluate the method and data I am using, as well as the results of classification in monolingual settings. In this section I present experiments to discriminate between language varieties in real-world classification settings which aim to answer one of the research questions raised at the beginning of this thesis **RQ2**: Can language varieties be integrated into real-world language identification systems?

Real-world language identification systems contain a larger number of languages than the experiments presented so far. To simulate a realistic scenario I included languages that are not pluricentric and I evaluate whether or not it is possible to discriminate between language varieties in a multilingual setting without substantial performance loss. After each experiment, I decrease the complexity of the classification model by grouping classes together and analysing the results obtained. The first experiment uses 17 languages, containing 8 language varieties and 3 closely related languages.

In the second experiment, I grouped the eight language varieties together resulting in four different classes. This results in an experimental setting that is frequently used by general-purpose language identification methods that do not make any distinction between varieties.

In the third and last experiment, I group Bosnian, Serbian and Croatian as one class: Serbo-Croatian. This was done for evaluation purposes, and by looking at the results, it seems clear that the identification of these three closely-related languages is another important challenge for language identification systems as discussed in Ljubešić et al. (2007); Tiedemann and Ljubešić (2012).

4.5.1 Towards Multilingual Classification

The integration of language varieties into broader identification settings has been mostly neglected. First, because including language varieties worsens performance of language identification systems and second because for many NLP applications, predicting the language of the text (irrespective of the language variety) is sufficient. This is, however, not the case for all NLP applications that might take advantage of language variety information for further processing.

A first experiment towards a multilingual classification setting was previously published in Zampieri et al. (2012). The results suggest that it is, to a certain extent, possible to include language varieties in a multilingual setting without a substantial loss in performance.

To evaluate this, I trained the likelihood estimation to discriminate between the six language varieties, namely: Spain and Argentina, France and Quebec, and Brazil and Portugal. The experiment uses the same setting of the previous experiments with 1,000 texts from each language variety split into 500 documents for training and 500 documents for testing. Results for this six-class experiment are presented in terms of accuracy, recall, precision and f-measure.

Algorithm	Likelihood Estimation
Features	Character n-grams (2-6), Word n-grams (1-2)
Number of Texts	1,000
Training/Test Split	50% test - 50% train
Document Length	max. 300 tokens

Table 4.14: Experimental Setting - Likelihood Estimation: Binary Classification for French

Feature	P	R	F
W 1-grams	0.917	0.905	0.911
W 2-grams	0.878	0.866	0.872
C 2-grams	0.898	0.880	0.889
C 3-grams	**0.947**	**0.933**	**0.940**
C 4-grams	0.910	0.890	0.899
C 5-grams	0.924	0.905	0.915
C 6-grams	0.935	0.932	0.933

Table 4.15: Likelihood Estimation: 6-Class Classification for French, Portuguese and Spanish

Performance drops in comparison to binary settings which is expected. Nevertheless the algorithm is still able to discriminate between these six language varieties with 94.7% precision and 94.0% f-score using character trigrams, which are often the best performing features in general-purpose language identification as well. These results pave the way for further investigations including more languages as presented below.

4.5.2 17-class Classification Experiment

The 17-class experiment is the largest experiment carried out in this section. For this experiment I used the likelihood estimation method and the same 1,000 texts from each language used in the previous experiments but this time split into 800 documents for training and 200 documents for testing. Language models were calculated using character trigrams, the features which performed best in the previous experiment. The average results for 17-class classification reached 88.9% f-measure. Results are presented in terms of precision, accuracy and f-measure in Table 4.17. The experimental settings are presented in Table 4.16.

The confusion matrix in Table 4.18 shows that when confronted with pluricentric languages, the algorithm tends to make generalizations towards one class and labels most texts as one of them. The only exception is the very good performance obtained by the algorithm when discriminating French texts from France and Canada with only 8 misclassified instances.

Algorithm	Likelihood Estimation
Features	Character *n*-grams (2-6), Word *n*-grams (1-2)
Number of Texts	1,000
Training/Test Split	20 % test - 80 % train
Document Length	max. 300 tokens

Table 4.16: Experimental Setting - Likelihood Estimation: Performance per Class using Character 3-grams for 17-class Multilingual Classification

Class	**P**	**R**	**F**	**Class**	**P**	**R**	**F**
Albanian	**1.0**	0.990	0.994	Greek	**1.0**	0.990	0.994
American	0.915	0.490	0.638	Macedonian	0.990	0.995	0.992
Argentina	0.737	0.995	0.846	Portugal	0.943	**1.0**	0.970
Bosnian	0.636	0.595	0.615	Quebec	0.961	**1.0**	0.980
Brazil	**1.0**	0.940	0.969	Romanian	0.995	0.995	0.995
British	0.653	0.960	0.777	Serbian	0.767	0.760	0.763
Bulgarian	0.994	0.990	0.992	Spain	0.992	0.655	0.789
Croatian	0.652	0.705	0.677	Turkish	0.995	**1.0**	**0.997**
France	1.0	0.955	0.976	**Average**	0.896	0.883	0.889

Table 4.17: Likelihood Estimation: Performance per Class using Character 3-grams for 17-class Multilingual Classification

The aforementioned generalization in favour of one language variety results in high recall and poor precision for one class and poor precision and high recall for its counterpart. This happens in the case of American and British English, where British is preferred; Brazilian and European Portuguese, in favour of the European class; Peninsular and Argentinian Spanish, preferring the Argentinian class and finally in favour of Quebecian French in comparison to the Hexagonal variety.

To evaluate the loss of performance in the real-world setting, I took the scores presented in Table 4.17 and I calculate the average results for each pluricentric language and by a Serbo-Croatian class consisting of Bosnian, Serbian and Croatian. All classes presented a loss in performance, which was an expected outcome. However, for French and Portuguese, as can be seen in Table 4.19, this loss of performance is not substantial (less them two percentage points).

It is important to note that while French and Portuguese scored well over 98% accuracy in the binary classification experiments, Spanish, English and the 3-fold classification of Serbo-Croatian obtained a poorer performance. This outcome suggests firstly that Portuguese and French present a stronger variation in the two corpora used for these experiments. Secondly, these results suggest that there seems to be a threshold that might indicate whether or not two or more varieties can be integrated into broader classification settings.

Doc	**Predicted**																
Lang	ALB	USA	ARG	BIH	BRA	GBR	BUL	HRV	FRA	GRC	MKD	POR	CAN	ROU	SRB	ESP	TUR
ALB	(198)	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
USA	0	(98)	0	0	0	102	0	0	0	0	0	0	0	0	0	0	0
ARG	0	0	(199)	0	0	0	0	0	0	0	0	0	0	0	0	1	0
BIH	0	0	0	(119)	0	0	0	53	0	0	0	0	0	0	28	0	0
BRA	0	0	0	0	(188)	0	0	0	0	0	0	12	0	0	0	0	0
GBR	0	8	0	0	0	(192)	0	0	0	0	0	0	0	0	0	0	0
BUL	0	0	0	0	0	0	(192)	0	0	0	2	0	0	0	0	0	0
HRV	0	1	0	40	0	0	0	(141)	0	0	0	0	0	0	18	0	0
FRA	0	0	1	0	0	0	0	0	(191)	0	0	0	8	0	0	0	0
GRC	0	0	0	0	0	0	0	1	0	(198)	0	0	0	0	0	0	1
MKD	0	0	0	0	0	0	1	0	0	0	(199)	0	0	0	0	0	0
POR	0	0	0	0	0	0	0	0	0	0	0	(200)	0	0	0	0	0
CAN	0	0	0	0	0	0	0	0	0	0	0	0	(200)	0	0	0	0
ROU	0	0	0	1	0	0	0	0	0	0	0	0	0	(199)	0	0	0
SRB	0	0	1	26	0	0	0	21	0	0	0	0	0	0	(152)	0	0
ESP	0	0	69	0	0	0	0	0	0	0	0	0	0	0	0	(131)	0
TUR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	(200)

Table 4.18: Likelihood Estimation: Full Confusion Matrix for Multilingual Classification

Class	**P**	**R**	**F**
English	0.784	0.725	0.753
Spanish	0.864	0.825	0.844
Portuguese	0.971	0.970	0.970
French	**0.980**	**0.977**	**0.979**
Serbo-croatian	0.685	0.687	0.686

Table 4.19: Likelihood Estimation: Average Performance for Pluricentric Languages in a Monolingual Setting

4.5.3 Classic Language Identification Setting

I proceed by modelling pluricentric languages as unique classes, namely: French (FRE), Portuguese (PTG), English (ENG) and Spanish (SPN). This is a classic language identification experiment such as those presented by Martins and Silva (2005) or Lui and Baldwin (2012). This results in a 13-class experiment reaching 92.2% accuracy and f-measure. The confusion matrix is presented in Table 4.20. The experimental settings are the same as presented in Table 4.16.

Doc	**Predicted**												
Lang	ALB	BIH	BUL	HRV	ENG	FRE	GRC	MAK	PTG	ROU	SRB	SPN	TUR
ALB	(199)	0	0	0	1	0	0	0	0	0	0	0	0
BIH	0	(112)	0	47	0	0	0	0	0	0	41	0	0
BUL	0	0	(199)	0	0	0	0	1	0	0	0	0	0
HRV	0	27	0	(144)	0	0	0	0	0	3	26	0	0
ENG	0	0	0	0	(200)	0	0	0	0	0	0	0	0
FRE	0	0	0	0	0	(199)	0	0	0	0	0	1	0
GRC	0	0	0	0	1	0	(198)	0	0	1	0	0	0
MAK	0	0	1	0	0	0	0	(199)	0	0	0	0	0
PTG	0	0	0	0	0	0	0	0	(200)	0	0	0	0
ROU	0	0	0	0	0	0	0	0	0	(200)	0	0	0
SRB	0	30	0	20	0	0	0	0	0	0	(150)	0	0
SPN	0	0	0	0	0	1	0	0	1	0	0	(198)	0
TUR	0	0	0	0	0	0	0	0	0	1	0	0	(199)

Table 4.20: Likelihood Estimation: Confusion Matrix for Multilingual 13-class Classification Without Language Varieties

In this setting, the model performs well for most classes except Bosnian, Croatian and Serbian. A degree of confusion is expected between Portuguese and Spanish, for example, but the model proved to be reliable in discriminating between the two languages.

4.5.4 Evaluating Serbo-Croatian

The main weakness of the results obtained by the likelihood estimation method is the discrimination between Bosnian, Croatian and Serbian. My last experiment in this section consists of grouping Bosnian, Serbian and Croatian together as a unique Serbo-Croatian (SBC) class. In an 11-class classification likelihood estimation scored 99.5% f-measure. The confusion matrix is presented in Table 4.21.

As pointed out by Palmer (2010), language identification performs almost perfectly when not dealing with similar languages. Grouping Bosnian, Serbian and Croatian as a unique class, confirms this hypothesis.

This section is a first step towards the evaluation of language varieties in broader language identification settings. The results show that it is possible to between distinguish language varieties in real-world settings with moderate loss in performance

Doc	**Predicted**										
Lang	ALB	BUL	ENG	FRE	GRC	MAK	POR	ROU	SBC	SPA	TUR
ALB	(198)	0	0	0	0	0	0	1	1	0	0
BUL	0	(200)	0	0	0	0	0	0	0	0	0
ENG	0	0	(200)	0	0	0	0	0	0	0	0
FRE	0	0	0	(200)	0	0	0	0	0	0	0
GRC	0	0	0	0	(200)	0	0	0	0	0	0
MAC	0	1	0	0	0	(199)	0	0	0	0	0
PTG	0	0	0	0	0	0	(200)	0	0	0	0
ROU	0	0	0	0	0	0	0	(200)	0	0	0
SBC	0	0	0	0	0	0	0	2	(198)	0	0
SPA	0	0	1	0	0	0	0	2	0	(197)	0
TUR	0	0	0	0	0	0	0	1	1	0	(198)

Table 4.21: Likelihood Estimation: Confusion Matrix for Multilingual Classification 11-class Classification with Serbo-Croatian as one Class

for some languages, specifically French and Portuguese. The results seem to indicate that for these two languages, it is possible to integrate at least the two varieties studied in real-word applications without strong loss in performance. French and Portuguese also have a strong presence in Africa and it is still an open question whether African varieties of French and Portuguese can be distinguished from the two varieties studied here and later integrated with real-world systems.

For other languages, such as English and Spanish, the results indicate a more important decrease in performance. It should be noted that in the binary settings, English and Spanish varieties also presented poorer results when compared to French and Portuguese varieties. This seems to indicate that the algorithm needs a certain threshold of confidence to distinguish varieties in real-world settings. When results are below a certain level, loss in performance becomes substantial. Future work may explore other settings and algorithms to estimate what this threshold would be.

In all cases, the performance is different from the one obtained in monolingual settings (including the Serbo-Croatian classification), which was an expected outcome.

The experiments presented here expand the state-of-the-art in language identification by evaluating the integration of language varieties and similar languages in real-world experiments. To the best of my knowledge this is exploring new ground and filling a gap in this area of research.

4.6 Bag-of-words Model

In this section, I compare the performance of systems using a combination of machine learning and bag-of-words to the performance of likelihood estimation using n-grams as features. I compare the results using BoW to the best n-gram language models as well as to word unigram models.[52] Bag-of-words have been widely used in text categorization problems as previously discussed in Chapter 3. The study published by Huang and Lee (2008) apply BoW similarity to discriminate between Chinese varieties and is to my knowledge one of the few studies to propose these features for language variety identification.

A number of experiments were carried out and are presented in this section, such as results obtained when discriminating between European and Brazilian Portuguese; Argentinian, Mexican, Peruvian and Peninsular Spanish and Hexagonal and Quebec French. Results presented in this section were previously published in Zampieri (2013).

The results presented in this section were obtained using 1,000 documents per language sampled from the corpora listed in Chapter 2 and split into 2 parts, each containing 500 documents, one half for training and one for testing. This amount of data was used to compare the performance of the machine learning methods to the likelihood estimation method presented in the previous sections. The accuracy results obtained using bag-of-words and three algorithms: SVM, MNB, and J48 are presented in Table 4.23.

Algorithm	SVM, MNB and J48
Features	Bag-of-words
Number of Texts	1,000
Training/Test Split	50 % test - 50 % train
Document Length	max. 300 tokens

Table 4.22: SVM, MNB, and J48: BoW Classification Results for Portuguese, Spanish and French Varieties

Language	**Classes**	**MNB**	**SVM**	**J48**
Portuguese	2	**0.988**	0.987	0.942
Spanish	4	**0.943**	0.936	0.865
French	2	**0.972**	0.955	0.950
Average	2.66	**0.968**	0.959	0.919

Table 4.23: SVM, MNB, and J48: BoW Classification Results for Portuguese, Spanish and French Varieties (Zampieri, 2013)

The best result is 0.988 accuracy obtained with MNB for Portuguese and the worst is 0.865 for Spanish using the J48 classifier. The best average performance was obtained by MNB with 0.968 accuracy. Table 4.24 presents the comparison between

[52] The difference between a word unigram model and a BoW model in my experiments is that unigrams are calculated using smoothing.

the results of the BoW approach (all obtained using MNB) and the best results obtained with likelihood estimation. Results obtained using BoW and the Naive Bayes classifier are 6.7 percentage points higher for Spanish (4 classes), 1 percentage point lower for Portuguese and 1.8 percentage points lower for French.

Language	**BoW + MNB**	**LM + LE**	**Best Feature**	**Difference**
Portuguese	0.988	0.998	C 4-grams	- 1.0 pp
Spanish	0.943	0.876	W 2-grams	+ 6.7 pp
French	0.972	0.990	C 3-grams	- 1.8 pp

Table 4.24: SVM, MNB and J48: Comparison Between BoW and Likelihood Estimation n-gram-based Method for Portuguese, Spanish and French (Zampieri, 2013)

It is important to point out that the best results obtained with the likelihood estimation with n-grams as features use different feature sets: two of them rely on characters and one relies on word bigrams. Portuguese varieties, for example, have moderate differences in orthography which favours character-based representation. On the other hand, word bigrams take syntax into account and this is an aspect of language that bag-of-words is not able to capture.

To allow for a more fair comparison, in Table 4.25 I present the best results obtained by MNB using BoW and the best results obtained by likelihood estimation using word unigrams.

Language	**BoW + MNB**	**LM + LE**	**Difference**
Portuguese	0.988	0.996	- 0.8 pp
Spanish	0.943	0.848	+ 9.5 pp
French	0.972	0.968	+ 0.4 pp

Table 4.25: SVM, MNB and J48: Comparison Between BoW and Likelihood Estimation Word unigram-based Method for Portuguese, Spanish and French (Zampieri, 2013)

BoW and word unigrams are very similar types of text representation. The main differences between word unigrams and the BoW approach presented is this section are probability calculation and smoothing. The likelihood algorithm uses Laplace probability distribution with add-one smoothing for unseen words whereas in the BoW model no smoothing technique is used.

As BoW methods are simpler than n-gram language models I believe that the results presented in this section are very interesting for the discrimination of language varieties, language identification and more broadly to text classification. The question of whether BoW can be used to train general-purpose language identification systems in real-world identification settings is still open.

With the exception of Huang and Lee (2008), BoW were not substantially explored for this task. The experiments present in this section fill this gap. Results show that the BoW method combined with machine learning classifiers achieves performance comparable to n-gram-based methods. A small loss of performance was observed for Portuguese and French in binary settings and a substantial performance gain was observed for Spanish in a multi-class classification setting.

4.7 Chapter Summary

This chapter presented the results obtained by several experiments using words and characters as features for language variety identification. The main objective of this chapter is to provide empirical evidence to answer research questions **RQ1**, **RQ2** and partially **RQ3**. To recapitulate, below I present the first three of the four research questions posed in the introduction of this thesis and a tentative answer to each of them.

- **RQ1:** Is it possible to automatically discriminate between language varieties with satisfactory performance?
  For the three major languages represented in my experiments the best results obtained using the likelihood estimation method reached 99.8% accuracy in discriminating between Brazilian Portuguese and European Portuguese texts using character 4-grams, 100% accuracy in discriminating Argentinian Spanish and Peruvian Spanish texts using character 4-grams, 100% accuracy discriminating between Argentinian Spanish and Mexican Spanish using character 4-grams, and 99% for Hexagonal French and Canadian French using character 3-grams. Based on these results, I can state that it is possible to automatically discriminate between language varieties using standard texts without name entity removal with close to perfect or perfect performance in binary settings. The results obtained for Spanish in a multi class classification setting are, however, lower. Two issues have to be taken into account when answering **RQ1**. One of them is the impact of text length. In the experiments presented in this chapter I used complete texts that are often several sentences long. These texts contain enough distinctive features for algorithms to discriminate between the varieties. The second aspect, is the impact of the named entities, which influences classification performance. These two issues are investigated in more detail in Chapter 5 using the DSL shared task results.

- **RQ3:** What are the most efficient features and algorithms to discriminate between language varieties?
  The main algorithm used in this chapter is the likelihood estimation method. In the several experiments described in this chapter, I showed that this method

was able to discriminate between language varieties with performance comparable to other state-of-the-art methods, e.g. the ones by Ljubešić et al. (2007); Tiedemann and Ljubešić (2012). I observed that the best features in most classification settings are character 3-grams and 4-grams. For multilingual settings, character 3-grams are the best features. Character 3-grams are also the best performing features in general-purpose language identification. An interesting finding of my experiments is that for language varieties, word-based n-gram models and bag-of-words delivered very good performance as well. This corroborates the findings of Huang and Lee (2008) who used top bag-of-word similarity to discriminate between Chinese varieties. The influence of named entities should be taken into account and the results presented in Chapters 4 and 5 contribute to this analysis. More empirical evidence on the best features and algorithms will be discussed in Chapter 5 and Chapter 6.

Chapter 5

Comparing Approaches: The DSL Shared Task

5.1 Introduction

This chapter compares different approaches to the discrimination of similar languages and language varieties based on the results of the two editions of the Discriminating between Similar Languages (DSL) shared task. The first edition was held at the VarDial workshop at the 2014 edition of COLING[53] and the second edition was organized in 2015 within the scope of the LT4VarDial workshop co-located with RANLP.[54] The information presented in this chapter is partially based on the content of the two shared task reports (Zampieri et al., 2014, 2015b) and my team's submission to the 2015 edition of the shared task (Zampieri et al., 2015a). Tables previously published in these papers are reproduced here with references to the original publications.

The idea behind the shared task came up while I was working on this dissertation. I noticed that there was no common data set to evaluate systems on language variety and similar language identification and that researchers in the field would benefit from such a resource. Moreover, no shared task with this focus had been proposed. There were a few similar shared tasks organized including the aforementioned DEFT2010 which focused only on French varieties (Grouin et al., 2010), the ALTW language identification shared task (Baldwin and Lui, 2010) for general-purpose language identification and the NLI shared task (Tetreault et al., 2013) for native language identification. Therefore in 2014, together with colleagues interested in language identification, dialects and language varieties, and comparable corpora, we proposed the organization of the first DSL shared task and the first workshop

[53] http://corporavm.uni-koeln.de/vardial/sharedtask.html

[54] http://ttg.uni-saarland.de/lt4vardial2015/dsl.html

on NLP for Similar Languages, Varieties and Dialects (VarDial).[55] Following the success of the first edition, we organized a second challenge and a second workshop in 2015.

Shared tasks are an intersting interesting way of comparing systems using a standardized data set and the same evaluation methodology and they often attract a high number of participants. This is evidenced by a number of well-established initiatives such as CLEF[56], Conference and Labs of the Evaluation Forum, and the Semantic Evaluation Exercises (SemEval).[57] The DSL shared task is the first international shared task to tackle the issue of discriminating between similar languages and it received a very positive response from the scientific community. The first edition featured 22 enrolled teams and 8 final submissions and the second edition featured 24 teams subscribed and 10 final submissions.

The aim of the shared task is to investigate systems' performance on discriminating a set of 13 different languages in 6 languages groups clustered by similarity.[58] For the purpose of the shared task we compiled the DSL corpus collection (Tan et al., 2014) that will be presented in more detail in this chapter. The DSLCC is the first comparable corpus collection designed for this purpose and it is freely available for the research community. To date three versions of the DSLCC have been released: v1.0, v2.0, and v2.1. In this chapter I describe the methodology behind the compilation of this useful resource.

This chapter addresses 2 of the 4 research questions of this thesis, namely **RQ2** and **RQ3**. Given the multilingual nature of the DSLCC dataset, the results of this chapter aim to contribute to answer **RQ2:** Can language varieties be integrated into real-world language identification systems?. As there were several systems participating in the two editions of the DSL shared task using a variety of computational approaches, the shared task provides me valuable information to answer **RQ3:** What are the most efficient features and algorithms to discriminate between language varieties?

This chapter is structured as follows:

- **Section 5.2** presents the DSLCC, the first corpus compiled for language vari-

[55] The reader will notice that throughout this chapter I use the plural *we* instead of the singular *I* as in the rest of the thesis. This indicates that the decisions related to the organization of the DSL challenge and the compilation of the dataset were taken jointly by all organizers.

[56] `http://www.clef-initiative.eu/`

[57] `http://alt.qcri.org/semeval2015/`

[58] The languages are different in the 2014 and 2015 challenges and language group codes were used only in the 2014 edition. The 2014 edition featured British and American English which were removed in 2015 when Macedonian and Bulgarian were included. The number of languages, however, remained constant.

ety and similar language identification.

- **Section 5.3** reports the results of the 2014 edition of the DSL shared task based on the shared task report published in Zampieri et al. (2014).
- **Section 5.4** presents the results of the second edition of the DSL shared task held in 2015 (Zampieri et al., 2015b).
- **Section 5.5** compares the performance of likelihood estimation to the other approaches proposed in the second edition of the DSL challenge.
- **Section 5.6** discusses the influence of named entities in this task.
- **Section 5.7** closes this chapter and discusses the main findings of the two editions of the shared task.

5.2 A Corpus for Language Variety Discrimination: The DSL Corpus Collection

The DSL corpus collection was initially compiled for the first DSL shared task and has been used for both the 2014 and 2015 editions. Three versions of the corpus collection have been released so far, the DSL v1.0, v2.0, and v2.1, and all of them are available for the research community. The idea to create such a resource goes beyond the scope of the shared task. We took the opportunity of organizing this evaluation campaign to create a dataset that could be redistributed and used by scholars and developers interested in the processing of similar languages and language varieties. To the best of my knowledge, a corpus with these characteristics was not available before the compilation of the DSLCC.

The DSLCC contains journalistic texts sampled from various existing corpora. The decision to work with journalistic texts is inspired by the discussion presented in Chapter 2 in which I contend that journalistic texts are the most accurate representation of the contemporary written standard of a language. Data for the DSLCC was collected from multiple sources including the SETimes Corpus[59] (Tyers and Alperen, 2010), HC Corpora (Christensen, 2014), and Leipzig Corpora Collection (Richter et al., 2006; Biemann et al., 2007).

There are several aspects to take into account when compiling language resources. One is to ensure that texts can be legally used under copyright law. Unless otherwise stated by the publishers, most texts are protected by copyright law and therefore cannot be redistributed without the permission of their respective copyright owners.

[59] Published in OPUS (Tiedemann, 2012)

In practice this means that anyone interested in compiling and distributing a corpus should first ask permission from the copyright owners of each text. This is, of course, infeasible for a multilingual corpus composed of various data sources such as the DSLCC. We had to make two decisions to avoid this limitation. First we sampled texts that were previously released in other sources under an open or permissive license that allowed us to distribute the material. Secondly, we did not include complete texts in the corpus collection, but short text excerpts comprising one or a few sentences at most. As the corpus collection was developed for research purposes and does not contain complete texts, fair use applies.[60]

As to the data, both versions of the DSL collection contain 18,000 randomly sampled training instances and 2,000 development instances for each language or language variety. For testing, DSLCC v1.0 contains 1,000 test instances for each language or variety, v2.0 contains 2,000 test instances divided into test set A (original texts) and test set B (texts with most named entities removed). The DSLCC v2.1 does not contain a test partition because it was compiled for the purpose of a qualitative study (unshared task) and not a quantitative one. The instances of the DSLCC v1.0 contain at least 20 tokens each without an upper limit, whereas the texts in the DSLCC v.2.0 and v.2.1 contained a minimum of 20 tokens and a maximum of 100. The decision to include texts with a maximum length of 100 tokens was taken to evaluate the impact of text length in classification and to make the task more challenging.

Group	Language/Variety	Code
A	Bosnian	*bs*
	Croatian	*hr*
	Serbian	*sr*
B	Indonesian	*id*
	Malay	*my*
C	Czech	*cz*
	Slovene	*sk*
D	Brazilian Portuguese	*pt-BR*
	European Portuguese	*pt-PT*
E	Argentine Spanish	*es-AR*
	Castilian Spanish	*es-ES*
F	British English	*en-GB*
	American English	*en-US*

Table 5.1: DSLCC v1.0: Language Groups.

The languages included in the DSLCC v1.0 are presented in Table 5.1 with the

[60] The second decision was taken not only because of copyright restrictions, but also to make the shared task even more challenging and interesting for participants.

standardized ISO 639-1 language codes,[61] the total number of documents and tokens written in each language. For language varieties the country code is appended to the ISO code[62] (e.g. *en-GB* refers to the British variety of English).

We decided to select sentences from corpora randomly and I regard the collection as a set of balanced comparable news corpora representative of each language variety. The DSLCC was distributed in tab delimited format; the first column presents a sentence in the language/variety, the second column states its group and the last column refers to its language code.

One of the two new additions in the DSLCC v2.0 and v2.1 is the class *others*. This class contains a mixed collection of documents written in Catalan, Russian, Slovene, and Tagalog. The other modification is that the information regarding language groups was excluded in v2.0 and v2.1 The languages included in the DSLCC v2.0 and v2.1 are presented in Table 5.2, and languages and varieties marked with * are only included in v2.1.

Language/Variety	**Code**
Bosnian	*bs*
Croatian	*hr*
Serbian	*sr*
Indonesian	*id*
Malay	*my*
Czech	*cz*
Slovak	*sk*
Brazilian Portuguese	*pt-BR*
European Portuguese	*pt-PT*
*Macanese Portuguese	*pt-MO*
Argentine Spanish	*es-AR*
Castilian Spanish	*es-ES*
*Mexican Spanish	*es-MX*
Bulgarian	*bg*
Macedonian	*mk*
Others	*xx*

Table 5.2: DSLCC v2.0 and v2.1: Languages Grouped by Similarity.

One important difference that the reader might notice from DSLCC v1.0 to DSLCC v2.0 and v2.1 is the absence of the English group. English is a very interesting example of a pluricentric language that fits well within the scope of the DSL shared task. As discussed in the 2014 shared task report (Zampieri et al., 2014), however,

[61] `http://www.loc.gov/standards/iso639-2/php/English_list.php`

[62] Note that for all the other experiments in this thesis, I used solely the aforementioned ISO 3166 alpha-3 country code.

problems were found in the English corpora that lead us to exclude the English group from both the 2014 edition and the 2015 edition. The problems include cross citation and cross referencing between English corpora that the original data sources in DSLCC did not take into account. This includes, for example, British texts that were republished by American websites and news agencies and tagged as American English.

5.3 DSL Shared Task 2014

In the next sections I will present the results obtained by all participants of the 2014 edition of the shared task who submitted final results as described in the 2014 shared task report (Zampieri et al., 2014).[63,64] The 2014 edition of the DSL shared task included 22 enrolled teams from different countries (e.g. Australia, Estonia, Holland, Germany, United Kingdom and United States). From the 22 enrolled teams, eight submitted their final results for evaluation and five of them submitted papers describing their systems. We provided the opportunity for teams to participate in two kinds of submissions:

- **Closed submission:** Using only the DSLCC v1.0 for training.
- **Open submission:** Using any dataset for training.

Most of the teams used the DSL corpus collection exclusively and therefore only participated in the closed submission track. Two teams compiled other datasets to participate in the open submission track.

Given that the dataset contained misclassified instances, group F (English) was not taken into account to compute the final shared task scores.

5.3.1 Participating Systems

The eight teams that submitted their final runs were invited to submit research papers describing their systems and findings. The top five teams in the closed submission track submitted their papers, namely: NRC-CNRC, RAE, UMich, UniMelb-NLP and QMUL. Brief descriptions of each team's approach are presented below. Teams are ranked by their performance in the closed submission track:

[63] Visit `https://bitbucket.org/alvations/dslsharedtask2014/downloads/dsl-results.html` for more detail on the shared task results or the aforementioned DSL shared task website.

[64] For a comprehensive evaluation of the two editions of the DSL shared task see Goutte et al. (2016).

- The NRC-CNRC team (Goutte et al., 2014) proposed a two-stage classification system to predict first the language group followed by the language or language variety of each instance. The language group is identified using a probabilistic classifier similar to Naive Bayes. Within each group, the languages or language varieties were identified using linear SVM classifiers. The SVMs were trained in a binary setting for groups B-F and one versus all for group A, which contains three languages (Bosnian, Croatian and Serbian).

- The RAE team (Porta and Sancho, 2014) used a hierarchical two-stage classifier to identify first the language group and then the language or language variety of each instance. It uses a Maximum Entropy (MaxEnt) classifier with word n-grams, characters n-grams and a so-called 'white list' of tokens, a list containing words that are exclusive to a language or variety, similar to one of the features that Ranaivo-Malançon (2006) proposed to discriminate between Malay and Indonesian texts.

- UniMelb-NLP (Lui et al., 2014) also proposed a two-stage approach similar to the two previously presented teams. The team explored different forms of text representations including delexicalized representations using a 12-tag universal POS tagger. The team used *langid.py* (Lui and Baldwin, 2012), a general-purpose language identification tool, which makes use of information gain (IG) to select the best features for classification. *langid.py* is generally regarded as being able to achieve performance superior to other language identification tools (e.g. *TextCat*). UniMelb-NLP was one of the two teams who compiled additional training corpora to participate in the open submission track as well.

- UMich (King et al., 2014) submissions used words and characters as features and applied information gain, parallel text feature selection, and a manual feature selection to select the best features for classification. UMich used implementations of three different algorithms available at Mallet (McCallum, 2002): Naive Bayes, Logistic Regression and SVM. UMich also participated in the open and closed submission tracks.

- The QMUL team (Purver, 2014) proposed the use of words and characters as features and a linear SVM classifier. QMUL investigated the influence of the cost parameter c (from 1.0 to 100.0), in the classifiers' performance.

The next section presents the results obtained by the five aforementioned systems plus the three teams who did not submit system descriptions: LIRA, UDE ,and CLCG.[65,66]

[65] Note that in 2014 I did not participate in any of the teams competing in the DSL challenge.

[66] As it happens in most shared tasks, only the best performing team end up publishing system

5.3.2 Results

Table 5.3 presents the results obtained by the eight teams that submitted their results for the closed submission track. Results are ranked according to average accuracy.

Team	**Accuracy**
NRC-CNRC	0.957
RAE	0.947
UMich	0.932
UniMelb-NLP	0.918
QMUL	0.906
LIRA	0.766
UDE	0.681
CLCG	0.453

Table 5.3: DSL Shared Task 2014: Closed Submission Results in an 11 Class Classification Setting (Zampieri et al., 2014)

The performance of five teams (NRC-CNRC, RAE, UMich, UniMelb-NLP, and QMUL) is higher than 90% accuracy and therefore comparable to the performance levels described in the literature for discriminating similar languages and language varieties (Tiedemann and Ljubešić, 2012; Lui and Cook, 2013). The performance obtained by these five teams is above the 88.9% accuracy baseline reported in Tan et al. (2014) before the official release of the DSLCC.[67] Three teams obtained substantially lower scores ranging from 45.33% to 76.64% accuracy.

	CLCG	**LIRA**	**NRC-CNRC**	**QMUL**	**RAE**	**UDE**	**UMich**	**UniMelb-NLP**
A	0.338	0.333	**0.936**	0.879	0.919	0.785	0.919	0.915
B	0.503	0.982	**0.996**	0.935	0.994	0.892	0.992	0.972
C	0.500	**1.000**	**1.000**	0.962	**1.000**	0.493	0.999	**1.000**
D	0.496	0.892	**0.956**	0.905	0.948	0.493	0.926	0.896
E	0.503	0.843	**0.910**	0.865	0.888	0.694	0.876	0.807

Table 5.4: DSL Shared Task 2014: Performance for Language Groups in the Closed Submission in an 11 Class Classification Setting (Zampieri et al., 2014)

Table 5.4 shows the performance of systems in discriminating each language within

description papers. However, as commented in the DSL 2014 report (Zampieri et al., 2014), valuable information could also be obtained by investigating approaches that do not achieve good performance.

[67] The baseline included results of group F, which were not included in the official DSL 2014 scores.

groups A to E in terms of accuracy. Teams are sorted alphabetically and the best score per language group is displayed in bold.

The top five systems plus the LIRA team obtained results above 90% accuracy for groups B (Malay and Indonesian) and C (Czech and Slovak). Half of the teams obtained perfect performance when discriminating Czech and Slovak texts suggesting that texts from these two languages are not as similar as we assumed when compiling the corpora for the DSL shared task.

The results from the shared task confirm that discriminating between Bosnian, Croatian, and Serbian is a very challenging task as discussed in Tiedemann and Ljubešić (2012). For group A, the best result was again obtained by the NRC-CNRC team with 93.5% accuracy. Corroborating the findings presented in Chapter 4, the language groups containing texts written in varieties of the same language, namely D (Portuguese) and E (Spanish), were the most difficult to discriminate.

Compiling language variety corpora is a laborious task. Common language resources used in NLP such as Wikipedia are mostly untagged regarding the country of origin of texts and texts can be edited by both native and non-native speakers, making them unsuitable for this task. It is important to point out that the DSL shared task organizers did not make any distinction about the kind of data to be used in the open submissions. The NLI shared task (Tetreault et al., 2013), for example, distinguished between open-training 1 and open-training 2. The first one allowed the use of any amount or type of training data, excluding the shared task dataset, whereas the latter allowed the use of the shared task dataset combined with any other additional data.

Given the difficulties in compiling suitable corpora, only two systems (UniMelb-NLP and UMich) compiled external language resources and submitted results for the open submission track. Accuracy results obtained by these two teams are presented in Table 5.5.

Team	**Accuracy**
UniMelb-NLP	0.880
UMich	0.859

Table 5.5: DSL Shared Task 2014: Open Submission Results (Zampieri et al., 2014)

These two results indicate that using external data sources did not increase performance for any of the two groups who participated in the open submission track. An interesting outcome is that the best submission from UniMelb-NLP was outperformed by the best UMich system by about 1.5% accuracy in the closed submission, but in the open submission UniMelb-NLP scored 2.1% better than UMich. As discussed in the shared task report (Zampieri et al., 2014), in my opinion this difference can be explained by investigating first the quality and quantity of the external training material these teams use, and second by looking at the robustness of the classification method used to deliver correct predictions across multiple datasets and domains. The influence of the domain and text types when discriminating between three English varieties was previously discussed by Lui and Cook (2013).

5.4 DSL Shared Task 2015

The 2015 edition of the DSL shared task was held at the LT4VarDial[68] workshop co-located with RANLP. On this occasion we observed a slight increase in participation,[69] receiving a total of 24 subscriptions, 10 final submissions, and 7 system description papers.

The organizers released v2.0. and v2.1. of the DSL corpus collection and offered the same two types of submission as the 2014 edition. Participants were allowed to use the previous version of the DSLCC for the open submissions. Teams could submit up to three runs to each submission track. To summarize the types of submission:

- **Closed submission:** Using only the DSLCC v2.0 for training.

- **Open submission:** Using any dataset including the DSLCC v2.0 for training.[70]

Two changes were introduced in the DSL shared task 2015. The first was to split the test set into test set A and B, and the second was the unshared task track which was not part of the 2014 edition. Regarding the division of the test set, we adopted the following criteria:

- **Test Set A:** Includes original texts retrieved from newspapers.

- **Test Set B:** Includes texts retrieved from newspapers with most named entities removed. Capitalized named entities were substituted by placeholders.

We substituted most named entities with placeholders to decrease topic bias in classification and to evaluate the extent to which proper nouns can influence classifiers'

[68] http://ttg.uni-saarland.de/lt4vardial2015/index.html

[69] The increase in participation is in my opinion very significant due to fact that in 2014 the DSL shared task was organized within the scope of a workshop co-located with COLING, one of the largest and the most well-established conference in computational linguistics. In 2015 the workshop was co-located with RANLP, which is a much smaller conference that exists since 2003. This confirms the interest of the community in the task.

[70] Training on DSLCC v1.0 also makes a submission open.

performance.[71] It is important to note that named entity recognition (NER) systems are language specific and vary substantially both in terms of performance and in terms of the type of named entities considered. Therefore, to take advantage of existing NER tools we would have to adapt 13 different named entity taggers (one for each language or language variety) to recognize the same type of named entities, to attribute the same set of tags and to achieve the highest possible performance, and finally evaluate each of them before the release of the test set. For pragmatic reasons, the DSL approach to NE removal is rather simple. We wrote a script to substitute all words that were capitalized but didn't occur in the beginning of sentences for placeholders *#NE#*. The results of a small evaluation on the substituted corpus indicates that this approach addressed around 85% of all named entities in the texts. This matter will be further discussed in Section 5.6.

As an example of the two different test instances, I show a Spanish short text first in its original version (as an example of the ones included in test set A), and then in its version with named entities substituted by placeholders *#NE#* (as an example of the instances included in test set B).

(5) La cinta, que hoy se estrena en nuestro pas, competir contra Hors la Loi, de Argelia, Dogtooth, de Grecia, Incendies, de Canad, Life above all, de Sudfrica, y con la ganadora del Globo de Oro, In A Better World, de Dinamarca.

(6) La cinta, que hoy se estrena en nuestro pas, competir contra #NE# la #NE#, de #NE#, #NE#, de #NE#, #NE#, de #NE# , #NE# above all, de #NE#, y con la ganadora del #NE# de #NE#, #NE# A #NE# #NE#, de #NE#.

Along with the shared task, in the DSL 2015 we proposed an unshared task track as well. This track was inspired by the unshared task in PoliInformatics held in 2014 (Smith et al., 2014). For this track, teams were allowed to use any version of the DSL corpus collection to investigate differences between similar languages and language varieties using NLP methods. This track aimed to be a qualitative linguistics analysis of the dataset based on a few pre-defined questions. The questions we posed to the participants were:

- Are there fundamental grammatical differences in a language group? What are they?
- What are the most distinctive lexical choices for each language?
- Which text representation is most suitable to investigate language variation?
- What is the impact of lexical and grammatical variation on NLP applications?

[71] We applied a Python script using regular expressions for this purpose.

We received positive feedback on the unshared task and we were enthusiastic to see how teams would approach this task. From the 24 teams that enrolled for the shared task, eleven of them subscribed for the unshared task track as well. However, none of these teams ended up completing the track and submitting a paper for it.

5.4.1 Participating teams

The participants of the 2015 DSL challenge used different classifiers and features taking advantage of the experience of the 2014 edition. Below is a short description of the approaches proposed by the eight teams who submitted system description papers.

- The BOBICEV (Bobicev, 2015) team applied a technique called prediction by partial matching (PPM), which to the best of my knowledge had not been used for this task before. According to the description provided by BOBICEV, PPM is based on conditional probabilities of the upcoming character given one or more previous characters.

- The BRUniBP team (Ács et al., 2015) also approached the task using a two-stage classifier. At the first stage (language groups) the method uses a set of 100,000 keywords as features. The second stage (languages and language varieties) uses character *n*-grams, word *n*-grams, TF-IDF score, and stopwords as features. The BRUniBP team compared the performance of the MaxEnt and SVM implementations available at scikit-learn (Pedregosa et al., 2011) and opted for using MaxEnt due to processing speed.

- The MAC team (Malmasi and Dras, 2015b) used an ensemble of liner SVM classifiers. As features, the MAC team used character *n*-grams (up to 6-grams), word unigrams, and word bigrams.

- The MMS team (Zampieri et al., 2015a) is my submission with colleagues to the DSL 2015 challenge. All results from the MMS team are marked with * in the official results due to the fact that I am one of the shared task organizers.[72] I took the opportunity to compare three approaches and I submitted each of them as a different run. More details are provided in Section 5.5. The best run obtained by MMS combined TF-IDF weighting and an SVM classifier, which was previously applied to native language identification (Gebre et al., 2013).

- The NLEL team (Fabra-Boluda et al., 2015) used a Naive Bayes classifier trained as a two-stage classifier in the open submission and a single multi-class classifier in the closed submission.

[72] I acknowledge that the MMS competed under the same conditions as the other teams.

- The NRC team (Goutte and Léger, 2015) included members of the NRC-CNRC team, which developed the best system in the 2014 DSL closed submission track. Following the successful approach proposed in the first DSL challenge, in 2015 they also used two-stage classification by training a first classifier to predict the language group, and then individual classifiers per language group to predict the languages or languages varieties within each group.

- The PRHLT team (Franco-Salvador et al., 2015) also proposed a two-stage approach for the task. The novelty of their system is the use of word and sentence vectors which is a relatively recent trend in NLP (Mikolov et al., 2013). The submission of the PRHLT team is to the best of my knowledge the first attempt to apply these methods to discriminating between similar languages.

- The SUKI team (Jauhiainen et al., 2015a) used token-based backoff, which was previously applied to general-purpose language identification by Jauhiainen et al. (2015b). The features used were different types of token representation (e.g. space-delimited tokens and lowercased tokens), and character *n*-grams (from 1 to 8).

5.4.2 Results

The results obtained by the nine teams who submitted their runs to the closed submission track on test set A are shown in Table 5.6.

Team	**Accuracy**
MAC	95.54
MMS*	95.24
NRC	95.24
SUKI	94.67
BOBICEV	94.14
BRUNIBP	93.66
PRHLT	92.74
INRIA	83.91
NLEL	64.04

Table 5.6: DSL Shared Task 2015: Closed Submission Results for Test Set A in a 14 Class Classification Setting (Zampieri et al., 2015b)

The best result was obtained by the MAC team scoring 95.54% accuracy, followed very closely by MMS and NRC, both having achieved 95.24% accuracy. Seven out of the nine teams who took part in the open submission submitted runs for test set B. The results are shown in Table 5.7.

We can observe an expected drop in accuracy with respect to the results obtained on test set A. Once again, the MAC team performed best achieving 94.01% accuracy, followed by SUKI and NRC with 93.02% and 93.01%, respectively.

Team	**Accuracy**
MAC	94.01
SUKI	93.02
NRC	93.01
MMS*	92.78
BOBICEV	92.22
PRHLT	90.80
NLEL	62.78

Table 5.7: DSL Shared Task 2015: Closed Submission Results for Test Set B in a 14 Class Classification Setting (Zampieri et al., 2015b)

Next I present the results of the open submission track. Only three teams participated in this track: NRC, NLEL, and OSEVAL. Their results are presented in Table 5.8.

An important aspect to notice is that unlike in the DSL shared task in 2014, when all open submission results were substantially lower than closed submission ones, two out of the three teams who participated in the open submission, NRC and NLEL, achieved better accuracy in the open submission than in the closed one on test set A.[73] This increase in performance is in my opinion related to the domain of the texts which impact classification performance as previously discussed in this thesis. The availability of the DSLCC v1.0, which is a training corpus of the same kind as the DSLCC v2.0, has provided teams with more comparable training material for the task. In contrast, teams that participated in the open submissions track in the 2014 edition did not have access to such an adequate resource and they had to compile their own additional training data from other sources. These sources were not of the same kind as the DSLCC which led to a decrease in performance.

Team	**Accuracy**
NRC	95.65
NLEL	91.84
OSEVAL	76.17

Table 5.8: DSL Shared Task 2015: Open Submission Results for Test Set A in a 14 Class Classification Setting (Zampieri et al., 2015b)

[73] OSEVAL did not participate in the closed submission.

Table 5.9 presents the open submission results for test set B. Once again we observed improved performance for the two teams who used previous versions of the DSLCC, namely NLEL and NRC.[74]

Team	Accuracy
NRC	93.41
NLEL	89.56
OSEVAL	75.30

Table 5.9: DSL Shared Task 2015: Open Submission Results for Test Set B in a 14 Class Classification Setting (Zampieri et al., 2015b)

5.5 Comparing Likelihood Estimation

In the DSL shared task 2015, each team was allowed to submit a maximum of three runs to each track (closed and open). This turned the DSL challenge into a very good opportunity to compare the performance of different approaches including the method presented in Chapter 3, Likelihood Estimation. To accomplish this, three colleagues and myself registered to participate in the competition as the MMS team. A detailed description of our findings and our systems is published in Zampieri et al. (2015a). Here I reproduce some of the information and results included in this paper.

We developed three systems based on my previous work with colleagues on language variety identification and related tasks. For the first two runs, *Run 1* and *Run 2*, we adapted systems that were previously proposed for Native Language Identification (NLI) (Gebre et al., 2013) by the Cologne-Nijmegen team within the scope of the (NLI) shared task 2013 (Tetreault et al., 2013). Cologne-Nijmegen obtained good results in the NLI challenge and was ranked eighth in the official evaluation, and first in 10-fold cross validation.

Both systems developed for *Run 1* and *Run 2* rely on the TF-IDF weighting scheme combined with two classifiers. For *Run 1*, we used an implementation of Logistic Regression available through the LIBLINEAR open source library (Fan et al., 2008) from scikit-learn (Pedregosa et al., 2011), fixing the regularisation parameter to 100.0. This regression algorithm has been used in different classification problems including, for example, my previous work on temporal text classification (Niculae et al., 2014). For *Run 2*, we used an SVM classifier (Joachims, 1998), which was successful in the 2014 edition of the DSL task (Goutte et al., 2014) and delivered a

[74] It is important to note that in their system description paper, the NLEL team reported having a bug, which is probably the explanation for the low performance of their closed submission runs.

slightly better performance than Logistic Regression in the NLI shared task. Finally, for *Run 3* we used Likelihood Estimation.

The approaches submitted in our three runs were modelled as a single-classifier, unlike the successful two-stage approach proposed by Goutte et al. (2014) which achieved the best performance in the 2014 edition of the DSL. The three approaches submitted by the MMS team can be summarized as follows:

- **Run 1** - Logistic Regression with TF-IDF Weighting
- **Run 2** - SVM with TF-IDF Weighting
- **Run 3** - Likelihood Estimation

Along with the three different algorithms, it is important to note that for each run we used different groups of features, all of them based on n-gram language models. For *Run 1* and *Run 2* we used character n-grams ranging from bigrams to 7-grams whereas for *Run 3* we only used 5-grams, which were the best performing features using Likelihood Estimation and the DSLCC.

Table 5.10 reports the official shared task results of the MMS team in terms of accuracy, highlighting the best results for each dataset.

Run	**Test Set A**	**Test Set B**
Run 1	94.09%	**92.77%**
Run 2	**95.24%**	**92.77%**
Run 3	94.07%	92.47%
Rank	2nd *out of 9*	4th *out of 7*

Table 5.10: DSL Shared Task 2015: MMS Team Overall Accuracy in a 14 Class Classification System (Zampieri et al., 2015a)

Results obtained by the three MMS systems are all very similar. Nevertheless, the SVM with TF-IDF weighting approach proposed for *Run 2* obtained slightly better overall performance than the other two approaches. As discussed in the previous section, our results follow the trend observed in all teams' submissions, which is the performance drop from test set A to test set B. This means that our systems, even though it uses character-based representations, also relies to a certain extent on named entities to discriminate between similar languages.[75]

Tables 5.11, 5.12 and 5.13 present the confusion matrices obtained by the three systems using the 2,000 gold test instances.

[75] I acknowledge that team MMS did not carry out any specific training with the blinded named entities. This would possibly increase the performance of our systems for test set B.

	bg	**bs**	**cz**	**es-AR**	**es-ES**	**hr**	**id**	**mk**	**my**	**pt-BR**	**pt-PT**	**sk**	**sr**	**xx**
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1578	0	0	0	241	0	0	0	0	0	0	181	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1774	226	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	227	1773	0	0	0	0	0	0	0	0	0
hr	0	132	0	0	0	1841	0	0	0	0	0	0	26	1
id	0	0	0	0	0	0	1979	0	21	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	30	0	1970	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1826	174	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	222	1778	0	0	0
sk	0	0	0	0	0	0	0	0	0	0	0	2000	0	0
sr	0	86	0	0	0	41	0	0	0	0	0	0	1873	0
xx	0	0	0	0	0	0	0	0	0	0	0	0	0	2000

Table 5.11: DSL Shared Task 2015: Confusion Matrix *Run 1* - Axis Y represents the actual classes and Axis X the predicted classes (Zampieri et al., 2015a)

	bg	**bs**	**cz**	**es-AR**	**es-ES**	**hr**	**id**	**mk**	**my**	**pt-BR**	**pt-PT**	**sk**	**sr**	**xx**
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1661	0	0	0	193	0	0	0	0	0	0	146	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1796	204	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	209	1791	0	0	0	0	0	0	0	0	0
hr	0	135	0	0	0	1843	0	0	0	0	0	0	21	1
id	0	0	0	0	0	0	1988	0	12	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	19	0	1981	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1844	156	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	166	1834	0	0	0
sk	0	0	1	0	0	0	0	0	0	0	0	1999	0	0
sr	0	86	0	0	0	41	0	0	0	0	0	0	1891	0
xx	0	0	0	0	0	0	0	0	0	0	0	0	0	2000

Table 5.12: DSL Shared Task 2015: Confusion Matrix *Run 2* - Axis Y represents the actual classes and Axis X the predicted classes (Zampieri et al., 2015a)

	bg	**bs**	**cz**	**es-AR**	**es-ES**	**hr**	**id**	**mk**	**my**	**pt-BR**	**pt-PT**	**sk**	**sr**	**xx**
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1623	0	0	0	198	0	0	0	0	0	0	179	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1623	377	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	88	1912	0	0	0	0	0	0	0	0	0
hr	0	205	0	0	0	1746	0	0	0	0	0	0	49	0
id	0	0	0	0	0	0	1980	0	20	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	8	0	1992	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1867	133	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	236	1764	0	0	0
sk	0	0	0	0	0	0	0	0	0	0	0	2000	0	0
sr	0	107	0	0	0	36	0	0	0	0	0	0	1857	0
xx	5	2	0	5	7	3	0	0	0	0	0	0	2	1976

Table 5.13: DSL Shared Task 2015: Confusion Matrix *Run 3* - Axis Y represents the actual classes and Axis X the predicted classes (Zampieri et al., 2015a)

Table 5.13 shows that Likelihood Estimation used for *Run 3* achieved higher scores when discriminating between language varieties than either of the two other methods. Likelihood Estimation correctly identified 1,912 Peninsular Spanish texts and 1,867 Brazilian Portuguese texts. On the other hand, it was the only method which did not obtain 100% accuracy when identifying languages from the xx group. The results suggest that Likelihood Estimation is well suited to discriminate between language varieties, which is evidenced by the good results obtained in binary classification for Portuguese varieties (Zampieri and Gebre, 2012), but it clearly does not handle unseen data very well. For example, excluding the xx group from the valuation, for test set A *Run 3* would outperform *Run 1* and be closer to *Run 2*. This performance limitation can be explained by the simplicity of the method. In my opinion, taking the results presented in Chapter 3, Chapter 4, and in this chapter into account, Likelihood Estimation is a good method for situations in which texts from all classes are represented at the training stage (no unknown languages).

The simplicity of Likelihood Estimation has direct implications for processing speed. One of the reasons for testing this method for the first time in Zampieri and Gebre (2012) was to investigate how a very simple method would perform compared to other machine learning methods. In Table 5.14 I take a closer look at processing speed comparing the three runs submitted by team MMS. Processing speed was calculated based on the performance of an Intel i7 processor PC with 8GB RAM.

Run 1	Training Time	1,698.955s
	Test Time	1.590s
	Feature extraction	615.231s
	Number of Samples	252,000
	Number of Features	3,625,320
Run 2	Training Time	235.888s
	Test Time	1.397s
	Feature Extraction	628.954s
	Number of Samples	252,000
	Number of Features	3,625,320
Run 3	Training Time	3.258s
	Test Time	0.424s
	Feature Extraction	78.353s
	Number of Samples	252,000
	Number of Features	1,933,531

Table 5.14: DSL Shared Task 2015: Features, Instances, and Processing Speed for MMS Team Submissions (Zampieri et al., 2015a)

Likelihood Estimation is significantly faster than the other two methods in the three stages: feature extraction, training and testing. The greatest difference can be seen in the training stage, even though the number of features used by Likelihood

Estimation is about 60% of the features used in the other two methods. Likelihood Estimation takes just over 3 seconds to train the model whereas the second fastest method takes almost 4 minutes. Runtime performance is desirable in the case of very large datasets or batch processing systems in which the system needs to identify the language of a set of N documents on the fly.

5.6 On the Influence of Named Entities

To close this chapter, I investigate the influence of named entities in identifying languages and language varieties using the DSL corpus collection. This investigation made use of the Naive Bayes algorithm along with the same features and methods that were presented in the DSLCC description paper (Tan et al., 2014). I used the data available from the DSLCC v1.0. The first run uses all word forms and in the second run named entities are substituted for place holders in both training and testing. Results are reported in Table 5.15 in terms of accuracy.

Class	**With NE**	**Without NE**
bs	0.917	0.925
hr	0.944	0.944
sr	0.955	0.954
id	0.993	0.992
my	0.995	0.981
cz	1.000	1.000
sk	1.000	1.000
pt-BR	0.944	0.933
pt-PT	0.934	0.887
es-AR	0.941	0.933
es-ES	0.755	0.721

Table 5.15: DSLCC v1.0. Baseline Results with and without Named Entities

It is important to point out that, as mentioned in the shared task report (Zampieri et al., 2015b), the named entity substitution method applied to the DSLCC only addresses capitalized named entities. This was done for pragmatic reasons as not all language represented in the dataset possess equally good named entity recognition (NER) software. On a small evaluation taking Portuguese and Spanish into account, we observed that the method is able to substitute over 80% of the NE in text as the vast majority of them are capitalized (e.g. names of cities and countries, people, and organizations).[76]

[76] The method works reasonably well for the languages of the DSLCC but it would not work for

The performance clearly decreases for the vast majority of classes. The only surprise here is Bosnian which presents a slight increase in performance in the second run. The baseline algorithm continues distinguishing Czech and Slovak texts with 100% accuracy which once again provides an indication that, in their written forms, these two languages present substantial differences that allow classifiers to distinguish them with perfect performance even without relying on named entities. The language varieties are, as expected, the most difficult languages to be distinguished and the performance for Spanish and Portuguese dropped substantially.

Figures 5.1 and 5.2 present accuracy results for the 2015 test sets A and B in the closed submission track respectively. I plotted for each language the mean accuracy across all submissions and the interquartile range, excluding outliers.

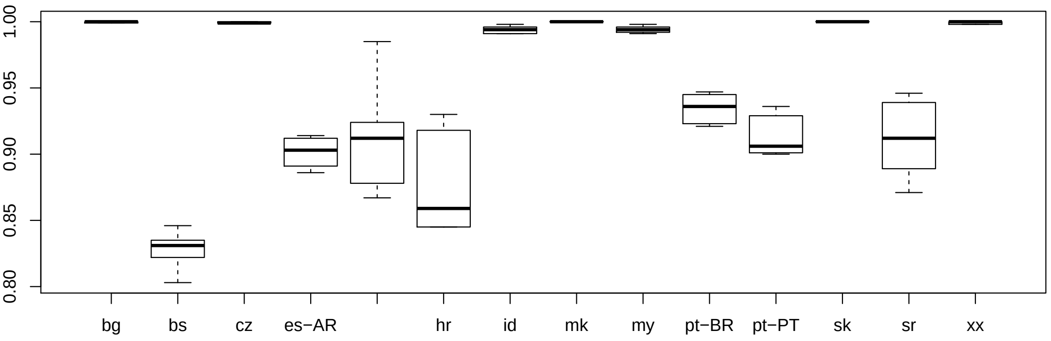

Figure 5.1: DSL Shared Task 2015: Accuracy per language: Closed Submission, Test Set A (Zampieri et al., 2015b)

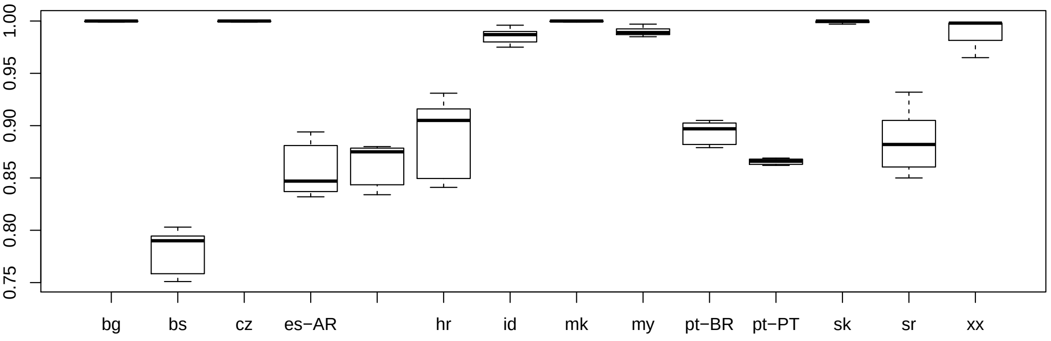

Figure 5.2: DSL Shared Task 2015: Accuracy per language: Closed Submission, Test Set B (Zampieri et al., 2015b)

languages which have different capitalization rules (e.g. German in which all nouns are written starting with a capital letter.)

Not all language pairs and groups of languages are equally difficult to distinguish from the rest. In the DSL shared task 2015 we were interested in investigating the impact of named entities in this task. Thus we divided the test set into test set A and test set B as previously mentioned in this chapter.

In the box plots we observe how performance of all teams vary between test set A and test set B. One example of this is the performance of the xx class which contains multiple languages. For this class, almost all teams obtained perfect results using test set A, but performance drops for test set B. The accuracy results obtained by the lowest performing system on the xx class falls to 95%. Other clear examples of performance drop that can be seen in the box plots are Brazilian Portuguese Indonesian, and Peninsular Spanish.

5.7 Chapter Summary

This chapter discussed the methods, datasets, and results obtained in the two editions of the DSL shared task. Organizing and promoting a successful shared task represents a lot of work for the organizers, from the compilation of the dataset to the evaluation and description of the findings. Nevertheless, in my opinion, shared tasks are a very interesting way comparing algorithms, computational methods, and features using the same dataset and evaluation methods.

Since there existed no shared task dealing with the problem of discriminating between similar languages and varieties, I believe that the DSL shared task filled an important gap in language identification and particularly in the discrimination between similar languages. This contribution will bring more attention to language identification and will allow other researchers to look in more detail into successful approaches for this task.

As previously discussed in this disseration, accurate methods for discriminating similar languages and language varieties can help to improve performance not only in language identification but also in a number of NLP tasks and applications such as part-of-speech tagging, spell checking, and machine translation. The best system in the 2014 edition of the DSL shared task obtained performance above the baseline described in Tan et al. (2014) and achieved performance of 95.71% accuracy for a set of 11 languages and varieties divided into 5 groups (A to E), using solely the DSL corpus collection (closed submission track). The best performance was obtained by using two-step predictions: first the language group and then the actual class, using characters and words as features.

The shared task also provided me with an interesting opportunity to compare likelihood estimation with other state-of-the-art algorithms. The results show that the performance of LE in this task is comparable to the performance of SVM classifiers and that the training time is substantially lower. The shared task also provides more evidence to answer **RQ2:** Can language varieties be integrated into real-world language identification systems? and **RQ3:** What are the most efficient features and algorithms to discriminate between language varieties? Given the results obtained by the best systems participating in the two editions of the DSL shared task I can state that for the languages available at the DSLCC it is possible to integrate language varieties in multilingual settings with high success rate. As to the most efficient algorithms for this task, the winners of both the 2014 and 2015 editions of the DSL challenge used SVM classifiers in a two-stage approach indicating that this is the best way to approach this task.

Another lesson learned from this shared task concerns the compilation of group F containing English texts. Researchers working with text, including the shared task organizers, often rely on previously annotated meta-data which sometimes may contain inaccurate information and errors. Corpus collection for this purpose should be thoroughly checked (manually if possible). This can be done by relying on human annotators and crowdsourcing which, however, unfortunately increase the costs of organizing such as event.

Chapter 6

Linguistic Variation: Looking Beyond Text Classification

6.1 Introduction

This chapter investigates differences in language varieties by analysing the most informative features in classification. The questions I aim to answer in this chapter are the following: is it possible to distinguish between, for example, Spanish language varieties solely based on grammatical patterns? Are there strong structural differences that allow algorithms to distinguish them automatically beyond the graphical representation of words? What are the most prominent structures that allow algorithms to differentiate, for example, Argentine Spanish from Peninsular Spanish texts? And finally, returning to **RQ4:** Can we use the information obtained from automatic classifiers to study differences between language varieties?

In this chapter I propose the use of features that go beyond word-based-methods. Inspired by the computational linguistics literature, in this thesis, I use the terms knowledge-rich and knowledge-poor to distinguish between two groups of features. Knowledge-rich refers to features which use implicit linguistic information such as morphology and syntax, whereas knowledge-poor is used to refer to features typically modelled by words and character combinations. I present a few case studies in this chapter dealing with varieties of Portuguese and Spanish. I choose these two languages because they were the pluricentric languages I explored the most throughout my thesis.

Chapter 6 is structured as follows:

- **Section 6.2** presents experiments discriminating between Spanish varieties using part-of-speech tags and morphological information.

- **Section 6.3** discusses how the output of classifiers can be used beyond automatic classification as a corpus-driven method to study language varieties.

- **Section 6.4** explores the use of the classification output to investigate differences between Brazilian and European Portuguese.

- **Section 6.5** investigates differences between four Spanish varieties using the collected corpora and the output of the classifiers. It presents two case studies: the use of demonstrative pronouns and verbal tenses.

- **Section 6.6** concludes this chapter and discusses its findings.

6.2 POS and Morphology: Spanish Varieties

At this stage, the four Spanish corpora described in Chapter 2 are used. The four corpora contain texts from Argentina, Mexico, Spain, and Peru, and they were annotated with POS and morphosyntactic information using FreeLing[77] (Padró and Stanilovsky, 2012). FreeLing is a developer-oriented library that provides language analysis services for a number of languages including Catalan, English, French and Spanish. The functions (services) available at FreeLing can be called by other text processing applications. These functions include tokenization, splitting, multiword recognition, word sense disambiguation, named entity recognition and, of course, part-of-speech tagging, which was used to annotate this data in conjunction with the morphological analyser function.

After annotating the data, it was necessary to represent the POS and morphosynctatic information in a way that the classification algorithms could handle. For this, I decided to represent the POS and morphological information as unique compound tags that could be subsequently arranged in n-grams (e.g. bigrams and trigrams) to be used in classification. I therefore concatenated the POS tags with morphological information to be able to represent tokens by using unique tags. Examples of the concatenated tags include *N-msc-sg* for noun masculine singular and *V-inf* for verbs in their infinitive form. A snapshot of the tagset and morphological information used in the annotation is presented in Table 6.1.

Sentences were therefore represented using only POS tags and morphological information disregarding word. An example of a sentence using the aforementioned text representation is displayed next.

(7) DET-msc-sg N-msc-sg PREP N-sg-prop CC PRON-poss-tonic-utr-pl-p3 NUM-card N-msc-pl DET-msc-pl N-msc-pl PREP DET-msc-pl N-msc-pl V-ind-pret-pl-p3 ADV V-inf PREP PRON-msc-sg N-msc-sg PREP PRON-poss-tonic-utr-pl-p3 N-msc-pl PREP N-msc-sg SENT

[77] http://nlp.lsi.upc.edu/freeling/

POS	Morph. Inf.	Example
N	msc sg	coche
N	msc pl	coches
N	fem sg	silla
N	fem pl	sillas
A	msc sg	bonito
A	msc pl	bonitos
A	fem sg	bonita
A	fem pl	bonitas
V	ind pres sg p1	hago
V	inf	hacer

Table 6.1: A Snapshot of the Spanish Tagset and Morphological Information (Zampieri et al., 2013)

Next I present the results obtained using POS distribution and morphological features. For this step, to provide sufficient amount of training data to LE, I used 1,000 documents from each variety, divided into 500 documents for training and 500 for testing. Accuracy results for all binary classification settings using the knowledge-rich features are presented in Table 6.2.

Feature	ARGxMEX	ARGxPER	MEXxPER	ESPxARG	ESPxMEX	ESPxPER	Average
PoS 2-grams	0.766	0.650	0.742	0.637	0.831	0.702	0.721
PoS 3-grams	0.815	0.670	0.753	0.673	0.821	0.741	0.746
PoS 4-grams	0.823	0.732	0.737	0.690	0.806	0.667	0.743
Average	0.801	0.684	0.744	0.666	0.819	0.703	0.736

Table 6.2: Accuracy Results of Spanish Classification with POS Tags and Morphological Information

The classification between Mexican and Peninsular Spanish texts obtained the best results in this setting, reaching 0.831 accuracy using compound tags (POS + morphology). These two varieties obtained the second best score for character and word-based features, which suggests that these two varieties have significantly distinctive features both at the word and at the grammatical level. The poorest results were obtained in the classification of Spanish and Argentinian texts. Argentinian and Spanish texts also had the worst performance using knowledge-poor features (see Section 4.3).

Although the results are significantly lower than those obtained with knowledge-poor features, the algorithm scored substantially better than the random 0.50 baseline. This indicates that the algorithm is able to identify patterns in the datasets using only sets of POS tags and morphological information. Moreover, named entities which usually help algorithms to identify varieties at the lexical level are not present in the experiments using POS tags, therefore not influencing the performance of the classifier. A careful linguistic analysis should be undertaken in order to understand what is behind these results in terms of differences between varieties. The most informative features of the classification can be used for this analysis and they will be investigated in more detail in this chapter.

To evaluate the relationship between the features explored here, I analysed the results with hierarchical clustering using the R language for statistical computing as presented in Zampieri et al. (2013). For each cluster, two p-values (between 0 and 1) are calculated using multiscale bootstrap resampling (Shimodaira et al., 2004). These values indicate how strongly the cluster is supported by data. The first dendogram is presented in Figure 6.1.

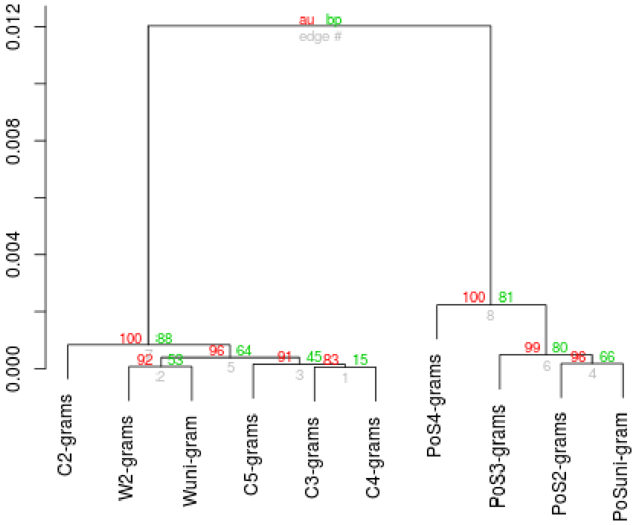

Figure 6.1: Cluster Dendogram: Spanish Classification (Zampieri et al., 2013)

The classic example to explain resampling is the world's population. Imagine that we want to calculate the average height of all members of a given population. It is impossible to measure everyone in that population, so some form of sampling to represent the entire population is required. However, from a single sample we can only calculate one mean which might lead to a biased result. What bootstrap resampling does is to introduce variability in the computation. This is done by extracting new samples from the existing data and calculating their means iteratively. The process is repeated many times and each time with a different sample.

The two p-values are: the AU (Approximately Unbiased), in red, computed by multiscale bootstrap resampling and BP (Bootstrap Probability) in green, computed by normal bootstrap resampling (see the dendograms in Figure 6.1, Figure 6.2, and Figure 6.3).

Using the aforementioned methods, I observed a direct relationship between the performance of knowledge-poor and knowledge-rich features for Spanish. Binary settings which obtained good performance using characters and words also present

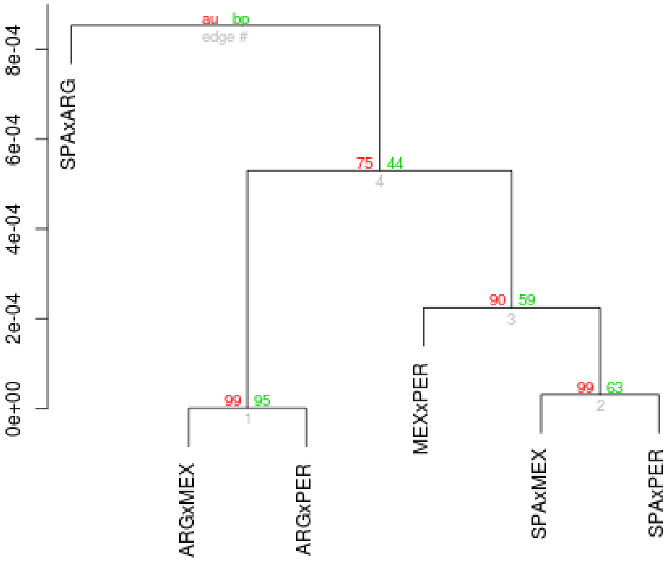

Figure 6.2: Cluster Dendogram: Knowledge-poor Results for Spanish

good results using morphosyntactic information. Figure 6.1 shows that the performance of the knowledge-poor and knowledge-rich approaches are grouped in two different branches of the dendogram, right and left respectively. This shows that the performance of the methods using these features differ significantly.

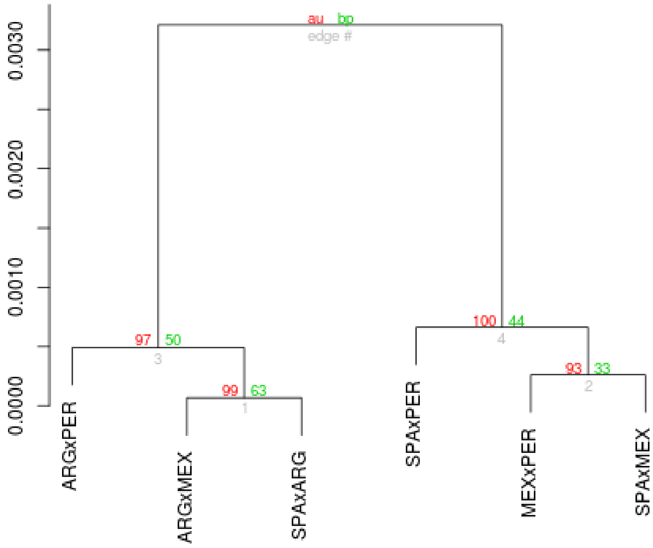

Figure 6.3: Cluster Dendogram: Knowledge-rich Features for Spanish

This relationship between results was also confirmed by the calculation of the p value (two tailed). For the values obtained in distinguishing between Peninsular Spanish and Argentine Spanish using knowledge-rich and knowledge-poor features the difference was not statistically significant which suggests there is a direct relationship between the two sets of values. As previously mentioned, this aspect should be better explored in future work through a careful linguistic analysis. The three following diagrams represent graphically the difference between the performance of knowledge-poor and knowledge-rich features.

6.3 Beyond Automatic Classification

This section deals with the information obtained as classification output and its use as a linguistic resource to study language variation. As previously discussed, this work focuses on pluricentric languages and argues that it is possible to use automatic classification to study differences between language varieties. Most work on automatic classification and language identification is concerned solely with how well the computational methods can distinguish languages and not with the differences that make this task possible.

In the next sub-sections I look in more detail into the differences between language varieties that allow algorithms to distinguish them automatically. It is, however, beyond the scope of my work to carry out a comprehensive linguistic investigation into the differences between these varieties. Nevertheless, I am convinced that the use of corpus-driven methods can provide interesting evidence for linguists to study these differences more extensively.

The two examples I focus on are Portuguese (Brazilian and European varieties) and Spanish, which contains samples from Spain, Argentina, Mexico and Peru. These two examples show that it is possible to carry out contrastive analysis by looking at the most informative features used to distinguish two languages as well as those obtained in experiments involving multiple languages.

6.3.1 Representing Language: Corpus-driven vs. Corpus-based

The classification methods proposed in this work use complete bodies of text to allow for generalizations. This makes it possible for the outcome of this study to be considered a source of information for a corpus-driven study which is substantially different from most corpus-based studies in the field of dialectology. The goal of studies in dialectology is conceptually similar to what I discuss here as such studies are also interested in studying and describing diatopical variation of languages. The main difference regarding the data used in each approach is that I use large-sized samples processed automatically instead of selected aspects of language, the sampling technique preferred by most dialectologists.

Moisl (2009) discusses two paradigms similar to the dichotomy discussed by Dipper (2008) in the scope of computational linguistics. Moisl describes the use of corpora for dialectology, in particular historical dialectology and he exemplifies this with the research question of classifying documents dialectally based on their most important linguistic characteristics. Given this example he divides the exploitation of corpora between theoretically-driven where the 'classification criteria are selected by the researcher on the basis of an independently-specified theoretical linguistic framework supported by existing case studies' and empirically-driven where 'classification criteria are algorithmically abstracted from the corpus data itself without reference to any theoretical linguistic framework'.

Accordingly, the experiments in this thesis are clearly more empirically-driven than theoretically-driven and corpus-driven rather than corpus-based. The description of an empirically-driven approach discussed by Moisl (2009) is consistent with the characterization of corpus-driven studies provided by McEnery and Hardie.

> Corpus-driven linguistics rejects the characterisation of corpus linguistics as a method and claims instead that the corpus itself should be the sole source of our hypotheses about language. It is thus claimed that the corpus itself embodies its own theory of language (Tognini-Bonelli, 2001, p.84-85). This notion of corpus-driven linguistics is closely associated with the work of scholars we will refer to as neo-Firthians (McEnery and Hardie, 2011, p.6).

Whether the approach presented here is more interesting to dialectology and/or contrastive linguistics than theoretically-driven ones, will not be fully evaluated here. What can be said is that these experiments constitute an attempt to investigate the variation between language varieties in a purely automatic way with features that are linguistically motivated and, furthermore, that these methods may also be applied to the study of dialects.

I propose the use of features that go beyond the linguistic surface. Inspired by the computational linguistics literature, I use the terms knowledge-rich and knowledge-poor to distinguish between them. Knowledge-rich refers to features which use more complex or implicit linguistic information such as morphology and syntax, whereas knowledge-poor features are used to refer to simple features typically modelled by words and character combinations. Differences in orthography are much more salient and can be easily captured by state-of-the-art classification algorithms. For example, the European Portuguese word *acto* has the same meaning as its correspondent *ato* in Brazilian Portuguese. To my understanding, this does not reflect any intrinsic systemic difference between the two varieties, but only a difference in orthography: Europeans use mute consonants *c*, *p* and *t* whereas Brazilians do not. This is from a linguistic point of view not of great interest, although from a computational linguistic perspective it does help algorithms to distinguish the two varieties automatically. By using abstract features, such as the combination of POS tags or morphological information, it is possible to investigate the extent to which the differences within language varieties lie at the structure level and are not simply orthographic conventions.

Another aspect of note is the use of named entities, which are often proper nouns that denote names of people, places, organizations, etc. It seems intuitive to think that Brazilian Portuguese texts will contain the word *Rio de Janeiro* in a much higher frequency than European Portuguese will. On the other hand the word *Lisboa* will be much more frequent in European Portuguese texts. These kind of culturally or thematically biased terms will also help algorithms to distinguish varieties automatically without any need to look at deeper or more intrinsic linguistic differences. Knowledge-rich features are also used to avoid this influence by disregarding words.

Another question that arises when looking at text samples is: taking the case of Brazilian and European Portuguese as an example, are we looking at two samples that have substantial differences in the *parole* but belong to the same linguistic system? Or are the differences so strong as to allow linguists to state they are two different linguistic systems genetically related? It is important to say that it would be beyond the scope of this work to be conclusive regarding the status of language varieties, dialects or languages in their own right. This task would be a much more complicated endeavour, and one which this thesis does not aim to address.

In the next sections I look more closely into the linguistic differences between language varieties yielded by the classification methods presented earlier.

6.4 Differences Between Portuguese Varieties

Articles that deal with Portuguese as a pluricentric language, divide Portuguese language varieties into three basic subgroups: The European variety, the Brazilian variety, and an African variety (Baxter, 1992). For my experiments I collected samples from Portugal and Brazil, on the basis that the Brazilian and the European varieties are the most representative because they are spoken by the whole population of their respective countries.[78] In African countries such as Angola and Mozambique, Portuguese co-exists with a number of national languages such as Kikongo and Umbundu in Angola and Swahili and Tsonga in Mozambique. Therefore, only a part of their population has Portuguese as their mother-tongue.

There are substantial differences between European and Brazilian Portuguese in terms of phonetics, syntax, lexicon and orthography. For the analysis of written texts, differences in syntax, lexicon and orthography are obviously more important. Orthography in these language varieties differs in two main aspects: graphical signs and mute consonants. Due to phonetic differences, Brazilian and European Portuguese use different orthographical signs for the same word, such as:

[78] Indigenous languages in the north of Brazil as well as Mirandese in the north of Portugal are the only exceptions. However, these have only a small number of speakers.

- *econômico* (BP); *económico* (EP): *economic* (EN)

Mute consonants are still used in the Portuguese orthography and are no longer used in Brazil:

- *ator* (BP); *actor* (EP): *actor* (EN)

Differences also appear at the syntactic level. Some contractions are used only in one of the varieties; for instance: *mo* (pronoun *me* + definite masculine article *o*) is exclusive to Portugal. Past verb tenses (perfect and imperfect) are used in different contexts, in each of them specific to the language variety. The use of pronouns also differs: the Brazilian variety tends to prefer the pronoun before the verb whereas the European variety uses it primarily afterwards:

- *eu te amo* (BP) and *eu amo-te* (EP): *I love you* (EN)
- *eu me chamo* (BP) and *eu chamo-mo* (EP): *I am called* or *My name is* (EN)

Lexical variation is also a distinctive characteristic of these varieties. Some words are frequent in one of the varieties and rare in the other: *nomeadamente* (EP), *namely* (EN) is widely used in Portugal and rare in Brazil. Additionally, there are cases in which each variety may heavily favour a different word in a set of synonyms, such as: *coima*(EP), *multa* (BP), *fine, penalty* (EN).

At this stage, it is important to mention the 1990 Orthographic Agreement, an international treaty that aims is to create an unified orthography for the Portuguese language in all Portuguese speaking countries. At the present moment, the orthographic agreement is an ongoing process for all members of the Community of Portuguese Speaking countries (CPLP),[79] or countries that have Portuguese as their official language, that should be concluded within a few years. The main point of the orthographic agreement is to improve the international status of the language. The agreement aims to substitute two official orthographic norms which are already in used: the Brazilian norm and the norm used in the remaining Portuguese-speaking countries.

Next I look more closely into the most informative features of the automatic classification of two Portuguese varieties, Brazil and Portugal, using lexical features. These lexical features were arranged as bag-of-words (BoW) and were classified using three algorithms: Multinomial Naive Bayes (MNB), J48 and Support Vector Machines (SVM). The output of the MNB classifier provided a set of 1,269 lexical items[80] each one of them having two scores. These two scores represent the individual probability of each word belonging to one of the two classes. WEKA's output

[79] CPLP from the Portuguese: *Comunidade dos Países de Língua Portuguesa.*

[80] At this stage I excluded 57 entries that were numbers, hence not relevant to a linguistic analysis.

is displayed in alphabetical order without any correlation between the two scores. To analyse the output of the classification experiments some ranking was needed. For this stage, I calculated the difference between the individual probabilities of each word belonging to the two different classes, BR for Brazilian and PT for Portugal using the following equation:

$$Difference = (w_i|BR) - (w_i|PT) \tag{6.1}$$

This step produced an ordered list containing 1,212 lexemes. The complete table is in annex A. In the following table, I include a snapshot containing the top 60 lexical items for the Brazilian class.

Table 6.3: Prominent Lexical Items in PT x BR Classification (Brazilian Axis - Short)

Word	**Brazil**	**Portugal**	**Difference**
de	0,01024248	0,0073774249	0,0028650551
R	0,0025541831	3,64E-005	0,0025178052
São	0,0031695558	7,13E-004	0,0024565502
e	0,0096901894	0,0073725745	0,0023176149
governo	0,0025155613	2,16E-004	0,0022997195
Brasil	0,00256062	2,79E-004	0,0022817233
Paulo	0,0029378255	6,98E-004	0,002239371
a	0,0095962099	0,0073749997	0,0022212102
o	0,0094983682	0,0073725745	0,0021257937
do	0,0093516057	0,0073289211	0,0020226846
ele	0,0028296847	8,51E-004	0,0019784433
LOCAL	0,0019568338	2,43E-006	0,0019544086
REPORTAGEM	0,0019568338	2,43E-006	0,0019544086
que	0,0091507728	0,0073652989	0,0017854738
O	0,0080436168	0,0063079165	0,0017357003
da	0,0090117346	0,0073119448	0,0016997898
em	0,0087993151	0,0072391891	0,001560126
DE	0,0015410066	4,85E-006	0,0015361562
Folha	0,0014251415	1,46E-005	0,0014105903
Ele	0,0015577427	1,77E-004	0,0013807039
FOLHA	0,0013504728	2,43E-006	0,0013480476
país	0,0023237401	0,0011155869	0,0012081532
SUCURSAL	0,0011676633	2,43E-006	0,0011652381
Luiz	0,0011818246	4,61E-005	0,001135746
no	0,0080397546	0,0069263397	0,0011134149

Table 6.3 – *Continued from previous page*

Word	**Brazil**	**Portugal**	**Difference**
Rio	0,0012500563	1,50E-004	0,0010996946
Lula	0,0011110181	3,15E-005	0,0010794907
para	0,008153045	0,0070936778	0,0010593672
DO	0,001047936	2,43E-006	0,0010455108
disse	0,0028039368	0,0017606872	0,0010432496
Segundo	0,0022490715	0,0012247204	0,001024351
janeiro	9,96E-004	4,85E-006	0,00099159
US	9,40E-004	9,70E-006	0,0009300944
com	0,0079045786	0,0069918198	0,0009127588
fato	9,15E-004	3,15E-005	0,0008838073
havia	0,0013272998	4,56E-004	0,0008713642
diretor	8,65E-004	2,43E-006	0,0008627013
Federal	8,72E-004	1,94E-005	0,000852162
BRASÍLIA	8,54E-004	2,43E-006	0,0008511148
deve	0,0018036343	9,65E-004	0,0008384091
projeto	8,33E-004	2,43E-006	0,0008305166
atual	8,24E-004	2,43E-006	0,0008215048
na	0,0075157866	0,006698372	0,0008174147
brasileiro	9,78E-004	1,75E-004	0,0008038033
novembro	7,94E-004	2,43E-006	0,0007918948
diz	0,0025322974	0,0017437109	0,0007885865
ela	0,0012989772	5,24E-004	0,0007751363
equipe	7,75E-004	2,43E-006	0,000772584
Justiça	0,0010273377	2,59E-004	0,0007678425
SP	7,65E-004	2,43E-006	0,0007622849
ação	7,35E-004	2,43E-006	0,0007326749
REDAÇÃO	7,33E-004	2,43E-006	0,0007301001
setor	7,33E-004	2,43E-006	0,0007301001
vem	0,0011110181	3,83E-004	0,0007278383
brasileira	8,21E-004	9,46E-005	0,0007267729
com	0,001924649	0,0012125945	0,0007120545
federal	7,12E-004	2,43E-005	0,0006876751
dezembro	6,85E-004	4,85E-006	0,0006800414
idéia	6,80E-004	2,43E-006	0,0006773171
time	6,85E-004	1,46E-005	0,0006703407

As expected, named entities play an important role in identifying the Brazilian class, hence the words: *Brasil, São Paulo, Lula.* Apart from that, we may also observe the aforementioned difference in the use of the mute consonants that are not present in the Brazilian orthography including *fato, diretor, projeto, atual* and *setor*. The use of the mute consonants are very important for the classifiers to distinguish between Brazilian and European texts. Another characteristic is the presence of the graphical sign in *idéia* that is not used in Portugal. Next I present the most important lexical items for the Portuguese class.

Table 6.4: Prominent Lexical Items in PT x BR Classification (European Axis - Short)

Word	**Brazil**	**Portugal**	**Difference**
DN	1,29E-006	0,002284528	-0,0022832406
Portugal	1,18E-004	0,0023475829	-0,002229143
euros	2,70E-005	0,001847994	-0,0018209588
Lisboa	5,15E-005	0,0017146086	-0,001663113
facto	1,29E-006	0,0015278691	-0,0015265817
Governo	2,18E-004	0,0017315849	-0,0015140159
aos	0,0022001506	0,003657185	-0,0014570344
num	9,00E-004	0,0023306066	-0,0014307205
numa	8,27E-004	0,0022117724	-0,0013852676
ainda	0,0029236642	0,0041640495	-0,0012403853
já	0,0034811043	0,0046248354	-0,0011437311
E	0,0017611504	0,0028932505	-0,0011321
desta	3,09E-004	0,0013944837	-0,0010855099
Esta	1,03E-004	0,0011762167	-0,0010732254
portugueses	3,48E-005	0,0010840595	-0,0010492999
Este	1,27E-004	0,0011689411	-0,0010414894
à	0,005074894	0,0060750984	-0,0010002045
onde	0,0013993936	0,0023766852	-0,0009772916
projecto	1,29E-006	9,63E-004	-0,0009615126
português	1,71E-004	0,0011252877	-0,0009540647
através	1,27E-004	0,0010792091	-0,0009517574
ou	0,0031541071	0,0040912938	-0,0009371867
objectivo	2,57E-006	9,34E-004	-0,000931123
actual	1,29E-006	9,09E-004	-0,0009081585
tal	2,77E-004	0,0011713663	-0,0008945773
Mas	0,0020778485	0,0029684313	-0,0008905828
ver	7,18E-004	0,0015763728	-0,0008580089
forma	0,0010942821	0,0019474268	-0,0008531447
equipa	2,57E-006	8,51E-004	-0,0008486666
portuguesa	5,92E-005	8,97E-004	-0,0008381
agora	0,0012513437	0,0020856625	-0,0008343188

Table 6.4 – *Continued from previous page*

Word	**Brazil**	**Portugal**	**Difference**
altura	2,21E-004	0,0010476816	-0,0008262504
assim	9,54E-004	0,0017800887	-0,0008261322
estes	4,76E-005	8,59E-004	-0,0008108834
qualquer	7,49E-004	0,0015569713	-0,00080771
António	2,57E-006	8,00E-004	-0,0007977376
tendo	2,30E-004	0,0010113038	-0,0007808609
ideia	1,29E-006	7,76E-004	-0,0007747731
estar	6,68E-004	0,0014284363	-0,0007602806
explicou	9,66E-005	8,39E-004	-0,0007425611
sector	1,29E-006	7,37E-004	-0,0007359701
ao	0,0056992784	0,0064316012	-0,0007323228
Fevereiro	1,29E-006	7,20E-004	-0,0007189937
ter	0,0027666025	0,003477721	-0,0007111185
País	1,18E-004	8,29E-004	-0,0007109747
depois	0,0016633087	0,0023718348	-0,0007085261
isto	7,72E-005	7,86E-004	-0,0007085178
nas	0,0022168867	0,0029247779	-0,0007078912
sobretudo	2,46E-004	9,43E-004	-0,0006975069
sempre	0,0010363495	0,0017315849	-0,0006952355
lhe	3,28E-004	0,0010185794	-0,0006902948
muito	0,0021589541	0,0028447467	-0,0006857926
face	4,12E-005	7,15E-004	-0,0006742342
Isto	2,96E-005	6,91E-004	-0,0006615689
vão	6,24E-004	0,0012732242	-0,0006488397
Europeia	1,29E-006	6,48E-004	-0,0006462381
Março	2,70E-005	6,72E-004	-0,0006447421
estas	2,19E-005	6,57E-004	-0,0006353406
nomeadamente	1,29E-006	6,35E-004	-0,0006341121

The Portuguese axis also shows examples of spelling differences such as in *projecto, facto, objectivo*, and *actual*. The adverb *nomeadamente* is an interesting case of lexical preference as it exists in both varieties but it is very frequent in Portugal and not often used in Brazil.[81]

This list of words obtained in the classification is similar to the keyword lists that are often used in corpus linguistics. For keyword calculation, one corpus (ideally

[81] For additional information about the syntactic behaviour and the frequency of -mente adverbs in Brazilian and European Portuguese see Baptista et al. (2012).

bigger) should be used as a reference corpus. Words are ranked based on their degree of *keyness* calculated with associative metrics like chi-square or log-likelihood. The *keyness* value is a score of how specific this word is to the corpus of study in comparison to the reference corpus. To compare the list of words obtained by the classifiers with this standard method in corpus linguistics, I calculated the keywords in my Brazilian corpus using the Portuguese corpus as a reference and using the log-likelihood implemented in the corpus processing software AntConc (Anthony, 2005, 2013).

Table 6.5: Keywords BR

Frequency	Keyness	Word
3665	2969.452	folha
4229	2373.680	brasil
5597	2147.849	ele
2480	1852.015	lula
2601	1797.970	h
4572	1610.655	paulo
1824	1506.114	us
1976	1442.306	reportagem
1825	1384.045	federal
1430	1172.243	sp
1257	1071.368	projeto
1205	1027.048	bilhões
2371	1023.886	rio
1213	1019.781	brasília
1863	994.140	brasileiro
1142	973.351	diretor
9619	869.869	sção
1004	855.731	br
1160	834.359	você
1085	817.934	pág
1228	810.933	luiz
947	807.149	setor
1123	804.045	time
1152	792.740	www
896	763.680	ação
943	760.488	sucursal
870	728.099	equipe
787	670.777	atual
785	669.072	ações

Table 6.5 – *Continued from previous page*

Frequency	**Keyness**	**Word**
799	667.754	paulista
1166	653.645	brasileira
748	637.537	bc
863	634.178	fato
4396	620.307	disse
741	618.470	prefeito
677	577.022	econômica
676	576.169	gt
2423	568.147	local
661	563.385	idéia
660	562.532	redação
652	542.869	prefeitura
2245	536.533	ela
660	530.508	psdb
686	514.312	tel
615	511.450	pmdb
2327	508.107	eu
994	506.617	natal
827	486.130	brasileiros
830	484.168	renda
870	474.042	pesquisa
758	472.844	argentina
533	454.287	págs
528	450.026	lt
524	446.617	projetos
523	445.764	objetivo
669	436.559	dólar
1748	423.898	eles
914	422.271	juros
4899	402.476	segundo
5331	398.571	governo

After comparing the 60 most informative features for the MNB algorithm with the 60 top ranked keywords for the Brazilian class, I observed an overlap of 43 words, or 71.6%. This outcome indicates a significant overlap between what is considered by classification algorithms as discriminative features and the keywords that are obtained by corpus processing tools.

This confirms my initial hypothesis that a well-designed classification experiment with comparable samples may level out meaningful characteristics of the data and obtain similar results to those calculated by corpus processing tools.

6.5 Differences Between Spanish Varieties

Methodologically, distinguishing between more than 2 varieties of Spanish allows me to explore data differently than what could be done with the results from the binary classification experiments on Portuguese. I use the R package FactoMineR to carry out factor analysis (Lê et al., 2008) on Spanish lexical items. Figure 6.4 presents the distribution of lexical items across the four different varieties.

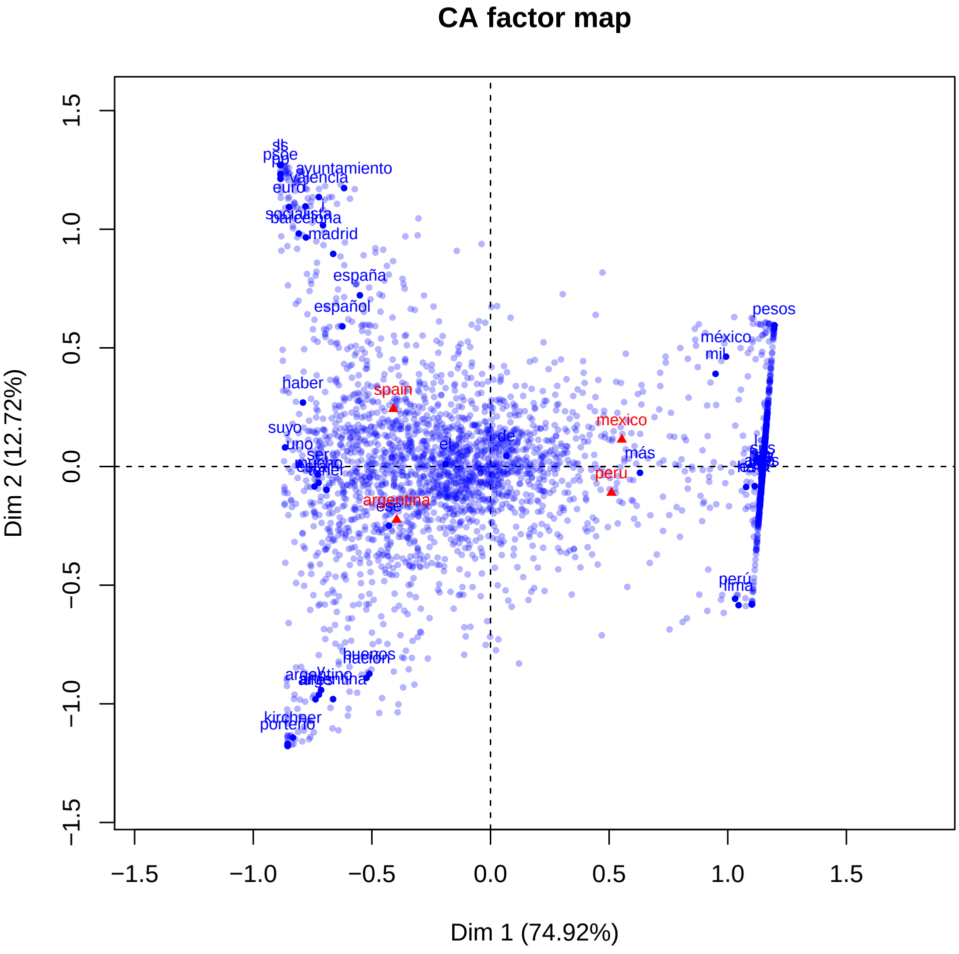

Figure 6.4: Lexical Features across four Spanish varieties

The word unigram list is available online so that these experiments can be replicated by other researchers (see Zampieri et al. 2013). The factor analysis reveals the influence of the most important lexical items in each sample (including many named entities) in classification. The presence of named entities as very informative features for classification is a known fact previously discussed in research with Portuguese and French as well as the multilingual experiments using the DSL corpus collection.

Even so, it is possible to find some interesting differences in the distribution of lexical items across the four samples that correspond to what has been discussed in the literature on the characteristics of Spanish varieties. One interesting case that will be discussed in this chapter in more detail is, for example, the distribution of the use of the demonstrative pronouns: *ese*, *este* and *aquel* in the different samples.

To diminish the influence of named entities, I subsequently use the output of the classification using POS tags and repeat the same steps described before. As with the word unigram list, the POS bigram list is also available online.[82]

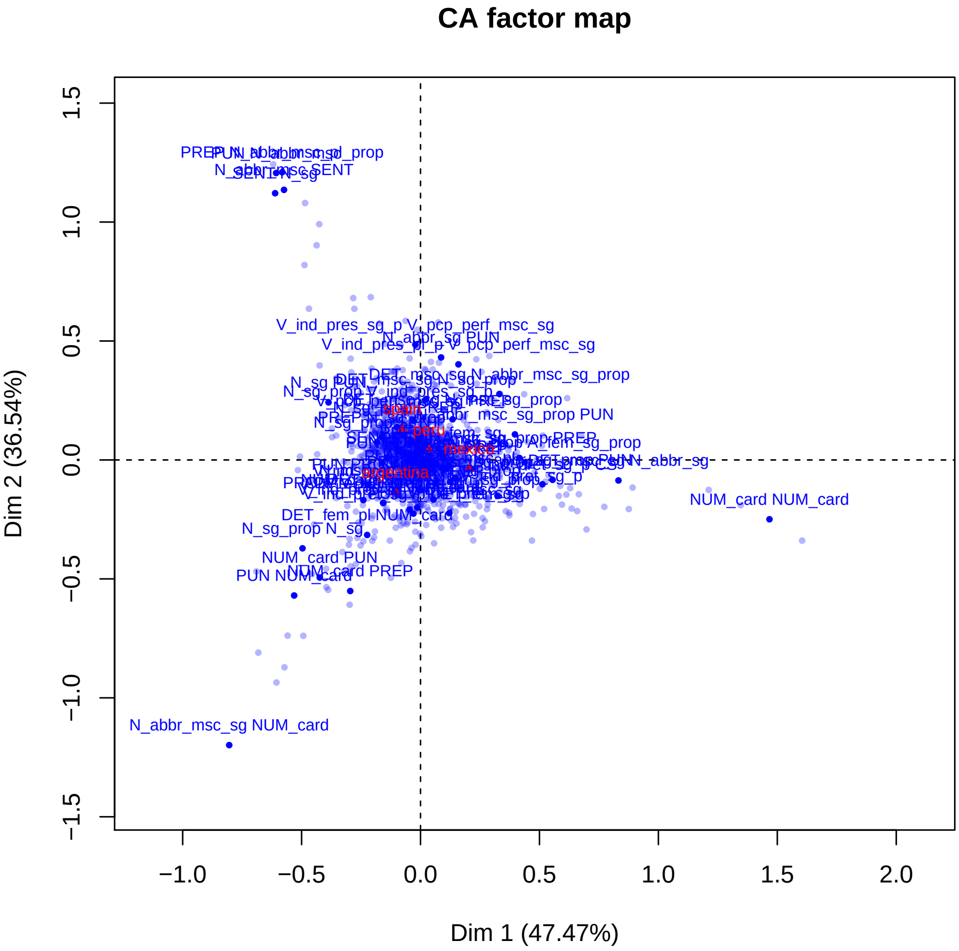

Figure 6.5: Morphosyntactic Features across four Spanish Varieties

The factor analysis presented in Figure 6.5 illustrates the distribution of the four samples, as expected, in four quadrants. In each quadrant the most prominent lexical items are shown and the more distant they are from the x and y axes the more important they are to the given variety. Not surprisingly, most of these lexical items

[82] http://www.dfki.de/~maza02/resources/spanishposbigram.txt

are named entities, which directly help in classification. Some examples include: *Madrid*, *España* and *español* in the Peninsular Spanish quadrant or *buenos* (probably from *Buenos Aires*), *Kirchner*, *proteño* and *Argentina* for the Argentinian quadrant.

The influence of named entities seems intuitive and provides a great aid for these discriminative methods that, from an NLP point of view, cannot be neglected or suppressed. As I wanted to look closer into systemic differences between these varieties, I carried out classification experiments that disregarded lexical items. The idea was to investigate whether it was possible to distinguish samples solely based on morphosyntacitc distribution. The results I presented earlier in this chapter show that the algorithm is able to discriminate varieties based solely on their POS distribution with moderate loss of performance but accuracy scores substantially above the baseline. This suggests that there are systemic differences in the corpora that the algorithm can capture. The next graph looks more closely at the distribution of morphosyntactic patterns using bigrams.

In this step I found the compound past construction with the auxiliary verb *haber* in Peninsular Spanish to be most prominent. This is evidenced by the distribution in Figure 6.6.

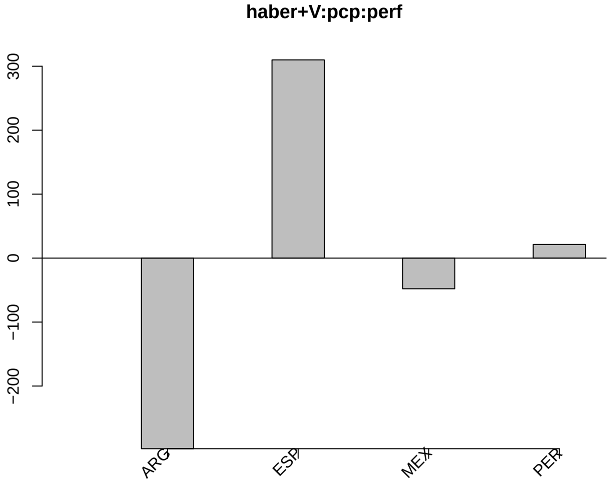

Figure 6.6: Distribution of *haber+V:pcp:perf* across four Spanish Varieties

The distribution of the construction with *haber* across the four varieties confirms the hypothesis. The value 0 represents the mean of the statistical distribution and the four bars represent the use of *haber+V:pcp:perf* in each of the varieties. It is possible to observe that the construction is mostly used in Peninsular Spanish texts, under represented in Mexico and even more so in Argentina. This is an example from the Peninsular Spanish corpus with *ha mejorado.*

(8) En Cubelles, que este año ha mejorado notablemente los niveles de calidad

de la arena y las aguas de sus playas, la Cruz Roja dispondrá por primera vez de una lancha.

Another example of the use of verbs in the past tense is the prominence of *Vpret* in the Argentinian sample, as presented in Figure 6.7. These two differences, in my opinion, represent a structural difference between Spanish varieties that would be very difficult to address using non-annotated data.

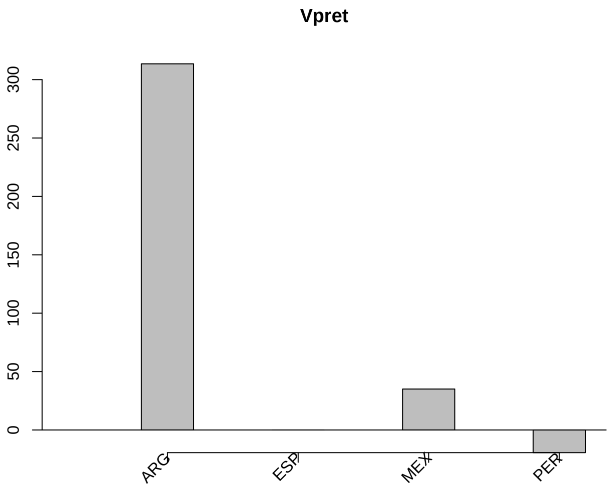

Figure 6.7: Distribution of Verbs *preterito* across four Spanish Varieties

Another interesting example of usage of the past tense from the Argentinian corpus is the word *elogió*.

(9) Oviedo se encuentra en la Argentina como parte de su gira proselitista y elogió ayer al matrimonio Kirchner: "És un binomio tan perfecto como pocas veces se dio en la historia de los gobiernos de las naciones".

The use of the past tense varies significantly across these samples with the compound form preferred in Spain while the simple past is favoured in Argentina. This confirms what was described by Squartini and Bertinetto (2000) in their article on simple past and compound past (henceforth SP and CP) in Romance languages. Squartini and Bertinetto (2000) discuss the development of these constructions in light of the work by Harris (1982). According to Harris (1982) there are four stages of development of SP and CP in Romance languages as follows:

1. CP is restricted to present states resulting from past actions;

2. CP occurs in specific circumstances, contexts marked as durative or repetitive. Comparable to the English usage *I have lived here*;

3. CP expresses 'the archetypal present perfect value of past action with present relevance';

4. CP expresses the preterial functions, while SP is restricted to 'formal registers'.

These structures and their syntactic, semantic or pragmatic behaviour were grammaticalized in each of the Romance languages at a different time and with different readings. The four stages of development described by Harris (1982) consider the languages and dialects in which these structures first appeared as follows:

Stage	Language
1	Southern Italian Dialects
2	Galician, Portuguese and a number of American Spanish Varieties
3	Castillian Spanish, Occitan
4	Standard French, Northern and Central Italian Dialects

Table 6.6: Stages of CP/SP development (Harris, 1982) (Squartini and Bertinetto, 2000)

Taking the work by Squartini and Bertinetto (2000) as a starting point it is possible to find the first difference in the grammaticalization of CP structure in Spanish which occurred with an older reading in American Spanish varieties and only later in the Castillian variety. To exhaustively investigate this with native speakers of two Spanish varieties is unfortunately impossible and outside the scope of this thesis. For the sake of brevity, I have selected a usage example similar to the one presented by Squartini and Bertinetto (2000) and Lope Blanch (1961) on the use of CP and SP in Mexican Spanish using the adverbial *ya* (already).

Examples from Argentina and Spain collected in the corpus were provided to two Spanish speakers (one from Spain and the other from Argentina). Their grammaticality judgement seems to confirm that in Argentinian Spanish the CP is used more often in durative contexts whereas in Spain the CP construction tends to be used as a past action relevant to the present.

The Argentinian Spanish native speaker, for example, judged example 10 grammatical and example 11 as non-grammatical referring to someone who no longer lives in Germany, whereas the Peninsular Spanish speaker considered both examples to be valid, yet also preferring the first one.

(10) Yo ya viví en Alemania.

(11) Yo ya he vivido en Alemania.

Example 10 would be judged grammatical (or preferred) if this person were still living in Germany when the sentence was uttered. The aforementioned 'present relevance' of these constructions can be observed by looking at the following examples. A similar example involves temporal expressions such as the case of *hoy (today)*

(12) Maradona ha jugado hoy.

(13) *Maradona jugó hoy.

For the Peninsular Spanish speaker, example 12 is judged to be grammatical whereas 13 not and therefore, in this case, the CP construction is preferred, which corresponds to the description of Squartini and Bertinetto (2000).

To confirm this hypothesis I searched the Corpus de Referencia del Español Acutal (CREA)[83] for examples of the quotative verb *declaró (to declare)* followed by the adverbs *hoy (today)* and *ayer (yesterday).*

Country	**Percentage**	**Frequency**
U.S.A	19.35	12
Bolivia	17.74	11
Mexico	16.12	10
Peru	9.67	6
Venezuela	9.67	6
Argentina	8.06	5
Guatemala	8.06	5
Spain	6.45	4
Equador	3.22	2
Cuba	1.61	1

Table 6.7: Distribution of *declaró hoy* in CREA

Country	**Percentage**	**Frequency**
Spain	85.92	916
Argentina	3.28	35
Costa Rica	1.78	19
Mexico	1.59	17
Colombia	1.40	15
Guatemala	1.21	13
Domenican Rep.	1.21	13
Honduras	1.12	12
U.S.A.	1.03	11
Other	1.40	15

Table 6.8: Distribution of *declaró ayer* in CREA

The quotative verb was chosen to avoid the durative meaning that some verbs have. The distribution is different across countries and it is possible to confirm that the Peninsular Spanish variety rejects the use of SP with the adverb *hoy* as mentioned in

[83] http://corpus.rae.es/creanet.html

example 12.[84] There are only four examples of this kind in the corpus from different cities (Barcelona, Madrid, Valladolid and Pamplona), but considering the size of the Peninsular Spanish sub-corpus we shall consider these occurrences of little relevance. This outcome seems to confirm that for past actions with present relevance the CP is preferred as in example 12 *Maradona ha jugado hoy.*

Once again, it should be noted that what I aim to accomplish in this chapter is to show how the output of classification methods relate to state-of-the-art theory in variational and contrastive linguistics. In the case of the verbs, the classification output and sampling techniques are sensitive to an important difference in the usage of verbs among varieties that seemed significant. With these results in hand I searched for the relationship between the quantitative output and the linguistic descriptions available in the literature. This is, in my opinion, the standard corpus-driven way of formulating hypotheses as discussed earlier in this thesis. As to the findings themselves, it would be beyond the scope of my work to hazard any conclusive remark on Spanish or Portuguese grammar. Nevertheless, I hope that the data and the examples discussed here are interesting for philologists and linguists.

Another interesting example that we are able to take a closer look at is the use of demonstrative pronouns across these four Spanish varieties. A set of pronouns proved to be a distinctive feature in the classification experiments, particularly *aquel, este* and *ese.* The graphs presented in Figure 6.8, Figure 6.9, and Figure 6.10 show the distribution of *aquel*, *este* and *ese* across the four Spanish samples.

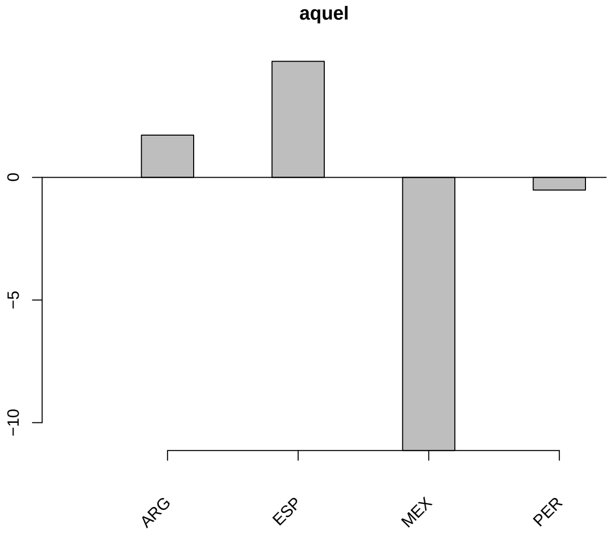

Figure 6.8: Distribution of *aquel* across four Spanish Varieties

[84] It is important to point out the frequencies in CREA are not smoothed or weighted, which means that the 11 examples for Bolivia in 6.7 represent much more than the 11 examples in American texts in Table 6.8.

The Argentinian Spanish variety, for example, prefers the use of *aquel* and *ese* whereas in the Peninsular Spanish samples the use of *aquel* and *este* is more prominent.

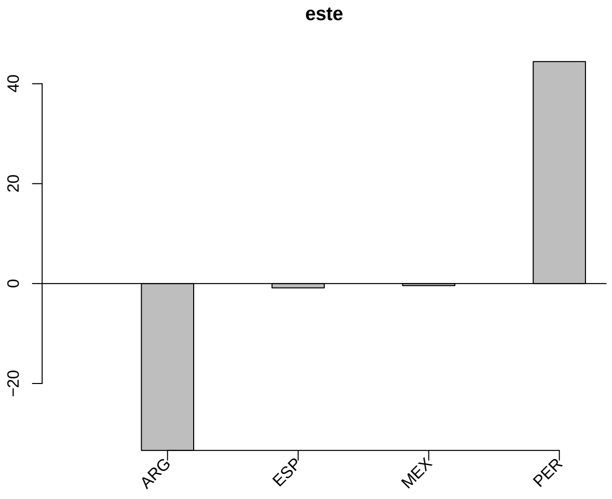

Figure 6.9: Distribution of *este* across four Spanish Varieties

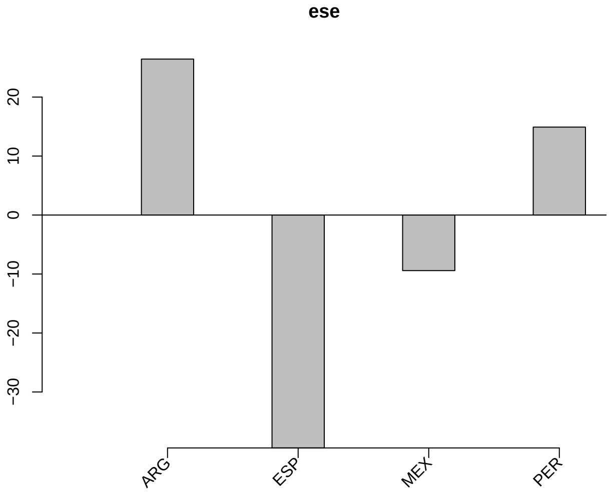

Figure 6.10: Distribution of *ese* across four Spanish Varieties

One explanation for this observation might indicate the transformation of the three-way system of demonstrative pronouns in Romance languages to a binary system similar to the English demonstrative pronouns *this* and *that.* My hypothesis by looking at these results and according to the literature presented on the following pages is that the demonstrative (or deitic) binary system will conserve the form *aquel* which is equivalent to *that* and consolidate either the use of *ese* or *este* comparable to the English form *this* depending on each variety.

Zulaica Hernández (2007) investigates the anaphoric and discourse properties of the three pronouns in Peninsular Spanish (both written and spoken). This study uses the spoken and written parts of the Spanish section of the CREA corpus. Unfortunately it does not provide any further indication on the use of these pronouns in other Spanish varieties. Nevertheless, the author concludes that for this variety the 'demonstrative pronoun *aquello* still maintains a distance - proximity feature, while the *esto* and *eso* are distance undefined or unspecified elements'.

The work of Zulaica Hernández (2007) also examines the aforementioned transformation of Spanish demonstrative pronouns from a tripartite to a binary system. His results also confirm the hypothesis. The author states that:

> It also allowed me to postulate a reduction of the tripartite system of Spanish demonstrative pronouns into a basic binary system whereby *esto* and *eso* would be grouped together as being unspecified with respect to proximity and *aquello* being the term most frequently used in modern Spanish to mark distance in the spatio-temporal axis (Zulaica Hernández, 2007, p.iii).

His results confirm what can be seen in the three figures, particularly in Figure 6.8 where it is possible to see the prominence of this pronoun in the Penilsular Spanish sample.[85]

The situation regarding *aquel*, *este* and *ese* may be related to that of Portuguese with *aquele*, *este* and *esse*. In spoken Brazilian Portuguese, for example, the pronouns *este* and *esse* are used interchangeably. The work from Pereira (2005) on Brazilian and European Portuguese supposes this, and the author states that:

> *Os nossos dados mostraram que na fala do brasileiro a forma 'este' apenas residual e está em vias de substituição.* - Our results show that in spoken Brazilian data, the form 'este' is being substituted (Pereira, 2005, p.5).

6.6 Chapter Summary

The first sections of this chapter presented results obtained by several experiments using linguistically motivated features such as POS tags, morphology, and hybrid

[85] It should be noted that both this work and Zulaica Hernández (2007) refer to all inflected forms of the pronouns. For example, I use *aquel* as the lemma whereas Zulaica Hernández (2007) uses *aquello*.

delexicalized representations. Using the information provided by the classification output, I looked in more detail at the most important features that are used to discriminate Brazilian and European Portuguese and four Spanish varieties. I compared the most informative lexical features that discriminate Brazilian and European Portuguese to the standard keyword methods used in corpus linguistics. With this comparison, I observed a significant overlap between the most important features in classification and the top ranked keywords used in corpus linguistics. This confirms the hypothesis that a well-designed classification experiment can be extremely useful for linguists interested in different levels of linguistic variation.

Regarding the Spanish classification experiments, I looked in detail at the results obtained when using POS tags and observed interesting linguistic features that differ across the four varieties. I discussed the cases of past tenses and demonstrative pronouns as examples of how to formulate hypotheses based on the output of automatic classification and also investigated these aspects in more detail using different forms of data visualization and analysis.

This chapter allows me to answer **RQ4:** Can we use the information obtained from automatic classifiers to study differences between language varieties? In my opinion, given what was discussed in this chapter, the information obtained from the classifiers can be a relevant resource for linguistic research. Representing the text with linguistically informative features instead of the classic word and character *n*-gram models has also proved to be an interesting strategy to study linguistic variation. The linguistic analysis contained in this chapter is, of course, preliminary and far from exhaustive, and it invites further investigation by linguists and philologists. Even so, I am convinced that this chapter provides information on the use of classification methods within the scope of contrastive linguistics research and offers meaningful examples on a level corresponding to the current standard of research in corpus-based language studies.

Chapter 7

Conclusion

This dissertation presented several experiments on language identification focusing on language varieties. The first chapter described how language identification has been substantially explored in the literature in NLP, while discriminating between closely related languages, varieties and dialects has not been substantially studied and still presents a challenge for most systems. The experiments described here contribute to the improvement of this situation and the insights derived from them fill a gap in the literature.

The experiments presented in this thesis applied different algorithms, features, and datasets as follows:

- **Algorithms:** Several algorithms were used in a series of experiments and their performance was evaluated. Examples of algorithms include the likelihood estimation discriminative algorithm, Support Vector Machines (SVMs) in their SMO implementation, Multinomial Naive Bayes (MNB), and the J48 decision tree algorithm. A comprehensive comparison of the systems that participated in the DSL shared tasks was also provided. In Chapter 4 I report impressive results of 99.8% accuracy in discriminating between Brazilian and European Portuguese texts using character 4-grams and likelihood estimation. Using a multilingual setting containing 14 classes (13 similar languages and language varieties plus a class including a collection of texts from different languages), the best system in the DSL challenge 2016 obtained 95.54% using SVM ensembles.

- **Features:** Unlike most language identification approaches, I used linguistically motivated features to investigate differences between language varieties. For the classification of Spanish texts, for example, the algorithms classified texts based solely on a combination of part-of-speech and morphology as described in Zampieri et al. (2013). For a multilingual dataset, in the version 2.0 of the DSL corpus collection (DSLCC), most named entities were replaced by place holders which allowed me evaluate the extent to which named entities and thematic bias influences the classifiers' performance.

- **Datasets:** The classification experiments dealt with a number of languages including four pluricentric languages with special focus on the Romance branch. Portuguese, Spanish, French and English were explored in detail with different national varieties for each of them. Furthermore, the DSLCC, a multilingual corpus of similar languages and language varieties, has been compiled and is freely available for other researchers to use for language identification or other purposes. To my knowledge, the corpus has already been used outside of the DSL shared task in related text classification tasks and as part of the NLP teaching curriculum in computational linguistics at Indiana University[86] and in students' projects in different courses at Stanford University.[87,88]

Moving beyond the classification methods themselves, in Chapter 6, I provided a description of how these methods can be used in contrastive linguistics research. I did so by comparing, for example, the most important features in classification with the keywords produced by popular associative metrics used in corpus linguistics. Methods such as these are of interest to linguists who work with linguistic variation and are interested in applying these methods on annotated corpora to investigate differences between language varieties in terms of lexicon or syntax.

Using the output of the experiments on Portuguese and Spanish, I looked more closely at the most informative lexical features for Portuguese and also the identification of Spanish varieties based on POS tags including two concrete examples on how the methods can be used to examine syntactic differences between two samples. The first of the two examples is the use of compound past and simple past, and the second one concerns the use of demonstrative pronouns.

An interesting finding of this study described in Chapter 6 is the strong evidence that language varieties that can be distinguished with high accuracy based on knowledge-poor features may also be distinguished based on their POS distribution with similar success. This seems to indicate that the results obtained by character and word n-grams are not a coincidence but generated by intrinsic systemic differences between language varieties.

Another interesting finding of my experiments is that in multilingual settings, discriminative algorithms need a certain threshold of confidence to distinguish varieties in real-world settings. When results are below a certain level, loss in performance

[86] `http://cl.indiana.edu/~md7/14/715/`

[87] `http://cs229.stanford.edu/proj2015/335_report.pdf`

[88] `http://nlp.stanford.edu/courses/cs224n/2015/reports/24.pdf`

becomes substantial. Future work may use other settings and algorithms to estimate this threshold.

7.1 Research Questions Revisited

- **RQ1:** Is it possible to automatically discriminate between language varieties with satisfactory performance?
  For the set of languages and language varieties I experimented with, results from Chapter 4 confirm that the task is feasible and that the performance varies between languages and language varieties. Performance depends on a number of factors including how similar the language varieties are, the number of language varieties to be discriminated, and whether they share the same orthography (e.g. Brazilian and European Portuguese and British and American English have moderate differences in orthography). As expected, the best results were obtained in binary settings. Chapter 4 offers results of up to 99,8% accuracy in discriminating between Brazilian and European Portuguese using words and characters as features and the likelihood estimation algorithm.(Zampieri and Gebre, 2012)[89] One further interesting outcome of my experiments, presented in Chapter 6, is the demonstrated correlation between performance using words and characters and results obtained using POS tags and morphosyntactic information for Spanish, which indicates that variation occurs beyond orthography and the lexicon (Zampieri et al., 2013).

- **RQ2:** Can language varieties be integrated into real-world language identification systems?
  My experiments confirm that the integration of language varieties into real-world language identification systems is possible with moderate loss of performance. Results presented in Chapter 4, in an experiment containing 17 languages, suggest that there seems to be a performance threshold below which performance drops and classifiers start to assign all documents from a language to one of the two classes (e.g. all Portuguese texts labelled as Brazilian Portuguese). For languages in which the algorithms perform well in monolingual settings (e.g. Portuguese), performance loss is not significant which allows researchers to integrate language varieties in broader language identification schemes. The results obtained by the teams who participated in the two editions of the DSL shared task (Zampieri et al., 2014, 2015b), presented in Chapter 5, also confirm that the task is feasible, particularly when using two-stage approaches, training classifiers to identify first the language

[89] Named entities were present in these texts.

and subsequently the language variety (Goutte et al., 2014).

- **RQ3:** What are the most efficient features and algorithms to discriminate between language varieties?
  The answer to this question can be summarized in two aspects:
  - Agorithms: In the DSL shared task, with a few exceptions, most of the best performing systems used variations of either Naive Bayes or Support Vector Machines (SVM). Systems using linear SVM were the best performing systems in the 2014 (Goutte et al., 2014) as well as in the 2015 editions of the challenge (Malmasi and Dras, 2015b). In my experiments using BoW presented in Chapter 4 and in Zampieri (2013), SVM also proved to be very efficient by outperforming Multinomial Naive Bayes and J48. The difference in performance between MNB and SVM was, however, small. In this thesis I further investigated LE, a Bayesian probabilistic classifier combined with Laplace smoothing I first proposed for Portuguese in Zampieri and Gebre (2012), this time comparing its performance and speed to SVM and Logistic Regression using the DSLCC dataset (Zampieri et al., 2015a). LE is simple, fast, and delivers performance comparable to the state-of-the-art in this task using only a few thousand instances from each class. One interesting aspect that requires further investigation is the use of sentence-to-vector and word-to-vector representations for this task. The use of deep learning has proved to be a trend in NLP in the past few years, but for this task, an approach based on these methods, did not perform better than other systems in the 2015 edition of the DSL shared task (Franco-Salvador et al., 2015). My hypothesis is that the 18,000 instances from each language available at the DSLCC training set are not enough training data for these methods which require more training material to perform well.
  - Features: General-purpose language identification usually achieves best results using character trigrams. Based on the results presented in this thesis, I contend that systems trained to discriminate between similar languages perform in many scenarios better using either higher-order character *n*-grams (4-grams, 5-grams, and even 6-grams) or word-based features such as unigrams or BoW (Zampieri, 2013). In Chapter 5, using the results and corpora from the DSL shared task I showed for the first time that named entities impact classification's performance, but their influence is not as great as one might imagine. We included a blinded NE test set in the second edition of the DSL shared task and the results obtained by all teams confirm my findings (Zampieri et al., 2015b). Finally, from a purely engineering perspective, my results suggest that the use

of linguistically-motivated features using morphosyntactic information or delexicalized representations does not bring improvements to system performance for this task. The use of these representations are, however, a relevant source of information for linguistic research (see **RQ4**).

- **RQ4:** Can we use the information obtained from automatic classifiers to study differences between language varieties?
  It is possible to use the information obtained by classifiers to study differences between language varieties in terms of lexicon and syntax. Chapter 6 details the use of Brazilian and European Portuguese samples to compare the most informative lexical features from the output of a Naive Bayes classifier to a keyword list. Keyword lists are typically used in corpus linguistics and they were produced comparing both samples using mutual information. There was a significant overlap (over 70%) between keyword lists and the list of most informative words from each sample. The use of linguistically motivated features also contributes in this direction as it is possible to study syntactic properties of language varieties by analysing them as confirmed by the examples carried out using Spanish varieties (Zampieri et al., 2013).

7.2 Future Work

There are a number of open questions and issues I see being investigated in the future. Some of these are:

- What is the optimal text representation that is both linguistically informative and delivers best performance in classification? I tested different ways of representing texts and I think my work has made a contribution in this direction. The features that achieve the best performance for this task are knowledge-poor features represented by characters and words. Models trained on knowledge-rich features did not obtain performance comparable to methods trained on knowledge-poor features. It should be investigated further whether mixed representations can be used to obtain both good performance and linguistically relevant information.

- A recent trend in NLP applications in the last years is the use of deep learning methods and the revival of neural networks applied to NLP tasks. These methods have been applied to a number of different NLP tasks from parsing to semantic modelling as well as to general-purpose language identification as proposed by Simões et al. (2014). The aim is to search for new forms of data and feature representations that allow algorithms to capture properties and patterns in the data that common machine learning algorithms cannot capture.

Research in Franco-Salvador et al. (2015), however, showed that using these techniques does not outperform common machine learning classifiers such as SVM. My hypothesis is that the DSL dataset is not as large as necessary to train robust deep learning models. In light of this, one possible future research direction is to investigate the different ways that neural networks and deep learning methods can be used to discriminate between similar languages and whether they are able to achieve performance superior to methods relying on other forms of machine learning and statistics.

- I would like to investigate how humans perceive differences in language varieties by carrying out experiments with human annotators such as those proposed by Goutte et al. (2016). Goutte et al. (2016) report that results obtained by several annotators when discriminating texts from Brazilian and European Portuguese and Argentinian and Peninsular Spanish were substantially lower than the classification performance. This again confirms the difficulty of this task and motivates the investigation of this issue in future work.

Apart from the open questions and research directions summarized above and following up the work presented in this dissertation, I am currently involved in further projects related to this topic. Some of these include:

- Applying distributional semantics methods for the identification of lexical variation between language varieties as described in Ljubešic and Fišer (2013) for similar languages.

- The organization of the third edition of the VarDial workshop and the DSL Shared Task which will once again enable systems to be evaluated using the same dataset and metrics. More details are presented next.

Regarding the DSL shared task, myself, along with the other researchers involved would like to organize a third edition of the challenge in 2016. The third edition of the DSL shared task is structured as two sub-tasks as follows:

- **Sub-task 1:** Language varieties and similar languages using an updated version of the DSL corpus collection (DSLCC) for training and two test sets.
  The following are languages included in this edition: Bosnian, Croatian and Serbian; Indonesian and Malay; Brazilian Portuguese and European Portuguese; Peninsular Spanish and Argentine Spanish; Bulgarian and Macedonian; Canadian French and Mainland French.

  - Test Set A: in-domain evaluation. The test set will follow the same distribution as the DSLCC dataset and it will contain a large number of 'unseen' languages not present in the training data in order to emulate a realistic language identification scenario.

  - Test Set B: out-of-domain evaluation with social media texts. We are interested to see how domain influences classification and how algorithms perform when training in standard texts (journalistic) and tested in non-standard data.

- **Sub-task 2:** Arabic dialect identification using literary texts. This initiative reflects the growing interest in processing Arabic dialects and more specifically in their identification evidenced by a number of studies published recently (Zaidan and Callison-Burch, 2014; Sadat et al., 2014). For this task, we will provide training and test data from the same domain. Dialects from the following countries will be included: United Arab Emirates, Bahrain, Kuwait, Qatar, Saudi Arabia and Oman.

Bibliography

Ács, J., Grad-Gyenge, L., and de Rezende Oliveira, T. B. R. (2015). A Two-level Classifier for Discriminating Similar Languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 73–77, Hissar, Bulgaria.

Adams, G. and Resnik, P. (1997). A Language Identification Application Built on the Java Client/server Platform. In *Proceedings of the Workshop on Research to Commercial Applications: Making NLP Work in Practice*, pages 43–47, Madrid, Spain.

Amine, A., Elberrichi, Z., and Simonet, M. (2010). Automatic Language Identification: An Alternative Unsupervised Approach Using a New Hybrid Algorithm. *International Journal of Computer Science and Applications*, 7:94–107.

Anstein, S. (2012). Comparing Variety Corpora with Vis-A-Vis - a Prototype System Presentation. In *Proceedings of KONVENS*, pages 243–247, Vienna, Austria.

Anstein, S. (2013). *Computational Approaches to the Comparison of Regional Variety Corpora: Prototyping a Semi-automatic System for German.* PhD thesis, University of Stuttgart.

Anthony, L. (2005). AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom. In *Proceedings of the International Professional Communication Conference (IPCC)*, pages 729–737, Limerick, Ireland.

Anthony, L. (2013). Developing AntConc for a New Generation of Corpus Linguists. In *Proceedings of the Corpus Linguistics Conference*, pages 14–16, Lancaster, United Kingdom.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., and Levitan, S. (2007). Stylistic Text Classification Using Functional Lexical Features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

Baldwin, T. and Lui, M. (2010). Multilingual Language Identification: ALTW 2010 Shared Task Data. In *Proceedings of Australasian Language Technology Association Workshop (ALTA)*, pages 4–7, Melbourne, Australia.

Baptista, J., Vieira, L. N., Diniz, C., and Mamede, N. (2012). Coordination of -mente Ending Adverbs in Portuguese: an Integrated Solution. In *Proceedings of the Internacional Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 24–34. Springer.

Baxter, A. (1992). Portuguese as a Pluricentric Language. In Clyne, M., editor, *Pluricentric Languages: Different Norms in Different Nations*, pages 11–43. CRC Press.

Beesley, K. (1988). Language Identifier: A Computer Program for Automatic Natural-Language Identification of On-line Text. In *Proceedings of the Annual Conference of the American Translators Association*, pages 57–54.

Ben-Hur, A. and Weston, J. (2010). A Users Guide to Support Vector Machines. *Data Mining Techniques for the Life Sciences*, pages 223–239.

Biemann, C., Heyer, G., Quasthoff, U., and Richter, M. (2007). The Leipzig Corpora Collection - Monolingual Corpora of Standard Size. In *Proceedings of Corpus Linguistics*, Birmingham, United Kingdom.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. (2013). TOEFL11: A Corpus of Non-Native English. Technical report, Educational Testing Service.

Bobicev, V. (2015). Discriminating between Similar Languages Using PPM. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 59–65, Hissar, Bulgaria.

Brown, R. D. (2012). Finding and Identifying Text in 900+ Languages. *Digital Investigation*, 9:S34–S43.

Brown, R. D. (2013). Selecting and Weighting N-Grams to Identify 1100 Languages. In *Proceedings of the 16th International Conference on Text Speech and Dialogue (TSD2013), Lecture Notes in Artificial Intelligence (LNAI 8082)*, pages 519–526, Pilsen, Czech Republic. Springer.

Brown, R. D. (2014). Non-linear Mapping for Improved Identification of 1300+ Languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–623, Doha, Qatar.

Carter, S., Weerkamp, W., and Tsagkias, M. (2013). Microblog Language Identification: Overcoming the Limitations of Short, Unedited and Idiomatic Text. *Language Resources and Evaluation*, 47(1):195–215.

Castro, I. (1991). *Curso de Historia da Lingua Portuguesa*. Universidade Aberta.

Cavnar, W. and Trenkle, J. (1994). N-gram-based Text Catogorization. *Proceedings of the 3rd Symposium on Document Analysis and Information Retrieval (SDAIR)*.

Ceylan, H. and Kim, Y. (2009). Language Identification of Search Engine Queries. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1066–1074, Singapore.

Chambers, J. and Trudgill, P. (1998). *Dialectology (2nd Edition)*. Cambridge University Press.

Cheng, N., Chandramouli, R., and Subbalakshmi, K. (2011). Author Gender Identification from Rext. *Digital Investigation*, 8(1):78–88.

Chew, Y. C., Mikami, Y., Nagano, R. L., et al. (2011). Language Identification of Web Pages Based on Improved N-gram Algorithm. *International Journal of Computer Science Issues*, 8(3):47–58.

Christensen, H. (2014). Hc corpora. `http://www.corpora.heliohost.org/`.

Ciobanu, A. M. and Dinu, L. P. (2014). Building a Dataset of Multilingual Cognates for the Romanian Lexicon. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 1038–1043, Reykjavik, Iceland.

Ciobanu, A. M. and Dinu, L. P. (2016). A Computational Perspective on Romanian Dialects. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 633–643, Portoroz, Slovenia.

Clyne, M. (1992). *Pluricentric Languages: Different Norms in Different Nations*. CRC Press.

Combrinck, H. and Botha, E. (1994). Text-based Automatic Language Identification. In *Proceedings of the 6th Annual South African Workshop on Pattern Recognition*.

Cook, P. and Lui, M. (2014). langid.py for Better Language Modelling. In *Proceedings of the Australasian Language Technology Association Workshop (ALTA)*, pages 107–112, Dunedin, New Zealand.

Cotterell, R. and Callison-Burch, C. (2014). A Multi-dialect, Multi-genre Corpus of Informal Written Arabic. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 241–245, Reykjavik, Iceland.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

de Amorim, R. C. and Mirkin, B. (2012). Minkowski Metric, Feature Weighting and Anomalous Cluster Initializing in K-Means Clustering. *Pattern Recognition*, 45(3):1061–1075.

Dehdari, J. (2014). *A Neurophysiologically-Inspired Statistical Language Model*. PhD thesis, Ohio State University.

Dipper, S. (2008). Theory-driven and Corpus-driven Computational Linguistics, and the Use of Corpora. In *Corpus Linguistics. An International Handbook.*, pages 68–96. Mouton de Gruyter.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics - Special Issue on Using Large Corpora*, 19(1).

Dunning, T. (1994). Statistical Identification of Language. Technical report, Computing Research Lab - New Mexico State University.

Elfardy, H. and Diab, M. T. (2013). Sentence Level Dialect Identification in Arabic. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 456–461, Sofia, Bulgaria.

Emerson, G., Tan, L., Fertmann, S., Palmer, A., and Regneri, M. (2014). SeedLing: Building and using a seed corpus for the Human Language Project. pages 77–85, Baltimore, United States.

Fabra-Boluda, R., Rangel, F., and Rosso, P. (2015). NLEL UPV Autoritas participation at Discrimination between Similar Languages (DSL) 2015 Shared Task. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial, pages 52–58, Hissar, Bulgaria.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Ferguson, C. (1959). Diglossia. *Word*, 15:325–340.

Franco-Salvador, M., Rosso, P., and Rangel, F. (2015). Distributed Representations of Words and Documents for Discriminating Similar Languages. In *Proceedings* *of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial, pages 11–16, Hissar, Bulgaria.

Frank, E. and Bouckaert, R. R. (2006). Naive Bayes for Text Classification with Unbalanced Classes. In *Proceedings of Knowledge Discovery in Databases (KDD)*, pages 503–510. Springer.

Gebre, B. G., Zampieri, M., Wittenburg, P., and Heskens, T. (2013). Improving Native Language Identification with TF-IDF Weighting. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 216–233, Atlanta, United States.

Giwa, O. and Davel, M. H. (2013). N-gram based Language Identification of Individual Words. In *Proceedings of the 24th Annual Symposium of the Pattern Recognition Association of South Africa*, pages 15–22, Johannesburg, South Africa.

Gold, M. (1967). Language Identification in the Limit. *Information and Control 10*, pages 447–474.

Good, J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika*, 40:237–264.

Goutte, C. and Léger, S. (2015). Experiments in Discriminating Similar Languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial, pages 78–84, Hissar, Bulgaria.

Goutte, C., Léger, S., and Carpuat, M. (2014). The NRC System for Discriminating Similar Languages. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 139–146, Dublin, Ireland.

Goutte, C., Léger, S., Malmasi, S., and Zampieri, M. (2016). Discriminating Similar Languages: Evaluations and Explorations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1800–1807, Portoroz, Slovenia.

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

Grefenstette, G. (1995). Comparing Two Language Identification Schemes. In *Proceedings of the 3rd International Conference on Statistical Analysis of Textual Data (JADT)*, Rome, Italy.

Groethe, L., De Luca, E., and Nürnberger, A. (2008). A Comparative Study on Language Identification Methods. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 980–985, Marrakesh, Marroco.

Grouin, C., Forest, D., Da Sylva, L., Paroubek, P., and Zweigenbaum, P. (2010). Présentation et Résultats du Défi Fouille de Texte DEFT2010 Où et Quand un Article de Presse a-t-il Été Écrit? In *Proceedings of the 6th Défi Fouille de Textes (DEFI) at TALN*, Montreal, Canada.

Hammond, M. (2007). Introduction to the Mathematics of Language. University of Arizona.

Harris, M. (1982). The past simple and the present perfect in romance. *Studies in the Romance verb*, pages 42–70.

Harris, Z. S. (1954). Distributional structure. *Word*.

Herring, S. C. and Paolillo, J. C. (2006). Gender and Genre Variation in Weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Hollenstein, N. and Aepli, N. (2015). A Resource for Natural Language Processing of Swiss German Dialects. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 108–109, Duisburg, Germany.

Huang, C. and Lee, L. (2008). Contrastive Approach towards Text Source Classification based on Top-Bag-of-Word Similarity. In *Proceedings of PACLIC*, pages 404–410, Cebu City, Phillipines.

Ingle, N. (1980). *A Language Identification Table.* Technical Translation International.

Ionescu, R. T. and Popescu, M. (2016). Native language identification with string kernels. In *Knowledge Transfer between Computer Vision and Text Mining*, pages 193–227. Springer.

Ionescu, R. T., Popescu, M., and Cahill, A. (2014). Can Characters Reveal Your Native Language? A Language-independent Approach to Native Language Identification. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1363–1373, Doha, Qatar.

Jarvis, S., Bestgen, Y., and Pepper, S. (2013). Maximizing Classification Accuracy in Native Language Identification. In *Proceedings of the 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA8)*, pages 111–118, Atlanta, United States.

Jauhiainen, T., Jauhiainen, H., and Lindén, K. (2015a). Discriminating Similar Languages with Token-based Backoff. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial, pages 44–51, Hissar, Bulgaria.

Jauhiainen, T., Lindén, K., and Jauhiainen, H. (2015b). Language Set Identification in Noisy Synthetic Multilingual Documents. In *Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, pages 633–643, Cairo, Egypt.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 137–142, Chemnitz, Germany.

Joachims, T. (2006). Training linear SVMs in Linear Time. In *Proceedings of Knowledge Discovery in Databases (KDD)*, pages 217–226.

Jurafsky, D. and Martin, J. (2009). *Speech and Language Processing (2nd Edition).* Prentice Hall, 2 edition.

Kabatek, J. and Pusch, C. (2009). *Spanische Sprachwissenschaft: Eine Enführung.* Gunter.

Katz, S. (1987). Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35:400–401.

Kelleher, J. D., Mac Namee, B., and D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked examples, and Case Studies.* MIT Press.

Kibriya, A., Frank, E., Pfahringer, B., and Holmes, G. (2004). Multinomial Naive Bayes for Text Categorization Revisited. In *Proceedings of the Australian Conference on Artificial Intelligence*, pages 488–499.

King, B., Radev, D., and Abney, S. (2014). Experiments in Sentence Language Identification with Groups of Similar Languages. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 146–154, Dublin, Ireland.

Kloss, H. (1952). *Die Entwicklung neuer germanischer Kultursprachen seit 1800.* IDS.

Kloss, H. (1967). 'Abstand languages' and 'Ausbau languages'. 9(7):29–41.

Kneser, R. and Ney, H. (1993). Improvde Clustering Techniques for Class-Based Statistical Language Models. In *Proceedings of EUROSPEECH*, pages 973–976, Berlin, Germany.

Kochmar, E. (2011). Identification of a Writer's Native Language by Error Analysis. Master's thesis, University of Cambridge, United Kingdom.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Koller, O., Abad, A., Trancoso, I., and Viana, C. (2010). Exploiting Variety-dependent Phones in Portuguese Variety Identification Applied to Broadcast News Transcription. In *Proceedings of INTERSPEECH*, pages 749–752, Makuhari, Japan.

Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412.

Koppel, M., Schler, J., and Zigon, K. (2005). Automatically Determining an Anonymous Author's Native Language. *Lecture Notes in Computer Science*, 3495:209–217.

Kotz, S., Kozubowski, T., and Podgorski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Exonomics, Engineering, and Finance*. Number 183. Springer.

Lado, R. (1957). *Applied Linguistics for Language Teachers*. University of Michigan Press.

Lê, S., Josse, J., Husson, F., et al. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of statistical software*, 25(1):1–18.

Leitner, G. (1989). *Core Grammar Versus Variety Grammar - the Case of English*, pages 163–183. Niemeyer.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Ljubešic, N. and Fišer, D. (2013). Identifying False Friends between Closely Related Languages. *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 69–77.

Ljubešic, N., Fišer, D., and Erjavec, T. (2014). Tweet-CaT: a Tool for Building Twitter Corpora of Smaller Languages. In *Proceedings of the Internacional Conference on Language Resources and Evaluation (LREC)*, pages 2279–2283, Reykjavik, Iceland.

Ljubešić, N., Mikelic, N., and Boras, D. (2007). Language Identification: How to Distinguish Similar Languages? In *Proceedings of the 29th International Conference on Information Technology Interfaces*, pages 541–546.

Lope Blanch, J. (1961). Sobre el Uso del Preterito en el Espanol de Mexico. *Studia Philologica*, pages 131–143.

Lui, M. (2014). *Generalized Language Identification.* PhD thesis, University of Melbourne.

Lui, M. and Baldwin, T. (2011). Cross-domain Feature Selection for Language Identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 553–561, Chiang Mai, Thailand.

Lui, M. and Baldwin, T. (2012). langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Meeting of the Association for Computational Linguistics (ACL)*, pages 25–30, Jeju, Korea.

Lui, M. and Cook, P. (2013). Classifying English Documents by National Dialect. In *Proceedings of Australasian Language Technology Workshop (ALTA)*, pages 5–15, Brisbane, Australia.

Lui, M., Letcher, N., Adams, O., Duong, L., Cook, P., and Baldwin, T. (2014). Exploring Methods and Resources for Discriminating Similar Languages. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 139–148, Dublin, Ireland.

Malmasi, S. (2016). *Native Language Identification: Explorations and Applications.* PhD thesis, Macquarie University.

Malmasi, S. and Cahill, A. (2015). Measuring Feature Diversity in Native Language Identification. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 49–55, Denver, United States.

Malmasi, S. and Dras, M. (2015a). Automatic Language Identification for Persian and Dari Texts. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 59–64, Bali, Indonesia.

Malmasi, S. and Dras, M. (2015b). Language Identification Using Classifier Ensembles. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial5, pages 35–43, Hissar, Bulgaria.

Malmasi, S. and Dras, M. (2015c). Multilingual Native Language Identification. *Natural Language Engineering*, pages 1–53.

Malmasi, S., Refaee, E., and Dras, M. (2015). Arabic Dialect Identification Using a Parallel Multidialectal Corpus. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 209–217, Bali, Indonesia.

Malmasi, S. and Zampieri, M. (2016). MAZA at SemEval-2016 Task 11: Detecting Lexical Complexity Using a Decision Stump Meta-Classifier. In *Proceedings of 10th Workshop on Semantic Evaluation (SemEval)*, pages 991–995, San Diego, United States.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Martinez, M. J. M. and Tan, L. (2016). USAAR at SemEval-2016 Task 11: Complex Word Identification with Sense Entropy and Sentence Perplexity. In *Proceedings of 10th Workshop on Semantic Evaluation (SemEval)*, pages 958–962, San Diego, United States.

Martins, B. and Silva, M. (2005). Language Identification in Web Pages. In *Proceedings of the 20th ACM Symposium on Applied Computing (SAC), Document Engineering Track*, pages 763–768, Santa Fe, United States.

McCallum, A. (2002). Mallet: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu`.

McEnery, T. and Hardie, A. (2011). *Corpus Linguistics*. Cambridge University Press.

Medlock, B. (2008). Investigating Classification for Natural Language Processing Tasks. Technical report, University of Cambridge - Computer Laboratory.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference of Learning Representations*, Scottsdale, United States.

Mitchell, T. (1997). *Machine learning*. McGraw-Hill.

Moisl, H. (2009). Using Electronic Corpora in Historical Dialectology Research: The Problem of Document Length Variation. In *Studies in English and European Historical Dialectology*, pages 67–90. Peter Lang.

Mubarak, H. and Darwish, K. (2014). Using Twitter to Collect a Multi-dialectal Corpus of Arabic. pages 1–7, Doha, Qatar.

Ney, H., Essen, U., and Kneser, R. (1994). On Structuring Probabilistic Dependence in Stochastic Language Modelling. *Computer Speech and Language*, 8:1–38.

Nguyen, D. and Dogruoz, A. S. (2013). Word Level Language Identification in Online Multilingual Communication. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 857–862, Seattle, United States.

Nguyen, D.-P., Gravel, R., Trieschnigg, R., and Meder, T. (2013). "How Old Do You Think I Am?" A Study of Language and Age in Twitter. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 439–448, Boston, United States.

Niculae, V., Zampieri, M., Dinu, L. P., and Ciobanu, A. M. (2014). Temporal Text Ranking and Automatic Dating of Texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 17–21, Gothenburg, Sweden.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved Part-of-speech Tagging for Online Conversational Text with Word Clusters. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 380–391, Atlanta, United States.

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 2473–2479, Istanbul, Turkey.

Padró, M. and Padró, L. (2004). Comparing Methods for Language Identification. *Procesamiento del Lenguaje Natural*, (33):155–162.

Palmer, D. (2010). Text Processing. In Indurkhya, N. and Damerau, F., editors, *Handbook of Natural Language Processing - Second Edition*, pages 9–30. CRC Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

Peirsman, Y., Geeraerts, D., and Speelman, D. (2010). The Automatic Identification of Lexical Variation Between Language Varieties. *Natural Language Engineering*, 16:469–491.

Pereira, H. B. (2005). 'Esse' Versus 'Este' no Português Brasileiro e no Europeu. Master's thesis, Universidade de São Paulo.

Piantadosi, S. T. (2014). Zipfs Word Frequency Law in Natural Language: A Critical Review and Future Directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word Lengths Are Optimized for Efficient Communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Platt, J. (1998). Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In Schoelkopf, B., B. C. and Smola, A., editors, *Advances in Kernel Methods Support Vector Learning*.

Popescu, O. and Strapparava, C. (2015). Semeval-2015 Task 7: Diachronic Text Evaluation. In *Proceedings of 9th Workshop on Semantic Evaluation (SemEval)*, pages 870–878, Denver, United States.

Porta, J. and Sancho, J.-L. (2014). Using Maximum Entropy Models to Discriminate between Similar Languages and Varieties. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 120–128, Dublin, Ireland.

Poutsma, A. (2001). Applying Monte Carlo Techniques to Language Identification. In *Proceedings of Computational Linguistics in the Netherlands*, pages 179–189, Twente, Netherlands.

Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., and Aletras, N. (2015). Studying User Income Through Language, Behaviour and Affect in Social Media. *PloS one*, 10(9).

Purver, M. (2014). A Simple Baseline for Discriminating Similar Language. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 155–160, Dublin, Ireland.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Ranaivo-Malançon, B. (2006). Automatic Identification of Close Languages - Case study: Malay and Indonesian. *ECTI Transactions on Computer and Information Technology*, 2:126–134.

Rangel, F., Stamatatos, E., Moshe Koppel, M., Inches, G., and Rosso, P. (2013). Overview of the Author Profiling Task at PAN 2013. In *Proceedings of the CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 352–365, Valencia, Spain.

Rehurek, R. and Kolkus, M. (2009). Language Identification on the Web: Extending the Dictionary Method. In *Proceedings of International Conference on Computational Linguistics and Intelligent Text Processing (CICLING). Lecture Notes in Computer Science (LNCS)*, pages 357–368. Springer.

Richter, M., Quasthoff, U., Hallsteinsdóttir, E., and Biemann, C. (2006). Exploiting the Leipzig Corpora Collection. In *Proceesings of the IS-LTC Language Technologies Conference*, Ljubljana, Slovenia.

Sadat, F., Kazemi, F., and Farzindar, A. (2014). Automatic Identification of Arabic Language Varieties and Dialects in Social Media. In *Proceedings of SocialNLP*, pages 22–27, Dublin, Ireland.

Sajjad, H., Darwish, K., and Belinkov, Y. (2013). Translating Dialectal Arabic to English. In *Proceedings of the 51st Annual Meeding of the Association for Computational Linguistics (ACL)*, pages 1–6, Sofia, Bulgaria.

Salloum, W., Elfardy, H., Alamir-Salloum, L., Habash, N., and Diab, M. (2014). Sentence Level Dialect Identification for Machine Translation System Selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 772–778, Baltimore, USA.

Salloum, W. and Habash, N. (2013). Dialectal Arabic to English machine translation: Pivoting through Modern Standard Arabic. In *Proceedings of the North American ACL*, pages 348–358.

Scannell, K. P. (2007). The Crúbadán Project: Corpus Building for Under-resourced Languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, pages 5–15, Louvain-la-Neuve, Belgium.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34:1–47.

Sebastiani, F. (2005). Text categorization. *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, pages 109–129.

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *Bell system technical journal*, 30(1):50–64.

Shimodaira, H. et al. (2004). Approximately Unbiased Tests of Regions Using Multistep-multiscale Bootstrap Resampling. *The Annals of Statistics*, 32(6):2616–2641.

Simões, A., Almeida, J. J., and Byers, S. D. (2014). Language Identification: a Neural Network Approach. *Proceedings of Slate.*

Smith, N. A., Cardie, C., Washington, A. L., and Wilkerson, J. D. (2014). Overview of the 2014 NLP Unshared Task in PoliInformatics. In *Proceedings of the Workshop on Language Technologies and Computational Social Science*, pages 5–7, Baltimore, United States.

Soares da Silva, A. (2010). Measuring and Parameterizing Lexical Convergence and Divergence between European and Brazilian Portuguese: Endo/Exogeneousness and Foreign and Normative Influence. *Advances in Cognitive Sociolinguistics*, pages 41–84.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Overview for the First Shared Task on Language Identification in Code-Switched Data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar.

Souter, C., Churcher, G., Hayes, J., Hughes, J., and Johnson, S. (1994). Natural Language Identification Using Corpus-based Models. *Hermes, Journal of Linguistics*, 13:183–203.

Squartini, M. and Bertinetto, P. M. (2000). The Simple and Compound past in Romance languages. *Empirical Approaches to Language Typology*, (6):403–440.

Stewart, W. (1968). A Sociolinguistic Typology for Describing National Multilingualism. In Fishman, J. A., editor, *Readings in the Sociology of Language*, pages 531–545. Mouton.

Takçı, H. and Güngör, T. (2012). A High Performance Centroid-based Classification Approach for Language Identification. *Pattern Recognition Letters*, 3:2077–2084.

Tan, L., Zampieri, M., Ljubešić, N., and Tiedemann, J. (2014). Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection. In *Proceedings of The Workshop on Building and Using Comparable Corpora (BUCC)*, pages 6–10, Reykjavik, Iceland.

Tetreault, J., Blanchard, D., and Cahill, A. (2013). A Report on the First Native Language Identification Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, United States.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, pages 2214–2218, Istanbul, Turkey.

Tiedemann, J. and Ljubešić, N. (2012). Efficient Discrimination Between Closely Related Languages. In *Proceedings of COLING*, pages 2619–2634, Mumbai, India.

Tillmann, C., Mansour, S., and Al-Onaizan, Y. (2014). Improved Sentence-Level Arabic Dialect Classification. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 110–119, Dublin, Ireland.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins.

Tomokiyo, L. and Jones, R. (2001). You're not from 'round here, are you?: Naive Bayes Detection of Non-native Utterance Text. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL)*, pages 239–246, Pittsburgh, United States.

Tonkim, E. and Tourte, G. E. (2016). *Working with Text: Tools, Techniques and Approaches for Text Mining*. Chandos Publishing, Elsevier.

Torres, L. S. and Aluísio, S. M. (2011). Using Machine Learning Methods to Avoid the Pitfall of Cognates and False Friends in Spanish-Portuguese Word Pairs. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL)*, pages 67–76, Cuiaba, Brazil.

Trieschnigg, D., Hiemstra, D., Theune, M., de Jong, F., and Meder, T. (2012). An Exploration of Language Identification Techniques for the Dutch Folktale Database. In *Proceedings of Eight International Conference on Language Resources and Evaluation (LREC)*, pages 47–51, Istanbul, Turkey.

Tromp, E. and Pechnizkiy, M. (2012). Graph-based n-gram Language Identification on Short Texts. In *Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning (Benelearn)*, pages 27–34, Ghent, Belgium.

Tsur, O. and Rappoport, A. (2007). Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16, Stroudsburg, United States.

Tyers, F. and Alperen, M. S. (2010). South-East European Times: A Parallel Corpus of Balkan Languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53, Valletta, Malta.

Van Rijsbergen, C. (1979). Information Retrieval. *Btterworths, London*.

Vogel, J. and Tresner-Kirsch, D. (2012). Robust Language Identification in Short, Noisy Texts: Improvements to LIGA. In *Third International Workshop on Mining Ubiquitous and Social Environments (MUSE)*, pages 43–50, Bristol, United Kingdom.

Vojtek, P. and Belikova, M. (2007). Comparing Language Identification Methods Based on Markov Processess. In *Proceedings of the International Seminar on Computer Treatment of Slavic and East European Languages (Slovko)*, pages 271–282, Bratislava, Slovakia.

Witten, I. and Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

Wong, S.-M. J. and Dras, M. (2009). Contrastive Analysis and Native Language Identification. *Proceedings of the Australasian Language Technology Association Workshop*, pages 53–61.

Xafopoulos, A., Kotropoulos, C., Almpanidis, G., and Pitas, I. (2004). Language Identification in Web Documents Using Discrete HMMs. *Pattern Recognition*, 37:583–594.

Xia, F., Lewis, C., and Lewis, W. D. (2010). The Problems of Language Identification within Hugely Multilingual Data Sets. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 2790–2797, Valletta, Malta.

Zaidan, O. F. and Callison-Burch, C. (2011). The Arabic Online Commentary Dataset: An Annotated Dataset of Informal Arabic with High Dialectal Content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 37–41, Portland, United States.

Zaidan, O. F. and Callison-Burch, C. (2014). Arabic Dialect Identification. *Computational Linguistics*, 40(1):171–202.

Zampieri, M. (2012). Evaluating Knowledge-poor and Knowledge-rich Features in Automatic Classification: A Case Study in WSD. In *13th IEEE International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 359–363, Budapest, Hungary.

Zampieri, M. (2013). Using Bag-of-words to Distinguish Similar Languages: How Efficient are They? In *Proceedings of the 14th IEEE International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 37–41, Budapest, Hungary.

Zampieri, M. (2016). Automatic Language Identification. In Tonkim, E. and Tourte, G., editors, *Working with Text: Tools, Techniques and Approaches for Text Mining*, pages 189–205. Chandos Publishing, Elsevier.

Zampieri, M. and Gebre, B. G. (2012). Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS*, pages 233–237, Vienna, Austria.

Zampieri, M. and Gebre, B. G. (2014). VarClass: An Open Source Language Identification Tool for Language Varieties. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3305–3308, Reykjavik, Iceland.

Zampieri, M., Gebre, B. G., Costa, H., and van Genabith, J. (2015a). Comparing approaches to the identification of similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, LT4VarDial '15, pages 66–72, Hissar, Bulgaria.

Zampieri, M., Gebre, B. G., and Diwersy, S. (2012). Classifying pluricentric languages: Extending the monolingual model. In *Proceedings of the Fourth Swedish Language Technlogy Conference (SLTC)*, pages 79–80, Lund, Sweden.

Zampieri, M., Gebre, B. G., and Diwersy, S. (2013). N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN*, pages 580–587, Sable d'Olonne, France.

Zampieri, M., Tan, L., Ljubešić, N., and Tiedemann, J. (2014). A Report on the DSL Shared Task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 58–67, Dublin, Ireland.

Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., and Nakov, P. (2015b). Overview of the DSL Shared Task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 1–9, Hissar, Bulgaria.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). Machine translation of Arabic Dialects. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 49–59, Montreal, Canada.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Zubiaga, A., San Vicente, I., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza, A., and Fresno, V. (2014). Overview of tweetLID: Tweet Language Identification at SEPLN 2014. In *Proceedings of the Tweet Language Identification Workshop (TweetLID)*, pages 1–11, Girona, Spain.

Zubiaga, A., San Vicente, I., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza, A., and Fresno, V. (2015). TweetLID: A Benchmark for Tweet Language Identification. *Language Resources and Evaluation*, pages 1–38.

Zulaica Hernández, I. (2007). *Demonstrative Pronouns in Spanish: a Discourse-based Study.* PhD thesis, Ohio State University.